Università degli Studi di Firenze
Dipartimento di Statistica "G. Parenti"

Dottorato di Ricerca in Statistica Applicata
XXII ciclo, SECS-S/01

Learning parametrico in presenza di soft-evidence. Valutazione probabilistica dell'età di individui viventi non adulti attraverso lo sviluppo del terzo molare

Iljà Barsanti

Tutor: **Prof.re Fabio Corradi**
Co-Tutor: **Prof.ssa Vilma Pinchi**
Coordinatore: **Prof.re Fabio Corradi**

Indice

Introduzione

Lo scopo di questa ricerca è classificare un individuo vivente non adulto in una determinata classe di età. Il crescente numero di immigrati sprovvisti di documenti di identità validi e la difficoltà nel risalire all'età anagrafica degli stessi, ha spinto le autorità competenti a rivolgersi ad esperti che siano in grado di attribuire un'età all'individuo oggetto di indagine.

Ciò che quindi si richiede all'esperto non è una stima dell'età vera e propria bensì la valutazione se un soggetto abbia raggiunto una determinata soglia d'età, che varia dai 14 ai 21 anni, consentendo di applicare correttamente le leggi e le normative vigenti nel Paese in cui il caso è trattato. La soglia di interesse a cui si pone particolare attenzione in questo lavoro è quella dei 18 anni, che in Italia discrimina l'individuo maggiorenne da quello minorenne.

Il metodo di attribuzione dell'età impiegato è quello che sfrutta l'informazione derivante dallo stato di mineralizzazione dei terzi molari, ritenuti i denti più idonei per la valutazione dell'età di individui che si collocano in un intorno della soglia dei 18 anni. Gli esperti odontologi valutano, quindi, il grado di maturazione dentale dei quattro terzi molari utilizzando la scala di classificazione dentale di Demirijan, ampiamente utilizzata per la sua chiarezza e semplicità.

La peculiarità di questo lavoro consiste nella possibilità che un esperto ha di esprimersi con incertezza dinanzi ad una valutazione. E' infatti usuale che prenda una decisione valutando in quale, fra un numero limitato di stati ordinati, si possa trovare l'unità osservata forzando la sua incertezza nella scelta di un solo stato. Nel presente lavoro, invece, viene data la possibilità all'esperto di indicare due o più stati adiacenti fra i quali di distribuisce l'incertezza, introducendo così la *soft evidence* caratterizzata dai gradi di fiducia assegnati agli stati.

Una variabile usualmente osservata congiuntamente allo sviluppo dentale dei terzi molari è il genere dell'individuo, il quale migliora l'attribuzione di in individuo ad una classe di età.

Poiché gli esperti non sono in grado di fornire identiche valutazioni sulle medesime radiografie dentali, questo suggerisce il fatto che ciascun modello classificatorio debba essere stimato separatamente per ciascun osservatore, non ammettendo quindi la scambiabilità degli osservatori.

Una volta che si è stimato il modello classificatorio *ad personam*, si procede al calcolo della predittiva per $l'(n+1)-esimo$ soggetto a partire dalla valutazione dentale fornita dall'esperto e dal genere dell'individuo. Mediante una regola decisionale si attribuisce così una classe di età all'individuo e, trattandosi di *learning* parametrico supervisionato si potranno valutare a posteriori le *performance* del modello sugli individui che costituiscono il *test data set*.

Nel caso la variabile di classe età sia dicotomica verrà posta particolare attenzione alla proporzione di falsi maggiorenni, vale a dire quegli individui che sono classificati come maggiorenni quando in realtà non lo sono, con conseguenze legali non trascurabili.

L'analisi dicotomica iniziale verrà poi ampliata considerando il caso di tre classi di età. E' infatti poco realistico un modello classificatorio che sia in grado di discriminare perfettamente individui di età in un piccolo intorno della soglia dei 18 anni utilizzando una variabile, come l'accrescimento dentale, che opera nel continuo. Per questo si è considerata una terza classe a cavallo di questa soglia con il ruolo di discriminare maggiormente gli individui che si trovano nelle due classi esterne. Il prezzo di questa riduzione degli errori, rispetto al caso dicotomico, risiede nel numero di individui classificati nella classe centrale, per i quali non si è in grado di stabilire se sono maggiorenni o minorenni.

Nel *Cap. 1* si introduce il problema di carattere forense e la classificazione dell'età di individui viventi non adulti mediante la valutazione dentale. Il metodo di classificazione di Demirijan viene presentato e confrontato con altre scale di classificazione dentale. Inoltre sono mostrati i lavori in letteratura in merito all'argomento trattato, generalmente basati sulla stima dell'età attraverso modelli di regressione lineare piuttosto che la assegnazione di individui ad una determinata classe di età. La *soft evidence* viene trattata come una valutazione naturale da parte di un esperto che meglio descrive l'incertezza dello stesso nel fornire il proprio parere su variabili di tipo ordinale. Una breve introduzione agli indici di riproducibilità e ripetibilità sottolinea come i modelli classificatori debbano essere stimati separatamente e come l'incertezza di valutazione, oltreché confermare la necessità dell'utilizzo della *soft evidence*, suggerisca un modello di transizione dell'incertezza stessa nel calcolo delle predittive.

Nel *Cap. 2* si introducono i concetti di *learning* con una accenno storico sull'apprendimento a partire dalla formula del teorema della probabilità condizionata di Bayes fino al *Machine learning* e le discipline ad esso correlate. Viene poi descritta la classificazione soffermandosi sul *Naive Bayes*, classificatore tanto semplice quanto efficace e competitivo, rispetto ad altri

classificatori che rilasciano il vincolo dell'assunzione di indipendenza condizionata degli attributi rispetto alla variabile di classe, proprio del *Naive Bayes*.

Nel *Cap. 3* si formalizza il problema, procedendo all'introduzione di assunzioni sulle distribuzioni di probabilità e sui legami di indipendenza fra le variabili di interesse. Si costruisce così il *Naive Bayes* modificato che altro non è che un *Naive Bayes* con l'introduzione della *soft evidence* e dell'insieme di covariate che maggiormente influenzano lo sviluppo dentale, in questo caso rappresentate dalla sola variabile genere. Si ricavano la verosimiglianza e la distribuzione a posteriori degli attributi in forma chiusa, per poi ottenere la predittiva, mediante la quale, scelta una regola decisionale, l'individuo verrà collocato in una determinata classe d'età. Inoltre verranno descritti alcuni indici di *performance* valutati sui *test data sets*, fra cui la riproducibilità, la ripetibilità, la capacità classificatoria del modello e la percentuale di errori commessi.

Nel *Cap. 4* si chiude il lavoro presentando i risultati inerenti la composizione del campione, gli osservatori ed infine la classificazione sia nel caso della variabile di classe dicotomica che tricotomica. L'accuratezza del modello classificatorio proposto è valutata nei due esperti sia mediante confronto diretto, condizionatamente alla tecnologia radiografica impiegata, sia mediante le prestazioni misurate in termini di percentuale di falsi minorenni prodotti e di capacità discriminatoria.

Capitolo 1
Assegnazione dell'età ad individui viventi non adulti

1.1 Introduzione

Negli ultimi anni la richiesta in ambito legale per la valutazione probabilistica dell'età di soggetti viventi, specie per quelli non adulti, è notevolmente aumentata. Questo fenomeno va letto in un contesto europeo in cui, in tempi recenti, si è stati testimoni dell'aumento del numero di immigrati privi di validi documenti di riconoscimento.

La valutazione dell'età assume perciò un ruolo rilevante nei processi, sia penali che civili, in cui l'identità dell'imputato è incerta e per il quale sia necessario verificare il raggiungimento o meno di determinate soglie d'età per consentire l'applicazione delle leggi. Queste problematiche si possono riscontrare anche nei casi di riconoscimento di paternità, di affidamento di minori provenienti dall'estero e nelle richieste d'asilo. In pratica, in tutte quelle situazioni ove siano coinvolti soggetti sprovvisti di documenti anagrafici validi oppure che presentano certificati di identità con irregolarità difficilmente risolvibili secondo la legislazione vigente nel paese che li ospita.

Da qui la necessità dei giudici e delle autorità competenti di ricorrere ad esperti che possano fornire una valutazione circa l'età del soggetto. In questo modo si garantisce l'applicazione delle leggi, delle misure correttive e delle procedure d'asilo o di affidamento, in funzione del superamento o meno di una soglia d'età che varia dai 14 ai 21 anni a seconda della legislazione e della materia trattata.

Da sottolineare come il giudizio di tali esperti richieda margini di incertezza differenti a seconda che il processo sia penale o civile e del tema

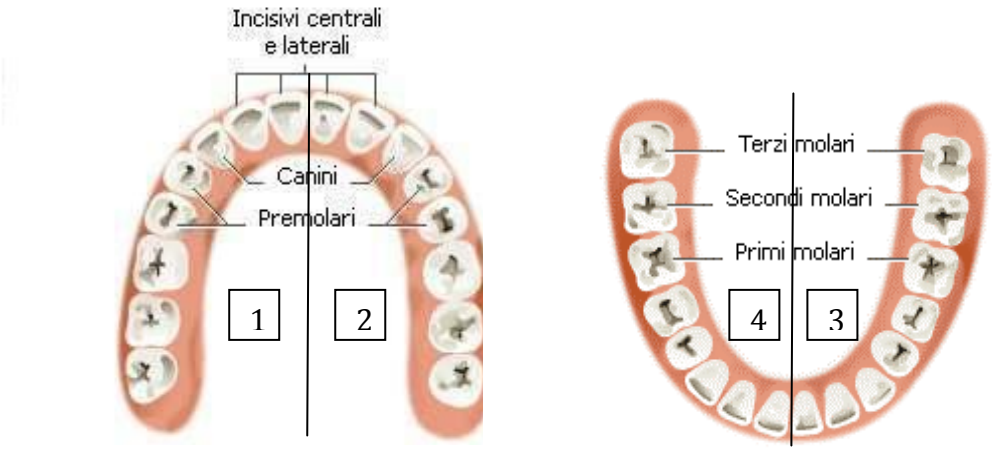
trattato. Questo significa che la valutazione probabilistica dell'età di un soggetto potrebbe essere differentemente impiegata secondo il livello di accettazione che varia da caso a caso.

1.2 Stima dell'età basata sullo sviluppo dentale: il metodo di Demirjian

Per l'assegnazione dell'età nel caso di individui viventi non adulti sono stati utilizzati vari criteri riguardanti il grado di maturazione fisica della persona. Fra questi, vi sono i metodi basati sullo sviluppo dentale, i quali possono comprendere l'esame di tutti i denti fino al secondo molare oppure del solo terzo molare, noto come "dente del giudizio".

Come illustrato nella *Fig 1.1* l'apparato dentale umano viene suddiviso in due arcate, *mascellare* o *superiore* e *mandibolare* o *inferiore*, le quali si dividono in semiarcate: superiore sinistra (1), superiore destra (2), inferiore destra (3) ed inferiore sinistra (4). Inoltre, ciascun dente viene classificato con una doppia numerazione: la prima, da 1 a 4, indicante la semiarcata di appartenenza mentre la seconda, da 1 a 8, identifica il dente a partire dall'incisivo centrale fino al terzo molare.

Figura 1.1. *Arcata mascellare (a sinistra) e mandibolare (a destra) nell'apparato dentale umano*

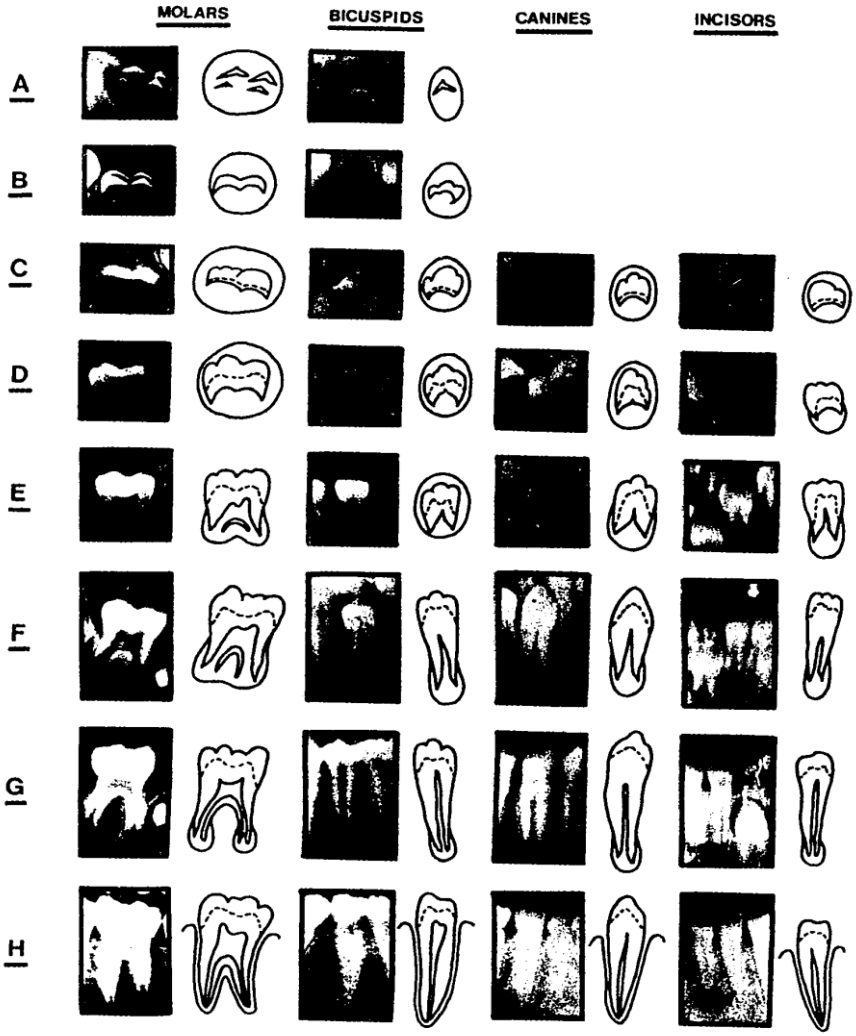


Nel 1973 il Dr. Arto Demirjian (Università di Montreal, Canada) ideò un metodo di assegnazione dell'età attraverso l'osservazione della maturazione dentale mediante radiografie ortopanoramiche (OPT). Demirjian et al. (1973) esaminarono un campione di soggetti di origine franco-canadese con età

compresa fra i 2 e 20 anni, osservando lo sviluppo dentale fino al secondo molare, vale a dire il grado di maturazione raggiunto da tutti i denti fatta eccezione dei terzi molari.

Il metodo di Demirjian si basa sulla classificazione dello sviluppo dentale in otto stati ordinati, come illustrato nella *Fig. 1.2*. Questi stati descrivono la morfologia della corona e della radice dentale, rispettivamente gli stati A-D e E-H, senza alcuna misurazione di lunghezza.

Figura 1.2. *Gli otto stati dentali secondo la classificazione di Demirjian (Demirjian et al., 1973, p.220)*



E' importante sottolineare come gli stati della classificazione di Demirjian siano volutamente non numerati per evitare l'idea che possano caratterizzare fasi di sviluppo dentale della stessa durata o siano legati a qualche misura appartenente alla scala a rapporti.

La proposta di Demirjian et al. (1973) non consiste solamente nella classificazione dei denti a seconda del grado di maturazione riconducibile ad una delle otto categorie suddette, ma comprende anche alcune tavole di conversione dentale. Mediante queste tavole si attribuisce, per ciascun dente e ciascuno stato, un punteggio la cui somma, per individuo, determina l'*indice di maturità globale* (GMI). Tale indice viene infine convertito in età anagrafica tramite ulteriori fattori di conversione.

La grande semplicità e chiarezza della scala illustrata in *Fig. 1.2* ha reso la classificazione di Demirjian più famosa dell'intera procedura di conversione. Infatti, sebbene tale classificazione nasca per i sette denti mandibolari, molti, fra cui Meinl et al. (2007), la utilizzano per valutare lo sviluppo dentale del terzo molare ritenuto di interesse per la maggior variabilità rispetto agli altri denti. Anche Pinchi et al. (2005) osservano il terzo molare per predire l'età fra i 16 e 19 anni poiché sotto i 16 anni esistono indicatori più attendibili come lo sviluppo dei denti fino al secondo molare e l'accrescimento osseo del polso e della mano (Cameriere et al., 2006). Al di sopra dei 19 anni Pinchi et al. (2005) considerano concluso il processo di maturazione dentale in quasi tutti gli individui. Inoltre questi ultimi autori sottolineano una significativa differenza fra le arcate dentali destre e sinistre in termini di correlazione fra età stimata e cronologica.

Un fattore fondamentale nello sviluppo dentale, che condiziona i punteggi delle tavole nel metodo proposto da Demirjian et al. (1973), è il genere. Infatti nei denti dei soggetti di sesso femminile il processo di mineralizzazione impiega più tempo a completarsi, iniziando prima e finendo più tardi di quello maschile. Pinchi et al. (2005), comunque, non ritengono che il genere comporti una significativa differenza negli stati D ed H, considerati i più affidabili indicatori d'età, rispettivamente per le soglie di 16 e 18 anni. Nel presente lavoro, come conseguenza di queste prime considerazioni e vista la composizione per età del campione esaminato (fra i 16 e i 22 anni) si è deciso di ricorrere alla valutazione dello sviluppo dentale dei soli terzi molari.

La possibilità che la maturazione dentale sia influenzata da eventuali patologie da cui il soggetto potrebbe essere affetto viene praticamente esclusa da Meinl et al. (2007). Ciò non impedisce la possibilità di prendere in esame altri fattori che possano incidere sulla crescita dei denti. A questo proposito l'AGFAD (*Study Group for Forensic Age Diagnostics*) e l'IOFOS (*International Organization for Forensic Odonto-Stomatology*) pongono delle linee guida per la qualità dei metodi di valutazione dentale asserendo che si debbano prendere in considerazione fattori quali lo stato di salute dell'individuo, il suo stato socio-economico e le origini genetico-geografiche.

Braga et al. (2005) indagano sugli standard geografico-specifici ritenuti influenti per la valutazione dell'età dentale, che si basa sull'usura del tessuto dentale e gengivale dovuta all'utilizzo. Gli autori giungono alla conclusione che l'etnia, la distribuzione dell'età cronologica e le caratteristiche della

regione di appartenenza non garantiscono un significativo miglioramento nelle previsioni sull'età.

Olze et al. (2004) analizzano il terzo molare mettendo a confronto tre etnie diverse: sudafricana, tedesca e giapponese. La popolazione sudafricana raggiunge gli stati della classificazione di Demirjian ad età inferiori rispetto a quelle della popolazione tedesca che, a sua volta, precede quella giapponese: tali differenze non permangono nello stato H, ritenuto un buon indicatore per il raggiungimento della soglia dei 18 anni. Gli autori concordano inoltre sul fatto che la maturità scheletrica è influenzata dallo stato socio-economico della popolazione in quanto una condizione socio-economica bassa ne ritarderebbe lo sviluppo. Infine, una maggiore dimensione del palato faciliterebbe l'eruzione del terzo molare mascellare (arcata superiore).

1.3 Confronto fra i metodi di valutazione dell'età basati su osservazioni dentali

Il metodo di Demirjian è ampiamente riconosciuto come uno dei più affidabili ed accurati metodi per la stima dell'età di soggetti viventi non adulti, garantendo un'elevata correlazione fra età cronologica ed età stimata. Inoltre, la classificazione a otto stati impiegata conduce a migliori valori di accordo fra più osservatori rispetto ad altri metodi utilizzati, come di seguito mostrato.

Olze et al. (2005) confrontano il metodo di Demirjian con altri quattro metodi basati sulla valutazione dello sviluppo dentale per i denti fino al secondo molare. La differenza sostanziale nelle classificazioni degli stati dentali sta nel fatto che, mentre la classificazione di Demirjian consiste nella mera descrizione morfologica del dente, le altre classificazioni si basano su frazioni di lunghezza future di corona e radice rispetto alla lunghezza finale. La critica spesso mossa contro queste classificazioni è l'incertezza nello stabilire la frazione di lunghezza raggiunta dalla radice o dalla corona quando in realtà non si conoscono ancora le lunghezze finali delle stesse. Tale critica è ancora più rilevante nel caso in cui si prendesse in considerazione lo sviluppo dentale del terzo molare la cui variabilità di forma e dimensione è maggiore rispetto agli altri denti.

Una volta attribuito lo stato dentale mediante una scala di classificazione che adotta un numero di categorie variabile dalle quattro categorie (Gustafson e Koch) alle 25 (Gleiser e Hunt), si ottiene l'età stimata impiegando opportune tavole di conversione. Utilizzando il metodo di Demirjian, Olze et al. (2005) ottengono la più alta correlazione fra età cronologica e stimata. Maber et al. (2006) affrontano il problema della stima d'età lavorando su tutti i denti fatta eccezione del terzo molare. Mettendo a confronto quattro differenti metodi di

classificazione dentale, gli autori arrivano alla conclusione che il più accurato, insieme a quello di Demirjian è il metodo di Willems, il quale utilizza i punteggi delle tavole di conversione di Demirjian opportunamente aggiustati.

Dhanjal et al. (2006) prendono in esame la riproducibilità di più osservatori nel valutare le stesse OPT in differenti istanti temporali, facendo uso di quattro diverse classificazioni dentali. L'indice di riproducibilità fra diversi osservatori sulle medesime OPT risulta più alto se si utilizza la classificazione di Demirjian, specie nel caso del terzo molare. Per quanto riguarda invece i restanti denti, si ottiene in generale una maggiore riproducibilità dei dati nei denti mandibolari (arcata inferiore) piuttosto che mascellari (arcata superiore).

La classificazione di Demirjian ha acquistato rapidamente notorietà prima ancora del metodo stesso di attribuzione dell'età mediante specifiche tavole di conversione. Essa è frequentemente utilizzata non solo per l'alto grado di accordo a cui può condurre ma anche per la chiarezza di valutazione ed il numero limitato di categorie che adotta. In realtà, a partire dalla classificazione di Demirjian, si sono poi utilizzati metodi di stima dell'età diversi da quelli che utilizzano le tavole di conversione, fra i quali modelli predittivi di regressione lineare (Pinchi et al., 2005; Cameriere et al., 2006; Cameriere et al., 2007) o regressione logistica (Cameriere et al., 2008).

Braga et al. (2005) propongono la classificazione per età mediante l'osservazione dei sette denti mandibolari applicata a tre campioni differenziati geograficamente: europeo, asiatico e africano. Le età dei soggetti, comprese fra i 4 e 16,5 anni, sono suddivise in 25 categorie di ampiezza pari a sei mesi. Gli autori mettono a confronto l'analisi delle corrispondenze combinata con la regressione (CAR) ed un modello di classificazione Bayesiana, mediante il quale l'assegnazione di un individuo ad una determinata classe d'età avviene in funzione della massimizzazione della probabilità condizionata a posteriori, rispetto agli stati dentali osservati. Qualora si introducesse nella classificazione Bayesiana la dipendenza fra gli stati dentali dei diversi denti, gli autori ottengono risultati più accurati rispetto al metodo CAR, vale a dire che le stime d'età così ottenute, per ciascuno dei tre campioni geografici, risultano meno distorte.

Cameriere et al. (2006) studiano modelli predettivi di regressione lineare che includono variabili esplicative relative all'apertura dell'apice dentale, che precede la nascita della radice. Inoltre gli autori verificano che l'osservazione delle OPT digitali non porta a differenze significative fra misurazioni dello stesso osservatore avvenute in tempi diversi, affermando che l'OPT digitale può essere utilizzata per produrre apprezzabili misure di riproducibilità intra-osservatore. Successivamente Cameriere et al. (2007) introducono nel modello di regressione lineare il fattore di nazionalità a livello europeo. L'incremento del coefficiente di determinazione lineare non risulta significativo, vale a dire la nazionalità non contribuisce in maniera significativa alla stima dell'età dei soggetti esaminati.

Recentemente Cameriere et al. (2008) hanno proposto un modello di classificazione per età utilizzando la regressione logistica dove la variabile risposta, che corrisponde alla probabilità che un individuo sia maggiorenne o minorenne, viene studiata in funzione di un indice di maturità dentale dei terzi molari mandibolari e del genere. Gli autori affermano che il genere non incrementa significativamente la *performance* del modello di classificazione basato sulla regressione logistica che include il solo indicatore di maturità dentale. Inoltre, viene analizzata a posteriori la sensibilità e specificità della classificazione prestando particolare attenzione ai falsi positivi (minorenni classificati erroneamente come maggiorenni), per i quali le conseguenze giuridiche hanno un peso non trascurabile.

Lucy et al. (2002) propongono di stimare la distribuzione per età di individui adulti mediante un modello di calibrazione, dove per *calibrazione* si intende l'operazione in cui uno strumento di misura viene regolato in modo da migliorarne l'accuratezza (Everitt, 2002, p.58). Utilizzando i metodi Bayesiani gli autori determinano la distribuzione a posteriori dell'età per effettuare stime puntuali e intervallari, a partire dall'osservazione di variabili ad essa associate. Fra queste, il *root dentine traslucency* (RDT), che è uno degli indicatori più affidabili per la valutazione dell'età dentale. Infine, confrontando i risultati con quelli derivanti dal corrispondente modello di regressione multipla, risulta che la calibrazione proposta conduce ad una maggiore accuratezza.

Riassumendo i lavori precedentemente citati, l'assegnazione dell'età per individui viventi non adulti si concretizza sostanzialmente in una stima d'età a seconda del genere del soggetto, del grado di maturazione dei denti esaminati e di altri fattori più o meno inclusi nel modello impiegato. La recente letteratura mostra come la classificazione di Demirjian sia quella più frequentemente utilizzata e ampiamente accettata. Questo è riconducibile non solo alla sua semplicità e chiarezza ma anche al fatto che essa fornisce indici di riproducibilità ed un coefficiente di correlazione fra età cronologica e stimata più elevati rispetto a quelli derivanti da altre classificazioni dentali. Sulla base della classificazione di Demirjian si ottengono successivamente le stime d'età attraverso delle specifiche tavole di conversione oppure facendo uso di modelli di regressione lineare o logistica.

1.4 Obiettivi della ricerca

L'obiettivo principale di questo lavoro è quello di classificare un soggetto per classe di età, dati il *Genere* e gli *Stati di Mineralizzazione Dentale* assegnati ai quattro terzi molari, e di associarvi una stima di probabilità che quantifichi l'incertezza dell'attribuzione. Il gruppo etnico di appartenenza non viene preso

in considerazione poiché i soggetti del campione esaminato sono tutti di etnia caucasica.

Una differenza sostanziale rispetto agli studi proposti fino ad oggi, fatta eccezione dei lavori di Braga et al. (2005) e Cameriere et al. (2008), è che nella presente ricerca si intende stimare la probabilità che un soggetto appartenga ad una particolare classe d'età anziché stimarne l'età stessa. Rispetto alla stima puntuale si è ritenuto più adeguato lo studio dell'assegnazione della classe d'età in termini di classificazione, proprio per il contesto legale all'interno del quale viene analizzato il fenomeno. Infatti, non bisogna dimenticare la natura forense del problema sottoposto ad analisi: non è importante conoscere l'età esatta del soggetto interessato, bensì sapere se questo abbia raggiunto o meno una determinata soglia d'età in funzione della quale garantire la corretta applicazione delle leggi.

In questo lavoro si è cercato, inoltre, di sottolineare l'importanza della procedura di stima basata sulle letture delle OPT che ciascun osservatore effettua. Questo significa che ogni modello deve essere stimato *ad personam*, vale a dire che la previsione della classe d'età dell'$(n+1)-esimo$ soggetto deve essere ottenuta impiegando il modello "tarato" su colui che ha compiuto la lettura. Ed è proprio grazie ad una riproducibilità inter-osservatori non perfetta che non si ammette, a differenza di molte metodologie adottate, l'esistenza della scambiabilità fra differenti osservatori.

Altro aspetto rilevante è il fatto che la valutazione fornita dall'esperto non è sempre certa. E' chiaro che la classificazione di Demirjian, per quanto semplice e funzionale, comporti delle attribuzioni degli stati dentali incerte. A questo proposito, una differenza sostanziale rispetto a tale classificazione è quella di introdurre l'incertezza circa l'attribuzione degli stati dentali, permettendo ad un osservatore di indicare due stati adiacenti, anziché uno soltanto, qualora lo reputasse opportuno. Tale valutazione corrisponde alla *soft evidence*, trattata nel prossimo paragrafo, e l'incertezza di attribuzione viene poi quantificata con opportuni "pesi probabilistici", detti *believes*.

La classificazione inizialmente proposta, coerentemente con la richiesta di discriminare un soggetto maggiorenne da uno minorenne, è dicotomica a seconda che l'individuo esaminato abbia raggiunto o meno la soglia dei 18 anni. Successivamente è stata effettuata un'ulteriore analisi per ridurre i problemi derivanti dall'adozione della soglia stessa.

Infatti, se dal punto di vista legale il passaggio da maggiorenni a minorenni è istantaneo ed avviene quando l'individuo ha superato il 365-esimo giorno del suo 17-esimo anno di età, lo sviluppo dentale è invece progressivo. Queste considerazioni hanno portato alla costruzione di una terza classe centrale, a cavallo della soglia puntuale dei 18 anni di età.

Inoltre si intende valutare i risultati ottenuti con il modello classificatorio in funzione dell'esperto che procede alla classificazione e della tecnologia radiografica che questi utilizza. A tal proposito è stata osservata la variabile *Tecnologia* della OPT, analogica o digitale, sottolineando come ogni esperto,

in sede processuale, possa ricorrere ad una determinata tipologia radiografica mediante la quale sia in grado di fornire maggiori garanzie di riproducibilità nelle valutazioni. La tecnologia non rientra, però, nella procedura di learning parametrico, vale a dire che le predittive sulla probabilità che un individuo appartenga ad una determinata classe d'età non sono state condizionate a tale variabile.

In conclusione, a partire dalla classificazione dello sviluppo dentale secondo la scala di Demirjian e dalla possibilità di assegnare per ciascun dente più di uno stato, l'osservatore valuterà le OPT del campione in oggetto. Queste valutazioni, congiuntamente al sesso dei soggetti esaminati, saranno necessarie alla stima dei parametri del modello di classificazione per classi d'età per ciascun esperto. Sarà quindi possibile verificare le prestazioni a posteriori del modello stesso anche in funzione della tecnologia utilizzata.

1.5 La soft evidence e i casi missing

Una volta che una variabile aleatoria si è realizzata, in genere è possibile conoscere con certezza con quale dei suoi stati si è manifestata. In talune circostanze tuttavia tale variabile non è osservabile con certezza. In questi casi l'osservatore riferisce circa lo stato o gli stati ritenuti possibili manifestazioni della variabile in oggetto, provvedendo altresì ad una valutazione di probabilità che ciascuno di essi si sia realizzato.

Shapiro (1977) sottolinea come in ambito medico un giudizio dicotomico del tipo "sano/malato" sia troppo limitato e tale fenomeno si può generalizzare al caso non dicotomico e a qualunque campo applicativo in cui si debbano fornire delle valutazioni, come nel caso presente. L'autore suggerisce, altresì, la possibilità di introdurre una valutazione probabilistica che possa migliorare l'abilità diagnostica di un medico o comunque dell'esperto chiamato in causa, ponendo così le basi per l'utilizzo della cosiddetta *soft evidence*.

A tal proposito viene introdotta una variabile aleatoria denominata *evidenza* (*evidence* in inglese), per la formalizzazione della quale si rimanda al *Cap. 3*. Per *evidence* si intende una variabile probabilisticamente connessa ad almeno un'altra di interesse tipicamente non osservata, in questo caso lo stato dentale (Pearl, 1988). Si deve, dunque, porre particolare attenzione nel differenziare i due concetti, da un lato, di evidenza legata alla "lettura" di una osservazione e, dall'altro, di variabile di interesse le cui realizzazioni non sono direttamente accessibili con l'osservazione. Si ha una *hard evidence* se si è certi riguardo lo stato assunto dalla variabile oggetto di interesse ed una *soft evidence* se si attribuisce una probabilità non nulla a più stati ritenuti possibili manifestazioni della variabile stessa.

Nella pratica forense non è infatti infrequente che un esperto esprima una valutazione che preveda più di uno stato della variabile di interesse. Da qui la necessità di introdurre tutte le forme di incertezza che potrebbero derivare dall'attività dello stesso.

Nel problema specifico, in alcuni casi l'osservatore si trova di fronte ad un'incerta assegnazione dello stato di appartenenza del dente, spesso legata alla complicanza di lettura della radiografia. L'utilizzo della *soft evidence* risponde dunque all'esigenza di formalizzare un'osservazione incerta che un esperto non è in grado di determinare con sicurezza. Questo ha condotto alla trattazione dell'osservazione dubbia mediante la *soft evidence*, trasformandola in un'attribuzione fra due stati adiacenti per i quali l'esperto ritiene si articoli l'incertezza.

Prima di portare a conclusione questo paragrafo è necessario specificare la terminologia adottata in letteratura, a partire da Rubin (1976), per quanto riguarda i dati a disposizione. La "lettura" di un'unità statistica che si concretizza in una delle due forme di evidenza introdotte, *hard* o *soft*, rappresenta un dato *osservato*, mentre nel caso in cui non si abbia alcuna forma di evidenza si parla di dato *mancante* o *missing data*. L'insieme formato sia dai dati osservati che da quelli mancanti costituisce l'insieme dei cosiddetti dati *completi*, ovvero quelli potenzialmente osservati.

L'autore considera due tipologie diverse di danti mancanti, quelli *Missing At Random* (MAR) e quelli *Missing Completely At Random* (MCAR). Per quanto riguarda i primi, il meccanismo generatore dei *missing data*, condizionatamente ai dati completi, non dipende dai dati mancanti. Questo equivale a dire che se per alcune unità statistiche tutte le variabili osservate assumono lo stesso valore allora anche le variabili non osservate hanno la stessa distribuzione statistica per tali unità. Nel caso di un processo generatore di *missing data* di tipo MCAR si ha l'indipendenza non solo dai dati mancanti ma anche da quelli osservati.

Nel presente lavoro i dati sono considerati MCAR ovvero si considera che l'assenza di un dato, che sia dovuta all'impossibilità di lettura della OPT o al dente fisicamente assente, è indipendente dall'assenza stessa e dai dati osservati condizionatamente all'età. Questo significa che nota l'età, il fatto che un dente non esista fisicamente oppure che l'OPT sia illeggibile è puramente un fenomeno casuale.

1.6 Riproducibilità e ripetibilità

La variabilità inter-osservatore è un fenomeno comune che si riscontra in tutti quei casi in cui siano richieste valutazioni a più osservatori circa fenomeni definiti su scala ordinale o nominale. In ambito medico, ad esempio, è

riconducibile a questa categoria la maggior parte delle indagini diagnostiche, le quali si basano sulla lettura di "oggetti" risultanti dall'elaborazione di strumentazioni apposite come le radiografie. Le valutazioni di differenti osservatori possono così differire in quanto ciascuna di esse è intrinsecamente legata all'esperienza precedentemente acquisita dall'esperto nella "lettura" del fenomeno (Sardanelli e Di Leo, 2008).

A questo punto della trattazione è d'uopo fare una precisazione sulla terminologia adottata, concordemente con quanto definito dall'organizzazione mondiale per la definizione di norme tecniche, l'ISO (*International Organization for Standardization*), che nel 1993 ha pubblicato la *Guide to the expression of uncertainty in measurement* (S.A.S.O, 2000). La *riproducibilità* viene definita come un indice di concordanza fra misurazioni effettuate sulle stesse unità sotto condizioni di misura diverse, in questo caso i due osservatori (Kotz e Johnson, 1985, p.378).

Fin dagli anni '60, quando si è presa coscienza dell'importanza dell'osservatore come fonte di variabilità non trascurabile, è nato l'interesse a misurare il grado di accordo fra osservatori differenti. Ciò comporta anche una riflessione sul concetto di scambiabilità non più fra osservazioni ma fra osservatori. E´ infatti importante sottolineare come la scambiabilità delle osservazioni all'interno di gruppi omogenei per sesso, dente ed eventualmente etnia e tecnologia della OPT, debba includere anche gli osservatori.

Questo tipo di assunzione è spesso implicita nelle analisi inferenziali in quanto si suppone scontata l'osservazione, vale a dire si assume che il "processo di lettura" delle variabili sia il medesimo per ciascun osservatore. Al contrario, poniamo attenzione alle letture effettuate da differenti strumentazioni che sospettiamo possano essere tarate diversamente. Appare evidente che il meccanismo inferenziale stimato facendo uso di una specifica apparecchiatura potrebbe non essere adatto ad effettuare previsioni nel caso se ne utilizzi una differente. Sembra perciò lecito analizzare le classificazioni che osservatori diversi possono effettuare e quindi chiedersi se un modello stimato sulla base delle osservazioni fornite da un esperto possa essere utilizzato da un altro le cui osservazioni differiscono dal primo in maniera cospicua. L'idea della costruzione di un modello *ad personam* deriva quindi dal fatto che la riproducibilità non è mai perfetta nella pratica. Tuttavia una bassa variabilità inter-osservatori fra due esperti potrebbe consentire l'utilizzazione dello stesso modello per entrambi.

Generalmente per misurare le discordanze fra due osservatori si ritiene che occorra calcolare indici di accordo depurati dall'effetto del caso, ovvero che tengano conto del fatto che alcuni degli accordi osservati potrebbero essere di natura casuale, ad esempio pensando che uno dei due osservatori tiri ad indovinare. Secondo Spitzer et al. (1967) una corrispondenza fra le valutazioni di due esperti significa che fra i loro pareri c'è associazione ma non necessariamente accordo. Non è infatti irrilevante la componente di accordo

casuale, la quale aumenta al diminuire del numero di categorie di assegnazione del fenomeno analizzato.

La variabilità inter-osservatori non è la sola sulla quale abbia senso indagare. Difatti, una volta separate le procedure di *learning* parametrico dei modelli per ciascun osservatore, si potrebbe includere nella "taratura" del modello stesso la variabilità intra-osservatore. Tale variabilità misura, nel tempo, quanto un osservatore sia in grado di fornire la medesima valutazione sulle stesse unità.

A questo proposito si definisce la *ripetibilità* come un indice di concordanza fra successive misurazione effettuate sulle stesse unità, sotto le medesime condizioni di misurare, vale a dire lo stesso osservatore (Kotz e Johnson, 1985, p.378). Tale indice è oggetto di interesse, come fonte di incertezza, per il calcolo delle predittive che includano modelli di transizione dell'evidenza e dei relativi *believes*, formalizzati nel *Cap. 3*.

La ripetibilità delle osservazioni è infatti importante per misurare la coerenza di un osservatore. Nel fornire le proprie valutazioni questi si avvale della propria esperienza circa la "lettura" dell'oggetto di interesse. Dunque la variabilità intra-osservatore può essere interpretata come una misura delle capacità valutative che l'osservatore ha appreso precedentemente e può risultare ragionevole includerla nel modello. Tale indice non vuole avere la pretesa di giudicare l'osservatore bensì è necessario per la costruzione di un modello più realistico, i cui risultati sono inevitabilmente legati alla "coerenza" dell'esperto che ha fornito le valutazioni. Proprio per questo si parla si modelli stimati *ad personam*.

Capitolo 2
Classificazione e modelli grafici: il Naive Bayes

2.1 Evoluzione dei metodi di apprendimento probabilistici

I metodi Bayesiani, che negli ultimi anni hanno riscontrato un crescente interesse, fanno risalire la loro origine al reverendo britannico Thomas Bayes (1701 – 1761). Noto matematico e famoso cultore della filosofia, Bayes deve la propria reputazione al teorema sulla probabilità condizionata, presente all'interno del suo articolo, *Essay towards solving a problem in the doctrine of chances*, pubblicato postumo nel 1763 (si veda Swinburne, 2002, p.122-149).

In forma sintetica si può riassumere il *Teorema di Bayes* nel seguente modo. Siano N eventi A_i una partizione dello spazio campionario Ω, vale a dire eventi tra loro incompatibili, $A_i \cap A_j = \varnothing$ per $\forall i, j$, ed esaustivi, $\bigcup_i A_i = \Omega$, e sia B un evento non nullo per il quale vale $P(B) > 0$. Si assuma di conoscere le probabilità marginali $P(A_i)$ e quelle condizionate $P(B \mid A_i)$ per $\forall i$. Allora la probabilità dell'*i-esimo* evento A_i dato l'evento B si ottiene mediante la cosiddetta *formula di Bayes*:

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{\sum_{j=1}^{N} P(A_j)P(B \mid A_j)} = P(A_i)\frac{P(B \mid A_i)}{P(B)}. \qquad (2.1)$$

Nell'ultima espressione della (2.1) è racchiusa l'idea intuitiva, ma di notevole interesse, che ha permesso di costruire le solide basi della statistica Bayesiana. La probabilità condizionata $P(A_i \mid B)$, detta *probabilità a posteriori*, si ricava a partire dalla marginale $P(A_i)$, nota come *probabilità a priori*, la quale viene "aggiornata" attraverso il rapporto $P(B \mid A_i)/P(B)$. Questo fattore trasforma la conoscenza iniziale (a priori) in una conoscenza finale (a posteriori) attraverso la probabilità condizionata $P(B \mid A_i)$, conosciuta in letteratura col nome di *verosimiglianza*.

Il Teorema di Bayes formalizza il concetto di *aggiornamento* della conoscenza probabilistica. Esso afferma, infatti, che la conoscenza a posteriori circa la realizzazione di un evento è proporzionale alla conoscenza a priori moltiplicata per la verosimiglianza:

$$P(a\ posteriori) \propto P(a\ priori) \times verosimiglianza\,. \tag{2.2}$$

La metodologia Bayesiana consente di fare previsioni su nuovi dati utilizzando la distribuzione a posteriori dei parametri ottenuta mediante il Teorema di Bayes. In breve, si supponga di avere un insieme di dati y e si introduca un modello di generazione dei dati, rappresentante la conoscenza probabilistica del fenomeno aleatorio analizzato, i cui parametri θ sono specificati da una distribuzione a priori $P(\theta)$. Per la (2.1) si ottiene la distribuzione a posteriori dei parametri che corrisponde ad un aggiornamento della distribuzione a priori ottenuto attraverso i dati:

$$P(\theta \mid y) = P(\theta)\frac{P(y \mid \theta)}{P(y)}\,. \tag{2.3}$$

Per prevedere nuovi dati $\hat{y}$ si procede quindi ad un'integrazione della distribuzione a posteriori (2.3), rispetto al parametro θ:

$$P(\hat{y} \mid y) = \int_{\theta} P(\hat{y} \mid \theta)P(\theta \mid y)d\theta\,. \tag{2.4}$$

Molti dei recenti sviluppi e progressi della statistica Bayesiana sono avvenuti in campo tecnologico, dove la velocità di elaborazione degli odierni *software* e la capienza delle nuove memorie *hardware* hanno permesso di processare una quantità di dati sempre più grande in tempi notevolmente ridotti. Un altro fattore che ha facilitato la diffusione dei metodi Bayesiani è stato lo sviluppo dei *metodi simulativi* che consistono nella generazione di valori di una variabile casuale il cui parametro segue una distribuzione probabilistica nota. E' stato così possibile riprodurre e risolvere numericamente problemi

riguardanti fenomeni condizionati da un alto numero di variabili casuali, le cui soluzioni analitiche erano fino ad allora troppo complesse o impossibili.

I miglioramenti tecnologici ed il crescente sviluppo di algoritmi applicati ai più svariati campi hanno consentito una rapida diffusione anche dell'*Intelligenza Artificiale*, termine coniato dallo scienziato informatico americano John McCarthy nel 1956. Nel 1990 Kurzweil (citato in Russel e Norvig, 2005, p.4) definisce l'intelligenza artificiale come "l'arte di creare macchine che eseguono attività che richiedono intelligenza quando vengono svolte da persone". Questa disciplina studia sistemi finalizzati a far svolgere al *computer* funzioni e ragionamenti tipici della mente umana. Si basa, dunque, sul concetto di *apprendimento* o *learning*, ovvero su di un processo di acquisizione di conoscenza, di competenze o particolari abilità attraverso l'esperienza.

Una delle aree di maggior sviluppo dell'Intelligenza Artificiale è il *Machine Learning*, che si occupa della realizzazione di sistemi informatici in grado di apprendere automaticamente dall'esperienza. Il lavoro di Samuel (1959), riguardante un programma in grado di giocare a dama, fu probabilmente il primo esempio di successo nel campo dell'apprendimento automatico che includeva le idee moderne sull'*apprendimento per rinforzo*. Con questo tipo di apprendimento l'algoritmo, ogniqualvolta produce un risultato, riceve un *feedback* di "rinforzo" che indica quanto sia buono o meno il risultato prodotto. Lo scopo dell'apprendimento per rinforzo è quello di generare azioni ottimali, vale a dire che massimizzano lo *score* totale risultante dai *feedbacks*, premi e penalizzazioni ricevute.

Fra le tecniche dalle quali il *Machine Learning* deve il recente sviluppo, si distingue il *Data Mining*, che ha come oggetto l'estrazione di informazioni ridotte da grandi quantità di dati e l'esplorazione di questi allo scopo di scoprirne schemi significativi, detti *pattern*. Il *Data Mining*, che nasce negli anni '90, deve la rapida ascesa a vari fattori di progresso tecnologico come la modalità di diffusione dei dati: non è infatti infrequente che al giorno d'oggi si possa accedere a masse rilevanti di dati rese disponibili sul *web* e tali da ridurre costi e tempi di indagine.

Gli argomenti di ricerca di cui si occupa il *Data Mining* spaziano dal *Market Analysis*, che si interessa alle strategie e politiche di mercato – Hu et al. (2000) propongono un algoritmo il cui obiettivo è ricercare gli insiemi di oggetti più frequentemente acquistati dai clienti nel fare la spesa – alla *Fraud Detection*, che studia le evasioni fiscali, i comportamenti fraudolenti negli acquisti on-line e le transazioni bancarie avvenute con carte di credito – Brause et al. (1999) sfruttano le reti neurali (si veda più avanti) per analizzare gli aspetti comuni nelle transazioni con carta di credito, considerando solo previsioni con un grado di accuratezza superiore al 99,9%, vista la bassa proporzione di fraudolenza riscontrata.

Molti problemi reali richiedono procedure risolutive con un continuo ed elevato grado di flessibilità che difficilmente è traducibile in codici informatici.

L'obiettivo del *Machine Learning* è proprio quello di migliorare la capacità di esecuzione automatica di un determinato compito, da parte di un computer, attraverso l'esperienza. Il concetto chiave è quello di apprendimento, definito da Mitchell (1997, p.2) nel seguente modo: "un programma di computer si dice che *apprende* dall'esperienza E rispetto ad una classe di compiti T e con una misura di *performance* P, se la sua *performance* per i compiti T, misurata da P, migliora con l'esperienza E".

Un altro campo in cui trovano ampio spazio le applicazioni del *Machine Learning* è quello delle *Reti Neurali*. Il termine biologico viene utilizzato anche in matematica applicata in relazione alle *Reti Neurali Artificiali*, vale a dire modelli matematici che rappresentano le interconnessioni fra costrutti, i cosiddetti *neuroni artificiali*, che in qualche misura imitano le proprietà dei neuroni viventi. Da quando nel 1943 il neurofisiologo e cibernetico Warren McCulloch ed il matematico Walter Pitts presentarono il primo modello di rete neurale si è stati spettatori in una notevole crescita della conoscenza sul sistema nervoso umano, le cui organizzazioni gerarchiche si sono rilevate utili ai fini risolutivi di molti problemi pratici.

Le reti neurali artificiali sono dunque dei sistemi di elaborazione dell'informazione che simulano il funzionamento del sistema nervoso nell'essere umano, il quale è costituito da un elevato numero di neuroni collegati fra di loro in una complessa rete. Il comportamento intelligente delle reti neurali deriva dalle numerose interazioni tra le unità, che sono raggruppate in strati. Alcuni strati ricevono informazioni dall'ambiente (*input*), altri emettono risposte nell'ambiente (*output*) ed altri ancora comunicano con le unità all'interno della rete. Si parla di *perceptrone* quando la rete neurale è formata da due strati, uno di ingresso e uno di uscita, e di *multistrato* nel caso in cui siano presenti strati di unità nascoste. Ciascun neurone svolge un'operazione che consiste nel diventare attivo, emettendo un segnale che viene trasmesso lungo i canali di comunicazione della rete, se la quantità totale di segnale ricevuto supera una cerca soglia. Tale azione avviene in parallelo con quella degli altri neuroni consentendo di trattare molte informazioni contemporaneamente.

Le reti neurali artificiali sono dunque uno strumento estremamente efficace nell'analisi di situazioni non predicibili analiticamente e si prestano con grande flessibilità alla modellizzazione di problemi di varia natura. Il legame input-output, ovvero la funzione di trasferimento del segnale (informazione) nella rete, non viene programmato ma è ottenuto attraverso un processo di apprendimento. A questo punto si introducono le forme di apprendimento, *supervisionato* e *non supervisionato*, che si utilizzano nel *Machine Learning* e nelle Reti Neurali.

Nel *supervised learning* si cerca di generalizzare l'informazione ottenuta dal campione, contenente sia i dati di input che di output, per costruire un modello che aiuti la "macchina" a prendere una decisione sulla base di nuovi dati di

ingresso e di una regola decisionale stabilita a priori. Sinteticamente, poste $X_1,...,X_N$ le variabili di input e Y quella di output, gli algoritmi nell'apprendimento supervisionato lavorano con la probabilità congiunta $P(X_1,...,X_N,Y)$. Con questo metodo si mira ad istruire un sistema informatico in modo da consentirgli di risolvere automaticamente dei compiti, fornendogli degli esempi. La costruzione del modello si basa, quindi, su un campione di "sostegno", il *training data set*, attraverso il quale la macchina si addestra e impara i legami esistenti fra le variabili di input e quelle di output.

La riuscita degli algoritmi supervisionati dipende dalla quantità dei dati di ingresso: troppi rallenterebbero l'algoritmo mentre pochi potrebbero portare ad un apprendimento scarso e per alcuni casi anche assente e quindi la "macchina" non risulterebbe idonea a dare valutazioni o fornire previsioni. Nel caso di variabili categoriali, il *training data set* dovrà quindi contenere un numero ragionevole di casi per ogni possibile elemento del cartesiano delle variabili di input esaminate. Infine, per valutare la bontà del processo di apprendimento si utilizza il *test data set*, composto da unità differenti da quelle presenti nel *training data set*, con il quale si testa il modello appreso verificando se ci sono riscontri fra i dati osservati e quelli previsti.

Un esempio di *supervised learning* è l'*Analisi Discriminante*, tecnica di analisi multidimensionale dei dati che nacque nel 1936 con l'articolo *The use of multiple measurements in taxonomic problems* (Fisher, 1936) del noto statistico Ronald Aylmer Fisher (1890 – 1962). Apparve così, per la prima volta, il metodo di analisi discriminante lineare per la classificazione di primati attraverso alcune misurazioni effettuate su reperti fossili. Il termine Analisi Discriminante racchiude quell'insieme di tecniche di classificazione che misurano l'importanza dei fattori nel determinare l'appartenenza di un'unità ad un gruppo specifico. Generalmente, si perviene ad una regola decisionale che è funzione di un numero limitato di variabili risultate significativamente discriminanti e tramite la quale si è in grado di attribuire un soggetto ad uno soltanto dei gruppi predefiniti.

Nell'*unsupervised learning*, invece, mancano le variabili di output e conseguentemente il *training data set*, vale a dire l'esempio attraverso il quale la "macchina" si addestra e impara. Gli algoritmi basati sull'apprendimento non supervisionato lavorano dunque sulla distribuzione di probabilità congiunta $P(X_1,...,X_N)$, che risulta diversa rispetto al caso supervisionato. Attraverso la stima della densità delle variabili di input si possono studiare i casi anomali come, ad esempio, quelli analizzati nelle transazioni fraudolente con carta di credito di cui si sono occupati, come già accennato, Brause et al. (1999). Rientra in quest'ottica la *Cluster Analysis* con la quale si cerca di ripartire i dati in un numero limitato di gruppi fra loro disgiunti, i cosiddetti *clusters*. Nel 1930 lo psicologo e statistico statunitense Robert Choate Tryon sperimentò una tecnica di *clustering* nel campo delle scienze sociali e

comportamentali ma si dovette aspettare fino agli inizi degli anni ‘60, con l’avvento dei moderni *computers*, perché la *Cluster Analysis* cominciasse ad essere sviluppata ed applicata nei più molteplici campi. E’ in definitiva una tecnica di riduzione dei dati che raggruppa le unità statistiche in base a misure di similarità. L’organizzazione dei dati nei gruppi è conseguita minimizzando le misure di associazione fra le unità contenute in un *cluster* e massimizzando quelle fra unità appartenenti a *clusters* diversi.

Si conclude questa parte introduttiva nel citare alcune applicazioni del *Machine Learning* e delle Reti Neurali che si sono sviluppate negli ultimi anni. I lavori spaziano dai programmi, come menzionato in precedenza, per i giochi da scacchiera – Tesauro (1995) analizza un programma chiamato TD-Gammon il cui algoritmo di apprendimento per rinforzo affronta la *Temporal Difference* nella ricezione dei premi e delle penalizzazioni, i quali subiscono un ritardo e vengono assegnati alla fine di un catena di output – a quelli per il riconoscimento della scrittura. Della scrittura dei numeri, utile per il riconoscimento di assegni bancari e codici postali, se ne sono occupati Ouchtati et al. (2007): gli autori presentano la lettura automatica di una sequenza di cifre dicotomizzate (bianco e nero) attraverso la tecnica della *Backpropagation* o propagazione all’indietro dell’errore. Tale propagazione avviene mediante un algoritmo di apprendimento supervisionato utilizzato comunemente nelle reti neurali e attraverso il quale si aggiornano i coefficienti di connessione della rete, compresi quelli che arrivano agli strati nascosti. Infine, si giunge a programmi che consentono un’autonomia di guida per veicoli su strada – Pomerleau (1989) ha ideato ALVINN (*Autonomous Land Vehicle In a Neural Network*), un sistema che sfrutta la *backpropagation* all’interno di una rete neurale a tre strati e che apprende la guida associando all’immagine della strada opportunamente rielaborata le decisioni prese dal guidatore.

I numerosi successi ottenuti nello studio dell’apprendimento umano, da un lato, e gli enormi miglioramenti *software* e *hardware* conseguiti in ambito tecnologico, dall’altro, hanno dato origine ad un’interessante interconnessione, sia dal punto di vista scientifico che filosofico, fra l’uomo e la “macchina”. E’ quindi nelle tecniche di apprendimento che risiede la fiducia di molti ricercatori e studiosi ai fini di risolvere problemi di carattere pratico la cui difficoltà computazionale era considerata irrisolvibile fino a qualche anno prima. Inoltre, l’apprendimento automatico risulta necessario laddove mancano esperti “umani” o, addirittura, le conoscenze e tecniche da questi utilizzate sono difficili da formalizzare.

2.2 La classificazione

Nella pratica odierna non è infrequente affrontare problemi in cui si debba classificare una o più unità statistiche, come nel *Machine Learning* o nel *Data Mining*, ovvero "attribuire" ad uno stato di una variabile categorica l'unità stessa, dove il termine "attribuzione" va inteso in termini probabilistici.
Ai fini di una formalizzazione del concetto di classificazione si introduca la variabile discreta C, detta *variabile di classe*, il vettore di *N* variabili discrete $\mathbf{X} = (X_1, ..., X_N)$, note col nome di *attributi*, ed un *training set* formato da *n* osservazioni $\{x_1^i, ..., x_N^i, c^i\}$ per ciascuna *i-esima* unità.

Si consideri adesso l'$(n+1)$-*esimo* caso per il quale siano osservati gli attributi $\mathbf{x}^{n+1} = \{x_1^{n+1}, ..., x_N^{n+1}\}$ ma non la classe c^{n+1} di appartenenza. L'obiettivo è dunque quello di costruire una regola di classificazione, ovvero una funzione f, chiamata *classificatore*, che associ ad ogni osservazione *i* una ed una sola classe c^i, dato l'insieme delle realizzazioni dei corrispondenti attributi $\mathbf{x}^i$.

Un classificatore semplice e ampiamente diffuso è basato sulla massimizzazione della probabilità di appartenere ad una classe condizionatamente agli attributi, vale a dire, posta la notazione semplificata $\mathbf{x}^{n+1}$ per $\mathbf{X}^{n+1} = \mathbf{x}^{n+1}$, ed in modo simile per le altre variabili:

$$\hat{c}^{n+1} = f(\mathbf{x}^{n+1}) = \underset{c}{\operatorname{argmax}} P(c \mid \mathbf{x}^{n+1}). \tag{2.5}$$

Nella pratica si ricorre alla sua formulazione alternativa mediante il teorema della probabilità condizionata di Bayes (2.1):

$$\hat{c}^{n+1} = \underset{c}{\operatorname{argmax}} \left[\frac{P(c) P\left(\mathbf{x}^{n+1} \mid c\right)}{P\left(\mathbf{x}^{n+1}\right)} \right] = \underset{c}{\operatorname{argmax}} \left[P(c) P\left(\mathbf{x}^{n+1} \mid c\right) \right]. \tag{2.6}$$

Sia $\boldsymbol{\gamma}$ il vettore dei parametri della distribuzione della variabile di classe *C* e $\boldsymbol{\theta}_c$ l'insieme dei parametri della distribuzione congiunta degli attributi condizionatamente a $C = c$, $\mathbf{X} \mid c$. Inoltre, senza perdita di generalità, siano $\gamma_c = P(c)$ e $\theta_c^{n+1} = P\left(\mathbf{x}^{n+1} \mid c\right)$. Allora una forma compatta per la (2.6) è:

$$\hat{c}^{n+1} = \underset{c}{\operatorname{argmax}} \left(\gamma_c \theta_c^{n+1} \right). \tag{2.7}$$

Il problema della classificazione dell'$(n+1)$-*esimo* caso richiede quindi la conoscenza dei parametri presenti nella (2.7). Ma, come spesso accade, tali parametri sono incogniti e per questo si ricorre alle corrispondenti stime.

La stima di θ_c^{n+1} può presentare delle problematicità dovute al fenomeno della *iperparametrizzazione*, vale a dire quando le osservazioni non sono sufficientemente numerose rispetto ai parametri. Infatti, può accadere che una determinata combinazione di attributi $\mathbf{x} = (x_1, ..., x_N)$, condizionatamente alla classe c, non sia osservata per alcuna unità. Quindi la stima di massima verosimiglianza della corrispondente probabilità condizionata risulterebbe nulla.

Precisamente, se si indica con $n_{\mathbf{x}|c}$ la frequenza congiunta associata ad una specifica combinazione $(\mathbf{x}, c) = (x_1, ..., x_N, c)$ e con n_c la frequenza marginale per $C = c$, allora la stima di massima verosimiglianza di θ_c^{n+1}, basata sulle *n* osservazioni del *training set*, si ottiene dal seguente rapporto di frequenze:

$$\hat{\theta}_c^{n+1} = \frac{n_{\mathbf{x}|c}}{n_c}. \tag{2.8}$$

Allo stesso modo si ricava la stima del parametro γ_c:

$$\hat{\gamma}_c = \frac{n_c}{n}. \tag{2.9}$$

Nel caso in cui l'osservazione $(\mathbf{x}, c)$ non sia presente nel campione, $n_{\mathbf{x}|c} = 0$, la stima di probabilità secondo la (2.8) risulterebbe nulla e questo risultato potrebbe far erroneamente pensare che l'evento $\mathbf{x}\,|\,c$ sia impossibile.

Per tentare di risolvere questo inconveniente si può introdurre un'ipotesi semplificatrice, assumendo l'indipendenza degli attributi condizionatamente alla variabile di classe, vale a dire $X_i \perp X_j \,|\, C = c,\ \forall i \neq j$. Tale assunzione comporta la fattorizzazione della probabilità condizionata $P(\mathbf{X}\,|\,C)$:

$$P(\mathbf{x}\,|\,c) = \prod_{h=1}^{N} P(x_h\,|\,c), \tag{2.10}$$

riducendo notevolmente il numero di stime dei parametri necessari al calcolo della (2.7). Infatti, sia $\boldsymbol{\theta}_{h|c}$ il vettore parametrico che rappresenta la distribuzione del singolo attributo X_h condizionatamente alla classe c e si ponga senza perdita di generalità $\theta_{h|c}^{n+1} = P(x_h^{n+1}\,|\,c)$. Allora il classificatore (2.7) si può riscrivere come:

$$\hat{c}^{n+1} = \underset{c}{\operatorname{argmax}}\left(\gamma_c \prod_{h=1}^{N} \theta_{h|c}^{n+1} \right). \tag{2.11}$$

Quest'ultima formulazione consente la riduzione dello spazio parametrico. Precisamente, sia k_h il numero degli stati possibili con i quali la variabile X_h può manifestarsi e $k_h - 1$ il numero di parametri che ne caratterizzano la corrispondente distribuzione di probabilità. Allora l'ipotesi (2.10) comporta che il numero totale dei parametri nel caso congiunto, che corrisponde a $\prod_{h=1}^{N} k_h - 1$, si riduca a $\sum_{h=1}^{N} (k_h - 1)$. Se disponessimo di *N* attributi a *k* stati allora i parametri da stimare sarebbero $N(k-1)$ anziché $k^N - 1$.

Il vantaggio dell'assunzione di indipendenza condizionata degli attributi rispetto alla variabile di classe non si limita alla sola riduzione dello spazio parametrico ma si riscontra anche nell'applicabilità delle stime stesse. Difatti, in un campione sufficientemente numeroso la stima $\hat{\theta}_{h|c}^{n+1}$ difficilmente conduce a valori nulli a differenza di quanto potrebbe accadere per la stima del parametro congiunto $\hat{\theta}_c^{n+1}$, secondo la (2.8).

In alternativa al classificatore (2.5) si potrebbe utilizzare una regola di classificazione f_π, la quale introduce un vincolo π sulla probabilità di classificazione:

$$\hat{c}^{n+1} = f_\pi(x^{n+1}) = \underset{c}{\operatorname{argmax}}\left(\gamma_c \prod_{h=1}^{N} \theta_{h|c}^{n+1} \right)$$

$$\text{se e solo se } \gamma_c \prod_{h=1}^{N} \theta_{h|c}^{n+1} \geq \pi\,. \tag{2.12}$$

Pertanto il classificatore (2.12) potrebbe non essere in grado di assegnare una classe a tutte le unità statistiche se il vincolo non venisse soddisfatto. Infine, si osservi come, nel caso la variabile di classe *C* sia dicotomica, una soglia probabilistica $\pi = 0.50$ nella (2.12) comporterebbe gli stessi risultati di classificazione della regola (2.5).

2.3 Il Naive Bayes

La classificazione è una procedura decisionale che comporta l'attribuzione di un'unità statistica ad una classe, vale a dire deciderne la collocazione, in funzione di una regola decisionale rappresentata dal classificatore. Fra i

classificatori comunemente impiegati nelle varie discipline, quello che utilizza la probabilità fattorizzata nella (2.10) è noto in letteratura col nome di *Naive Bayesian Classifier*, o semplicemente *Naive Bayes*, per la sua semplicità derivante dall'ipotesi di indipendenza condizionata degli attributi rispetto alla variabile di classe (Duda e Hart, 1973; Langley et al., 1992). A dispetto di questa semplificatrice assunzione il *Naive Bayes* sorprende per la sua efficacia e competitività rispetto ad altri classificatori (Friedman e Goldszmidt, 1996; Friedman et al., 1997).

Friedman (1997) sottolinea la differenza fra l'accuratezza di un classificatore, che consiste nella valutazione a posteriori della percentuale di corretta classificazione osservata sul *test data set*, e l'accuratezza di stima, misurata come errore quadratico medio della differenza fra valore vero e predetto sui vari *training sets*. Secondo Domingos e Pazzani (1996) il segreto della *performance* del *Naive Bayes* è racchiuso nel fatto che esso può ottenere una migliore accuratezza di classificazione anche quando le stime di probabilità contengono errori maggiori rispetto ad altri classificatori. Gli autori, come altri fra cui Rish (2001), si interrogano sull'impatto che le dipendenze degli attributi esercitano sulla *performance* del *Naive Bayes* e misurano, per ciascuna coppia di attributi, la dipendenza condizionata rispetto alla variabile di classe. L'analisi effettuata, oltre a verificare una miglior *performance* del *Naive Bayes* anche laddove sussistono chiari legami di dipendenza, mette in evidenza una bassa correlazione fra dipendenza media degli attributi e la differente accuratezza fra *Naive Bayes* e gli altri classificatori. Quindi gli autori concludono che la dipendenza fra gli attributi non sembra influenzare l'accuratezza di un classificatore.

Friedman (1997) mostra come, a causa di un *trade-off* fra la distorsione e la varianza della stima, queste due componenti agiscono con intensità differenti sulla misura di accuratezza di stima e su quella di classificazione. Dunque una più accurata stima di probabilità non necessariamente porta ad una migliore *performance* di classificazione. Viceversa, piccoli errori di classificazione non implicano che le corrispondenti probabilità siano state stimate in maniera accurata. Questo spiega come mai un'alta distorsione di stima del *Naive Bayes*, mitigata da una bassa variabilità, possa condurre comunque ad un'elevata accuratezza di classificazione, rendendolo competitivo rispetto ad altri classificatori i cui errori di stima sono più accettabili.

Domingos e Pazzani (1997) scrivono che sotto la funzione di perdita 0-1, detta anche *misclassification rate*, il *Naive Bayes* risulta ottimale se vale l'indipendenza condizionata degli attributi rispetto alla variabile di classe. Nei casi in cui non valessero tali indipendenze presupponendo quindi una struttura di dipendenze meno vincolata e più adattabile ai problemi reali, tale classificatore rimarrebbe comunque ottimale sotto certe condizioni[1]. Infatti, la

[1] Gli autori presentano un semplice esempio con una variabile di classe dicotomica e tre attributi, due dei quali perfettamente dipendenti mentre il terzo indipendente dagli altri due. La regione di assegnazione di un'unità ad una delle due classi, secondo il *Naive Bayes*, rimane

distorsione della stima incide maggiormente nell'accuratezza di stima piuttosto che in quella di classificazione mentre, al contrario, la varianza della stima esercita un'influenza maggiore sull'accuratezza di classificazione, specie nei campioni di piccole dimensioni. Gli autori propongono condizioni necessarie e sufficienti di ottimalità per il *Naive Bayes* con il proposito di approfondirle affinché possano essere efficientemente verificate nella pratica.

Per concludere questa parte sul classificatore *Naive Bayes* si citano alcuni campi e corrispondenti applicazioni in cui il *Naive Bayes* è utilizzato con risultati soddisfacenti. In ambito informatico Hellerstein et al. (2000) affrontano il problema della qualità dei servizi telematici, valutando i tempi di trasmissione dell'informazione, mentre Kim e Hwang (2008) propongono un approccio automatico per scovare messaggi indesiderati, generalmente commerciali e noti col termine di *spam*, che oltre a contenere informazioni pubblicitarie non richieste possono alterare i risultati di ricerca. In campo epidemiologico Geenen et al. (2004) studiano la febbre suina cercando di discriminare quali animali ne sono affetti, mentre per quanto riguarda la classificazioni del testo Kim et al. (2003) propongono un classificatore *Naive Bayes* dove le distribuzioni di probabilità condizionate degli attributi rispetto alla variabile di classe sono delle Poisson invece delle classiche Multinomiali.

2.4 Rappresentazione di un sistema probabilistico tramite modelli grafici

Il classificatore *Naive Bayes* basa la propria fortuna sull'ipotesi di indipendenza condizionata sebbene si sia accennato al mantenimento della sua efficacia anche nel caso tale assunzione non sia più vera. Si è quindi interessati a trovare uno strumento che sia idoneo a derivare la struttura di dipendenza sottostante le variabili analizzate nel caso si dovesse ricorrere a classificatori basati su strutture di dipendenza più complesse.

Le *reti Bayesiane* rispondono a questa esigenza in quanto costituiscono uno strumento appropriato ed efficace per rappresentare e manipolare i legami di dipendenza fra i caratteri. Nelle reti Bayesiane si combinano elementi di *Teoria dei Grafi*, di seguito accennata, e di *Teoria della Probabilità*, la quale

---

prossima a quella ottimale (Domingos e Pazzani, 1997, *Fig. 1*, p.10), sottolineando l'elevata accuratezza di questo classificatore anche in circostanze in cui sono presenti chiari legami di dipendenza fra attributi. Più in generale, ma sempre considerando una variabile di classe dicotomica, si osserva la regione di ottimalità del *Naive Bayes* (Domingos e Pazzani, 1997, *Fig. 2*, p.13) coerentemente con il *Corollario 1* al *Teorema 1* (Domingos e Pazzani, 1997, p.12).

permette di quantificare l'incertezza attraverso le tavole di probabilità condizionate (CPT).

2.4.1 Cenni sulla Teoria dei Grafi

Un *grafo* è una coppia $G=(V,E)$ dove *V* è un insieme finito di *vertici* o *nodi*, che corrispondono agli oggetti di analisi, mentre $E \subseteq V \times V$ è un sottoinsieme costituito da *archi* caratterizzanti le relazioni binarie fra coppie ordinate di vertici (Cowell et al., 1999). Se $(\alpha,\beta) \in E$ e $(\beta,\alpha) \in E$ allora si dice che la coppia ordinata (α,β) è un arco *indiretto* o *non orientato* e i due nodi sono detti *vicini*. Altrimenti, se $(\alpha,\beta) \in E$ e $(\beta,\alpha) \notin E$, l'arco è *diretto* o *orientato* e i nodi α e β sono detti rispettivamente *genitore* e *figlio*. Se fra due nodi esiste un arco, diretto o indiretto, i due nodi sono detti *uniti*. Il grafo è quindi una struttura idonea per rappresentare relazioni binarie fra caratteri sia nel caso di associazione (arco indiretto) sia nel caso di causazione (arco diretto).

Dato un grafo *G*, $\widehat{G}$ è la *versione indiretta* o *scheletro* di *G* dove gli archi diretti sono sostituiti con i corrispondenti indiretti. Si parla di *sottografo* $G_A=(A,E_A)$ di $G=(V,E)$ se $A \subseteq V$ e $E_A \subseteq E \cap (A \times A)$. Se vale $E_A = E \cap (A \times A)$ allora G_A viene detto sottografo *indotto* da A ed è caratterizzato da tutti gli archi di E che interessano i nodi inclusi in A.

Gli insiemi che si vanno introducendo hanno nomenclatura derivante dai relativi termini inglesi: l'insieme costituito dai vicini di β si indica con $ne(\beta)$, l'insieme dei genitori e dei figli sono rappresentati rispettivamente da $pa(\beta)$ e $ch(\beta)$, il *Markov Blanket* $BL(\beta)$ è costituito dai vertici genitori, figli e genitori dei figli, $BL(\beta) = pa(\beta) \cup ch(\beta) \cup pa(ch(\beta))$, mentre per *intorno* $bd(\beta)$ si intende l'estensione dell'insieme dei vicini $ne(\beta)$ a quello dei genitori (ma non ai figli), $bd(\beta) = ne(\beta) \cup pa(\beta)$.

Un *cammino* di lunghezza *n* da α a β è una sequenza di vertici $\{\alpha = v_0, v_1, ..., v_n = \beta\}$ a due a due uniti e tali che $\forall i \in \{1,2,...,n\}$ $(v_{i-1}, v_i) \in E$. Un cammino si dice diretto se è costituito da almeno un arco diretto. Se esiste un cammino che conduce da α a β ed un cammino, non necessariamente distinto dal primo, da β ad α, allora i due nodi si dicono *connessi*. La relazione di connessione induce la *classe di equivalenza* $[\beta]$, dove $\alpha \in [\beta]$ significa che α e β sono connessi. Una classe di equivalenza, che può essere costituita anche da un solo nodo, è detta *componente forte* del grafo.

Un grafo è detto *completo* se ogni coppia di vertici che lo costituisce è unita. Un sottoinsieme di vertici è completo se induce un sottografo completo. Qualora aggiungendo al sottografo completo un nodo del grafo iniziale questi

non risultasse più completo, allora si parlerebbe di sottografo *completo minimale* o *clique*.

Un *ciclo* di ordine *n* è un cammino di lunghezza *n* dove il primo e l'ultimo nodo coincidono, $\{\alpha = v_0, v_1, ..., v_n = \alpha\}$. Un grafo è detto *aciclico* se non contiene alcun ciclo, *a catena* se non contiene cicli diretti. Fra i grafi a catena ricopre un ruolo fondamentale il *grafo diretto aciclico* (DAG), ovvero un grafo aciclico costituito da soli archi orientati. Un grafo che non ha cicli diretti, che siano grafi indiretti o DAGs, è detto *grafo a catena*.

In un DAG D si definisce $an(\beta)$ l'insieme degli *antenati* di β, ovvero quei vertici che attraverso un cammino conducono a β ma non il viceversa, mentre l'insieme dei *discendenti* $de(\beta)$ è costituito dai vertici raggiungibili da β ma non il viceversa. Un insieme A è detto *ancestrale* se qualunque vertice di A ha un intorno incluso in A stesso, ovvero $\forall \beta \in A \quad bd(\beta) \subseteq A$ ed il corrispondente grafo indotto G_A è chiamato *grafo ancestrale*.

Un grafo $G = (V, E)$, costituito da $|V| = N$ nodi e $|E| = m$ archi, si dice *albero* se soddisfa le seguenti condizioni, fra loro equivalenti:

1. G è aciclico e connesso;
2. G è aciclico e $m = N - 1$;
3. G è connesso e $m = N - 1$;
4. $\forall \alpha, \beta \in V$ esiste un unico cammino che li congiunge.

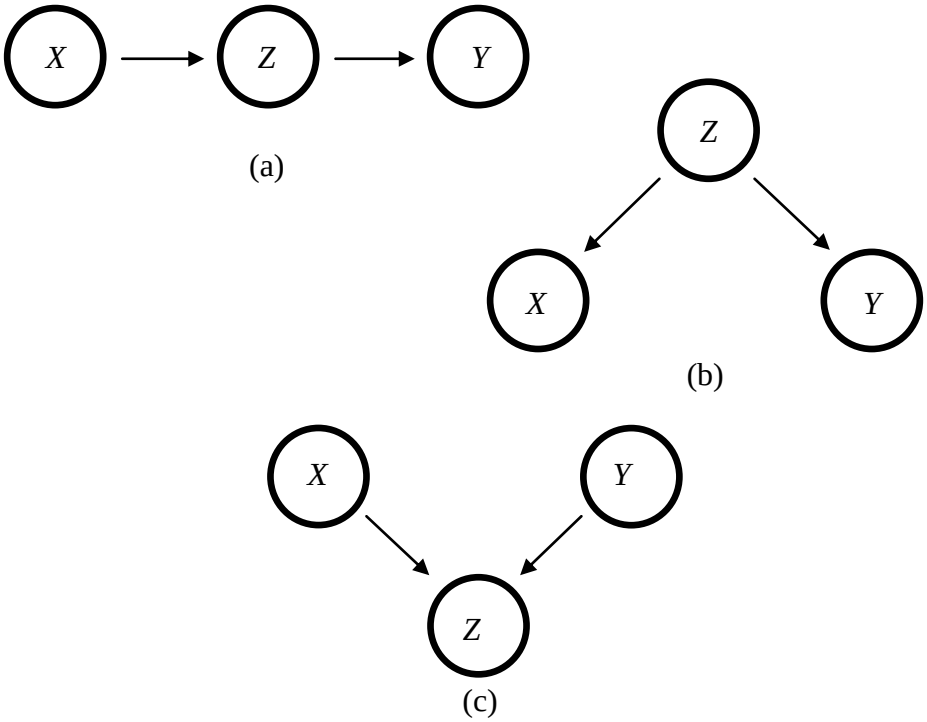
Un albero, utilizzato per rappresentare relazioni gerarchiche fra oggetti, è quindi costituito da nodi che hanno un solo genitore ed almeno un figlio: uniche eccezioni sono la *radice* che corrisponde ad un nodo senza genitori e le *foglie* che rappresentano nodi privi di figli.

Un *polialbero* è invece un albero i cui nodi possono avere più di un genitore mentre un grafo aciclico non connesso caratterizzato dall'unione di più alberi o polialberi viene detto *foresta*.

2.4.2 Grafi di indipendenza

Si considerino adesso grafi orientati aciclici costituiti da nodi rappresentanti caratteri discreti i cui archi orientati corrispondono alle relazioni causali. Jensen et al. (2007) definiscono l'*evidenza* di una variabile l'affermazione circa gli stati della variabile stessa: *hard evidence* se la variabile è *istanziata*, vale a dire è noto lo stato con cui si è manifestata, *soft evidence* in caso contrario.
In un DAG D l'evidenza si può trasmettere attraverso le variabili a seconda della struttura relazionale definita dagli archi orientati. Le differenti strutture di base sono mostrate nella seguente figura:

Figura 2.1. *Strutture di base dei DAG: a) seriale, b) convergente e c) divergente*



Con riferimento alla *Fig. 2.1*, il caso a) rappresenta una struttura *seriale* in cui la variabile *X* causa *Z*, che a sua volta causa *Y*. L'evidenza fra *X* e *Y* viene così trasmessa attraverso *Z* se questa non è istanziata. Altrimenti, noto lo stato con cui *Z* si è manifestata, le variabili *X* e *Y* sarebbero fra loro indipendenti in quanto le informazioni sui relativi stati dipenderebbero esclusivamente dallo stato di Z indipendentemente da quello manifestatosi per l'altra variabile, in altre parole $X \perp Y \mid Z$.

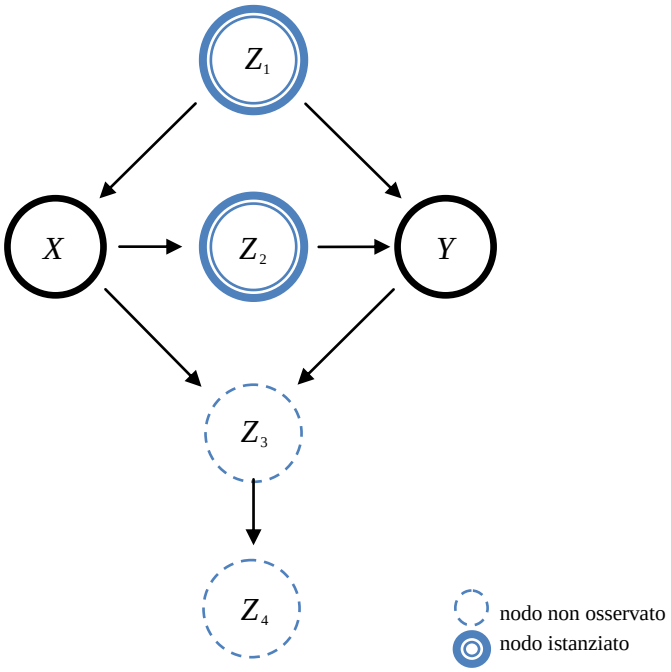
Il caso b), in cui ricade anche il *Naive Bayes*, presenta una struttura *divergente* dove la variabile *Z* è la causa comune di *X* e *Y*. La trasmissione dell'evidenza fra *X* e *Y* passa attraverso *Z* nel caso in cui tale variabile non sia nota. Se invece *Z* fosse istanziata allora le variabili *X* e *Y* sarebbero indipendenti fra di loro.

Infine, il caso c) tratta una struttura *convergente* dove *Z* è causata sia da *X* che da *Y*. Diversamente dai due casi precedenti, se *Z* non fosse nota non sarebbe possibile trasferire l'evidenza da *X* a *Y* e viceversa. Perché ciò avvenga la variabile *Z*, oppure uno dei suoi discendenti, deve essere istanziata.

La descrizione delle precedenti strutture consente di introdurre un concetto determinante per la definizione di Rete Bayesiana. Dato un DAG D, due variabili distinte *X* e *Y* (si veda la *Fig. 2.2*) sono *d-separate* se in ogni cammino fra X e Y esiste una variabile intermedia Z tale che si verifichi una delle seguenti condizioni:

1. Z è istanziata ed è in connessione divergente (Z_1) o seriale (Z_2);
2. Z e le sue discendenti non sono istanziate e Z è in connessione convergente (Z_3 e Z_4).

Figura 2.2. *Rappresentazione grafica del concetto di d-separazione*



Da quanto detto risulta che il *Markov Blanket* $BL(X)$ d-separa ciascuna variabile X dalle rimanenti. Simile, la definizione in un grafo indiretto: due variabili *X* e *Y* si dicono *separate* da una terza variabile Z se ogni cammino fra *X* e *Y* passa necessariamente da Z .

Nota la struttura del DAG, il concetto di d-separazione permette di identificare le indipendenze condizionate esistenti fra le variabili. Per verificare se due variabili *X* e *Y* sono d-separate da una terza variabile *Z*, oppure da un insieme di variabili, si prende in considerazione il grafo ancestrale indotto da $\{X,Y,Z\}$. Si costruisce successivamente il corrispondente *grafo morale*, vale a dire lo scheletro del grafo derivante dall'aggiunta di archi fra genitori con figli in comune. Si può dunque affermare che *X* e *Y* sono d-separate da *Z* se nel grafo morale ottenuto tutti i cammini fra *X* e *Y* devono necessariamente passare da *Z*.

Si conclude questa parte sulle indipendenze condizionate citando le tre *independence Markov properties*, che Whittaker (1990) dimostra essere equivalenti, e ricordando che l'indipendenza condizionata $X \perp Y \mid Z$, significa

che, se *Z* fosse istanziata, da *Y* non si potrebbero estrarre ulteriori informazioni su *X* e viceversa (si vedano i casi a) e b) della *Fig. 2.1*). Tali proprietà, applicabili ai grafi non orientati, si possono estendere anche ai grafi orientati: un DAG possiede le proprietà di indipendenza di Markov di cui gode il corrispondente grafo morale. Dunque, senza perdite di generalità ed indicando con Z_{rest} l'insieme delle variabili al di fuori di quelle citate nella relazione di indipendenza condizionata, le suddette proprietà affermano che:

Pairwise Markov property. Due variabili non adiacenti *X* e *Y* sono indipendenti condizionatamente ad una delle restanti variabili:

$$X \perp Y \mid Z_{rest}. \tag{i}$$

Local Markov property. Una variabile *X*, condizionatamente al proprio intorno, è indipendente da qualunque altra restante variabile:

$$X \perp Z_{rest} \mid bd(X). \tag{ii}$$

Global Markov property. Due variabili *X* e *Y*, o insiemi disgiunti di variabili, sono indipendenti condizionatamente ad una terza variabile Z_{sep}, o insieme disgiunto dai due precedenti, che separa le due variabili:

$$X \perp Y \mid Z_{sep}. \tag{iii}$$

2.4.3 Reti Bayesiane

Un DAG D è detto *I-map* di un modello di dipendenza *M* se ogni condizione di d-separazione mostrata in D corrisponde ad una relazione di indipendenza condizionale in *M*. D si dice *I-map minimale* di *M* se, comunque si elimini una delle direzioni degli archi di D, questi perde la proprietà di essere *I-map* di *M* (Jensen et al., 2007).

Data una distribuzione di probabilità *P* su un set di variabili *V*, un DAG $D = (V, E)$ è chiamato *Rete Bayesiana* di *P* se e solo se è *I-map* minimale di *P*. Una Rete Bayesiana è dunque uno strumento adatto a rappresentare efficacemente la distribuzione congiunta di probabilità di un insieme di variabili discrete, la cui struttura di dipendenza è caratterizzata da un DAG. Infatti, dalla definizione di Rete Bayesiana discende la fattorizzazione della probabilità congiunta $P(\mathbf{X}) = P(X_1, ..., X_N)$ in:

$$P(\mathbf{X}) = \prod_{i=1}^{N} P(X_i \mid pa(X_i)). \tag{2.13}$$

In genere non esiste una sola struttura di indipendenza che codifica la distribuzione di probabilità congiunta di interesse, ma se ne possono utilizzare altre appartenenti alla stessa classe di equivalenza. A tal proposito si definisce *v-struttura* di un DAG una terna (X,Y,Z) in connessione convergente su Z (si veda il caso c) della *Fig. 2.1*) nella quale i due nodi X e Y non sono connessi. Dunque due DAG D_1 e D_2 sono detti *d-equivalenti* se hanno lo stesso scheletro e le stesse v-strutture.

Nelle Reti Bayesiane si combinano così elementi della Teoria dei Grafi e della Teoria della Probabilità. Quest'ultima permette di quantificare l'incertezza delle variabili discrete rappresentate dai nodi della rete, attraverso le *tavole di probabilità condizionate* (CPT), necessarie per la trasmissione dell'evidenza nota in letteratura sotto il nome di *propagazione* (Pearl, 1982).

Affinché questo processo abbia inizio le CPT devono essere *inizializzate*, vale a dire si devono stimare i parametri che le caratterizzano mediante metodi di apprendimento: il numero dei parametri per ciascuna CPT cresce in maniera esponenziale col numero di genitori del nodo interessato. Il processo di trasmissione dell'evidenza consiste nel calcolo delle probabilità marginali e successivamente nell'aggiornare le probabilità condizionate delle variabili in accordo con l'evidenza a disposizione[2].

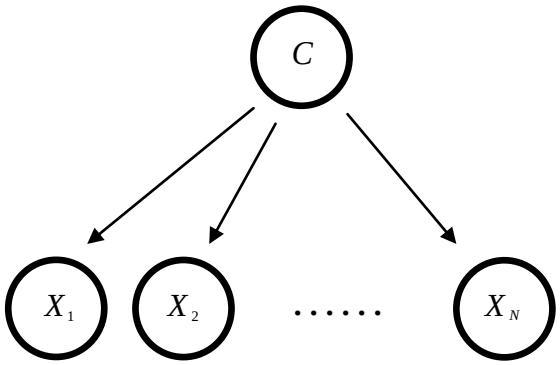
La Rete Bayesiana corrispondente al *Naive Bayes*, nota come *Naive Bayesian Network*, presenta una struttura semplificata (si veda la *Fig. 2.3*) dovuta all'ipotesi di indipendenza degli attributi condizionatamente alla variabile di classe (genitore di ogni attributo), vale a dire:

$$P(X_i \mid C, X_j) = P(X_i \mid C),\ \forall i \neq j\,, \tag{2.14}$$

che equivale alla pairwise *Markov property* (i).

[2]Nel caso di reti Bayesiane con struttura di dipendenza ad albero, l'algoritmo di propagazione dell'evidenza comunemente utilizzato è il *Junction Tree* (Cowell et al., 1999; Jensen et al., 2007), il quale aggrega i nodi della rete in *cliques* per poi impiegare la tecnica di propagazione proposta da Pearl nel 1982, il *Message Passing*, e basata sul reciproco scambio di informazioni fra nodi.

Figura 2.3. *Rappresentazione grafica del Naive Bayesian Network*



Il *Naive Bayesian Network* dovrebbe essere preso in maggiore considerazione non solo per la sua efficacia rappresentativa ma anche per la velocità computazionale sia in termini di inizializzazione delle CPT che di propagazione dell'evidenza. Inoltre, a partire da questa struttura di indipendenza semplificata, si può ricorrere a classificatori più complessi.

2.5 Un'estensione del Naive Bayes: il TAN

Friedman e Goldszmidt (1996) e Friedman et al. (1997) affrontano i limiti del *Naive Bayes* imposti dal vincolo sulle indipendenze condizionate (2.14). Gli autori, al fine di migliorare la *performance* del *Naive Bayes*, introducono la possibilità che ciascun attributo possa avere come secondo genitore, oltre alla variabile di classe, un altro attributo. Questo nuovo classificatore, denominato TAN (*Tree Augmented Naive Bayesian Classifier*), rispetto al *Naive Bayes* mantiene la semplicità computazionale e la robustezza degli errori nelle stime di probabilità dovuti a campioni di piccola dimensione riducendo, allo stesso tempo, la distorsione di stima.

Pazzani (1995) studia in che modo le dipendenze fra attributi possano esser misurate affinché si costruisca un classificatore più accurato. L'algoritmo impiegato è il BSEJ (*Backward Sequential Elimination and Joining*), il quale valuta ad ogni passo se ciascun attributo debba essere eliminato oppure aggregato ad un altro generando così un nuovo attributo. Alla fine della fase di eliminazione-congiunzione l'algoritmo stabilisce se il nuovo classificatore è significativamente più accurato di quello precedente fino a raggiungere una struttura di dipendenze condizionate ottimale. L'autore afferma che l'algoritmo BSEJ è in grado di costruire un classificatore che, rispetto al *Naive Bayes*,

produce una migliore accuratezza di classificazione, pur mantenendo quelle caratteristiche di semplicità e competitività che lo hanno reso famoso.

Infine, Zhang (2004) sostiene che utilizzando una funzione di perdita 0-1, il cosiddetto *misclassification rate* che definisce l'errore come numero di incorrette classificazioni (Domingos e Pazzani, 1997), si ottiene una miglior *performance* del *Naive Bayes*. L'autore sostiene, infatti, che tale funzione di perdita non penalizza, a differenza dell'errore quadratico medio, la scarsa accuratezza delle stime di probabilità. Egli fornisce allora una nuova spiegazione alla sorprendente *performance* del *Naive Bayes* affermando che essa non è influenzata dalle indipendenze condizionate bensì deriva dalle distribuzioni delle stesse.

Zhang (2004) prende in esame una variabile di classe dicotomica ed un classificatore ANB (*Augmented Naive Bayesian Classifier*), che equivale al TAN senza il limite sul numero di attributi genitori. L'autore dimostra che la classificazione che ne deriva si ottiene, per ciascuna delle classi della variabile di classe, dalle distribuzioni delle dipendenze locali degli attributi, dove per locale si intende il condizionamento ai nodi genitori.

Capitolo 3
Learning parametrico

3.1 Introduzione

Nel *Cap*. *1* si è descritto il problema di carattere forense affrontato in questo lavoro, vale a dire l'assegnazione di individui viventi non adulti a classi di età di interesse, e più in generale gli obiettivi di ricerca. Nel *Cap*. *2* si sono dunque introdotti i concetti teorici di apprendimento e classificazione necessari per la trattazione del problema, i quali verranno formalizzati in questo capitolo.

Inizialmente si introdurranno le variabili discrete oggetto di studio: la variabile di classe *C*, che corrisponde alla classe di età misurata in anni compiuti, l'attributo *X* caratterizzante lo sviluppo dentale del terzo molare, l'insieme **S** di covariate che influenzano maggiormente lo sviluppo dentale, e l'evidenza *E* necessaria a formalizzare l'incertezza sulle osservazioni dentali. Successivamente verranno formulate delle ipotesi sulle distribuzioni delle variabili e sui legami di dipendenza condizionata esistenti. Verrà poi esplicitata la verosimiglianza dei parametri caratterizzanti lo sviluppo dentale condizionatamente alla variabile di classe e all'insieme delle covariate **S**, ottenendone così la corrispondente distribuzione a posteriori in forma chiusa. Questo permetterà di calcolare la predittiva della variabile classe di età, condizionatamente ad **S** ed alla valutazione dentale, conseguendo l'obiettivo di classificare un soggetto in una determinata classe di età mediante l'impiego delle regole decisionali (2.11) o (2.12).

La parte conclusiva del capitolo presenterà alcuni indici di *performance* per la valutazione delle prestazioni dei modelli stimati per i singoli esperti, ed una proposta per la valutazione di una predittiva che tenga conto della ripetibilità di un osservatore.

3.2 Variabile di classe e attributi osservabili con incertezza mediante uso di soft evidence

Siano *T* la variabile discreta che rappresenta l'età di un individuo in anni compiuti e $\{\tau_1, \tau_2, ..., \tau_{Q-1}\} \subset \mathbb{N}$ l'insieme delle soglie di età di interesse. Allora si costruisca la variabile casuale di classe $C = \{c_1, c_2, ..., c_Q\}$, le cui *Q* classi sono definite come $c_i = \{t : \tau_{i-1} \leq t < \tau_i\}$, dove $\tau_0 = 0$ e $\tau_Q = \infty$.

Si considerino ora *N* attributi *k*-dimensionali della variabile di classe *C* corrispondenti agli sviluppi dentali dei terzi molari, $\mathbf{X}_h = (X_{h1}, X_{h2}, ..., X_{hk})$, dove X_{hj} rappresenta una variabile casuale di Bernoulli indicante se si è verificato o meno lo stato *j* per ciascun *h-esimo* attributo. Poiché la variabile $\mathbf{X}_h$ si può manifestare attraverso uno solo dei *k* stati, allora la corrispondente distribuzione di probabilità equivale ad una multinomiale per una singola osservazione $Mu_k(\mathbf{x}_h \mid \boldsymbol{\theta}_h, 1)$ di parametri $\boldsymbol{\theta}_h = (\theta_{h1}, \theta_{h2}, ..., \theta_{hk})$, dove $0 \leq \theta_{hj} \leq 1$ e $\sum_{j=1}^{k} \theta_{hj} = 1$ per $\forall h$ con $1 \leq h \leq N$.

Il vettore parametrico $\boldsymbol{\theta}_h$ rappresenta quindi la probabilità associata a ciascun stato di sviluppo dell'*h-esimo* terzo molare secondo la classificazione di Demirjian (*Fig. 1.2*) e si assume abbia a priori una distribuzione Dirichlet $Dir_k(\boldsymbol{\theta}_h \mid \boldsymbol{\alpha}_h)$ con iperparametri noti $\boldsymbol{\alpha}_h = (\alpha_{h1}, \alpha_{h2}, ..., \alpha_{hk})$. Sia infine $\mathbf{S}$ l'insieme delle covariate che influenzano gli attributi $\mathbf{X}_h$.

La natura dell'attributo $\mathbf{X}_h$ rende l'osservazione di tale variabile affetta da incertezza e questo comporta l'impiego della *soft evidence* introdotta nel *Cap. 1*. Sia dunque $\mathbf{E}_h = (E_{h1}, E_{h2}, ..., E_{hk})$ una variabile casuale discreta *k*-dimensionale associata al corrispondente attributo $\mathbf{X}_h$ e dove E_{hj} rappresenta una variabile casuale di Bernoulli che indica se lo stato *j* dell'*h-esimo* attributo è ritenuto una *possibile* manifestazione. Questo significa che $1 \leq \sum_{j=1}^{k} E_{hj} \leq k$ e, di conseguenza, la variabile $\mathbf{E}_h$ non è distribuita come una multinomiale.

L'evidenza $\mathbf{E}_h$ è caratterizzata dai *believes* $\mathbf{b}_h = (b_{h1}, b_{h2}, ..., b_{hk})$, dove $0 \leq b_{hj} \leq 1$ e $\sum_{j=1}^{k} b_{hj} = 1$ per $\forall h$ e che l'osservatore attribuisce ai *j-esimi* stati come possibili manifestazioni dell'attributo $\mathbf{X}_h$. Il vettore dei *believes* $\mathbf{b}_h$ rappresenta quindi il grado di fiducia che l'osservatore associa alla possibile realizzazione dell'attributo $\mathbf{X}_h$ per ciascuno dei *k* stati.

Il caso in cui l'osservatore indica un solo stato j^* come possibile realizzazione dell'attributo $\mathbf{X}_h$ escludendo tutti gli altri, equivale ad

un'osservazione diretta sulla variabile stessa. Questo tipo di evidenza è nota come *hard evidence* e si ha che se $\exists! j^* : b_{hj^*} = 1$ allora $b_{hj} = 0 \quad \forall j \neq j^*$ e $\sum_{j=1}^{k} E_{hj} = 1$. Se invece l'osservatore ritiene che più di uno stato dell'attributo $\mathbf{X}_h$ sia plausibile, assegnando corrispondenti *believes* non nulli, si parla di *soft evidence* e $1 < \sum_{j=1}^{k} E_{hj} \leq k$. Più in generale ci riferiremo alla variabile $\mathbf{E}_h$ con il termine di *soft evidence.*

Inoltre, si assume che il dato mancante, dovuto all'assenza del terzo molare oppure ad impedimenti nella lettura della OPT, non comporta alcuna valutazione da parte dell'osservatore nemmeno sfruttando le informazioni derivanti dai restanti attributi osservati sul medesimo individuo. L'imputazione è che l'osservatore si trovi in una situazione di massima incertezza in cui non è in grado di assegnare una maggior fiducia ad uno stato piuttosto che ad un altro e questo si concretizza nel considerare tutti gli stati dell'attributo $\mathbf{X}_h$ equamente possibili. Dunque i dati mancanti vengono trattati come casi particolari della *soft evidence* in cui la fiducia dell'osservatore si ripartisce equamente fra i *believes*, $\sum_{j=1}^{k} E_{hj} = k$ e $b_{hj} = 1/k$.

3.3 Assunzioni del Naive Bayes modificato

Nel *Cap. 2* si è visto come l'assunzione di indipendenza condizionata di *N* attributi, rispetto ad una variabile di classe *C*, equivalga alla costruzione del *Naive Bayes*. Il classificatore impiegato in questo lavoro, che equivale ad un'estensione del *Naive Bayes*, deve tener conto di due nuovi gruppi di variabili: le covariate e le *soft evidence*. Le covariate contenute in $\mathbf{S}$ si considerano sempre osservate e tutte le stime dei parametri del modello sono condizionate ad esse. In aggiunta alle normali ipotesi del *Naive Bayes* verranno di seguito presentate altre assunzioni che possano includere sia le covariate che le *soft evidence*.

Sia $\mathbf{X} = (\mathbf{X}_1, ..., \mathbf{X}_N)$ la congiunta degli *N* attributi e senza perdita di generalità sia $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N)$ la matrice dei parametri che caratterizzano tale distribuzione. Supponendo che gli attributi sono osservati, allora una naturale fattorizzazione di $P(\mathbf{X}, C, \mathbf{S} \mid \boldsymbol{\Theta})$ è:

$$P(\mathbf{X}, C, \mathbf{S} \mid \boldsymbol{\Theta}) = P(\mathbf{X} \mid C, \mathbf{S}, \boldsymbol{\Theta}) P(C, \mathbf{S}). \tag{3.1}$$

La probabilità congiunta $P(C,\mathbf{S})$ può essere riscritta come $P(C)P(\mathbf{S})$ nel caso di indipendenza fra la variabile di classe C e l'insieme delle covariate $\mathbf{S}$ oppure fattorizzata naturalmente senza imporre l'indipendenza:

$$P(C,\mathbf{S}) = P(C \mid \mathbf{S})P(\mathbf{S}). \tag{3.2}$$

E' ovvio che il caso di indipendenza rientra nella (3.2) se $P(C \mid \mathbf{S}) = P(C)$.

Adesso si introduca l'assunto caratterizzante il *Naive Bayes* di indipendenza condizionata degli attributi rispetto alla variabile di classe C con l'estensione del condizionamento anche all'insieme di covariate $\mathbf{S}$, vale a dire $\forall i \neq j \;\; \mathbf{X}_i \perp \mathbf{X}_j \mid C,\mathbf{S},\boldsymbol{\Theta}$:

$$P(\mathbf{X} \mid C,\mathbf{S},\boldsymbol{\Theta}) = \prod_{h=1}^{N} P(\mathbf{X}_h \mid C,\mathbf{S},\boldsymbol{\theta}_h). \tag{3.3}$$

Questa assunzione permette di stimare separatamente i vettori parametrici $\boldsymbol{\theta}_h$, condizionatamente alla classe C e all'insieme $\mathbf{S}$.

Nel caso in cui la natura degli attributi non permetta l'osservazione diretta sugli stessi, allora si deve introdurre nella (3.3) la *soft evidence*. Dunque, posta $\mathbf{B} = [b_{hj}]_{Nxk}$ la matrice dei *believes* forniti dall'osservatore per gli N attributi, estendiamo la (3.3) considerando $(\mathbf{X}_i,\mathbf{E}_i) \perp (\mathbf{X}_j,\mathbf{E}_j) \mid C,\mathbf{S},\boldsymbol{\Theta},\mathbf{B}$ $\forall i \neq j$, e quindi:

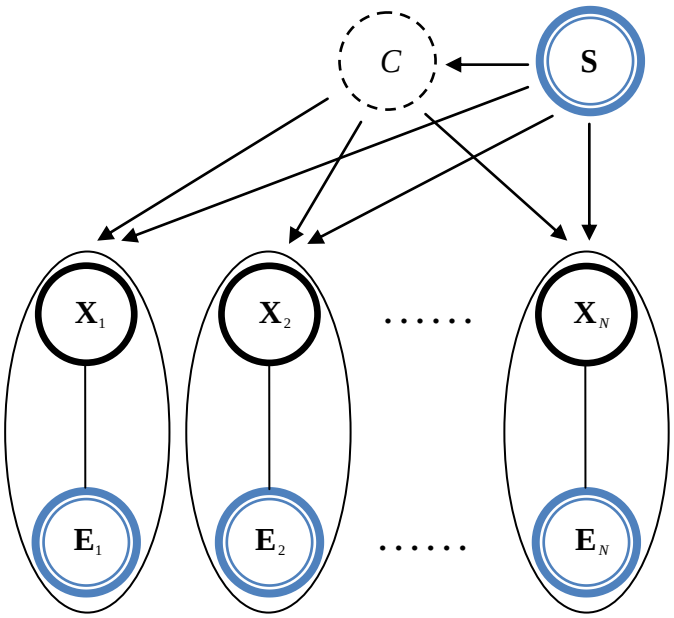
$$P(\mathbf{X},\mathbf{E} \mid C,\mathbf{S},\boldsymbol{\Theta},\mathbf{B}) = \prod_{h=1}^{N} P(\mathbf{X}_h,\mathbf{E}_h \mid C,\mathbf{S},\boldsymbol{\theta}_h,\mathbf{b}_h). \tag{3.4}$$

Considerando congiuntamente la (3.2) e la (3.4) si ha:

$$P(\mathbf{X},\mathbf{E},C,\mathbf{S} \mid \boldsymbol{\Theta},\mathbf{B}) = \prod_{h=1}^{N} P(\mathbf{X}_h,\mathbf{E}_h \mid C,\mathbf{S},\boldsymbol{\theta}_h,\mathbf{b}_h)P(C \mid \mathbf{S})P(\mathbf{S}). \tag{3.5}$$

Rispetto al *Naive Bayes*, l'estensione del classificatore (3.5) consiste nella introduzione dell'insieme delle covariate $\mathbf{S}$ e delle *soft evidence* $\mathbf{E}_h$. La struttura di dipendenza può essere visualizzata attraverso il grafo della *Fig. 3.1*:

Figura 3.1. *Grafo a catena della struttura di dipendenza del classificatore Naive Bayes modificato*



Rispetto alla *Fig. 2.3*, dove il *Naive Bayes* è rappresentato mediante un DAG, la *Fig. 3.1* raffigura il *Naive Bayes* modificato con l'aggiunta del nodo $\mathbf{S}$, caratterizzante l'insieme delle covariate, e i nodi delle *soft evidence* $\mathbf{E}_h$. Il grafo che ne deriva è un grafo a catena per la presenza di archi indiretti fra le coppie di nodi attributo $\mathbf{X}_h$ ed i corrispondenti nodi evidenza $\mathbf{E}_h$. Le coppie di nodi $(\mathbf{X}_h, \mathbf{E}_h)$ possono essere considerate singoli nodi (rappresentati dagli ovali) e quindi la *Fig. 3.1* si differenza dalla *Fig. 2.3*, per la sola presenza del nodo delle covariate $\mathbf{S}$. Attraverso questa struttura di dipendenza si vuole così trasmettere l'informazione osservata nelle *soft evidence* $\mathbf{E}_h$ e in $\mathbf{S}$ (nodi in blu) fino alla variabile di classe C non osservata.

Si indichi con $\mathbf{X}_{h|qs} \,|\, \boldsymbol{\theta}_{h|qs}$ la variabile casuale condizionata $\mathbf{X}_h \,|\, C, \mathbf{S}, \boldsymbol{\theta}_h$, che corrisponde ad una multinomiale su una singola osservazione, $Mu_k\left(\boldsymbol{\theta}_{h|qs}, 1\right)$, condizionatamente a tutte le classi di C, indicizzate da q, e a ciascuna combinazione degli stati delle covariate presenti in $\mathbf{S}$, indicizzate da s. Per poter effettuare la classificazione si deve dunque fare inferenza sui vettori parametrici $\boldsymbol{\theta}_{h|qs}$ attraverso i corrispondenti *believes* condizionati $\mathbf{b}_{h|qs}$.

Il legame probabilistico esistente fra l'attributo $\mathbf{X}_{h|qs}$ e la relativa *soft evidence* $\mathbf{E}_{h|qs}$ si può esprimere come:

$$P(\mathbf{X}_{h|qs}, \mathbf{E}_{h|qs} \,|\, \boldsymbol{\theta}_{h|qs}, \mathbf{b}_{h|qs}) = P(\mathbf{X}_{h|qs} \,|\, \mathbf{E}_{h|qs}, \boldsymbol{\theta}_{h|qs}) P(\mathbf{E}_{h|qs} \,|\, \mathbf{b}_{h|qs}). \qquad (3.6)$$

Posta:

$$P(\mathbf{X}_{h|qs} \mid \mathbf{E}_{h|qs}, \boldsymbol{\theta}_{h|qs}) \propto \boldsymbol{\theta}_{h|qs}\mathbf{E}_{h|qs}, \tag{3.7}$$

dove $\boldsymbol{\theta}_{h|qs}\mathbf{E}_{h|qs}$ è un vettore *k*-dimensionale il cui *j-esimo* elemento corrisponde al prodotto $\theta_{hj|qs}E_{hj|qs}$, allora:

$$P(\mathbf{X}_{h|qs}, \mathbf{E}_{h|qs} \mid \boldsymbol{\theta}_{h|qs}, \mathbf{b}_{h|qs}) \propto \boldsymbol{\theta}_{h|qs}\mathbf{b}_{h|qs}. \tag{3.8}$$

La (3.7) equivale ad affermare che vi è una probabilità positiva che uno stato dell'attributo $\mathbf{X}_{h|qs}$ si sia verificato solo se su di esso è stata riscontrata un'evidenza. Invece dalla (3.8) il vettore dei *believes* appare come un indicatore degli stati che non si sono verificati mentre per i restanti stati la probabilità è proporzionale al *belief* stesso moltiplicato al corrispondente parametro $\theta_{hj|qs}$.

Infine, si assume l'indipendenza delle distribuzioni a priori dei vettori parametrici condizionati $\boldsymbol{\theta}_{h|qs}$:

$$f(\boldsymbol{\theta}_{1|qs}, ..., \boldsymbol{\theta}_{N|qs}) = \prod_{h=1}^{N} f(\boldsymbol{\theta}_{h|qs}). \tag{3.9}$$

3.4 Verosimiglianza a struttura polinomiale

Da ora in avanti si prendano in considerazione gli attributi e relativi parametri, le *soft evidence* e corrispondenti *believes*, condizionati alla variabile di classe *C* e all'insieme delle covariate $\mathbf{S}$ mediante gli indici, rispettivamente, *q* ed *s*.

La funzione di verosimiglianza $L(\boldsymbol{\theta}_{h|qs}; \mathbf{b}_{i,h|qs})$ nella variabile $\boldsymbol{\theta}_{h|qs}$, nota la valutazione dell'osservatore $\mathbf{b}_{i,h|qs}$ per l'*i-esima* osservazione, si ottiene marginalizzando rispetto ad $\mathbf{X}_{h|qs}$ la congiunta espressa dalla (3.8):

$$L(\boldsymbol{\theta}_{h|qs}; \mathbf{b}_{i,h|qs}) = \sum_{j=1}^{k} P(X_{hj|qs}, \mathbf{E}_{h|qs} \mid \boldsymbol{\theta}_{h|qs}, \mathbf{b}_{i,h|qs}) = \sum_{j=1}^{k} b_{i,hj|qs}\theta_{hj|qs}. \tag{3.10}$$

Si osservi che nel caso l'osservazione sia *missing* allora la verosimiglianza (3.10) si riduce a $1/k$.

Sia $\mathbf{B}_{h|qs} = [b_{i,hj|qs}]_{n_{qs}xk}$ la matrice con n_{qs} righe che corrispondono ai *believes* condizionati $\mathbf{b}_{i,h|qs}$, dove $\sum_q \sum_s n_{qs} = n$:

$$\mathbf{B}_{h|qs} = \begin{bmatrix} b_{1,h1|qs} & b_{1,h2|qs} & \dots & b_{1,hk|qs} \\ b_{2,h1|qs} & b_{2,h2|qs} & \dots & b_{2,hk|qs} \\ \dots & \dots & \dots & \dots \\ b_{n_{qs},h1|qs} & b_{n_{qs},h2|qs} & \dots & b_{n_{qs},hk|qs} \end{bmatrix}, \tag{3.11}$$

allora la verosimiglianza per n_{qs} osservazioni condizionatamente indipendenti assume la forma di un prodotto di polinomi:

$$L(\boldsymbol{\theta}_{h|qs}; \mathbf{B}_{h|qs}) = \prod_{i=1}^{n_{qs}} L(\boldsymbol{\theta}_{h|qs}; \mathbf{b}_{i,h|qs}) = \prod_{i=1}^{n_{qs}} \sum_{j=1}^{k} b_{i,hj|qs} \theta_{hj|qs} \ . \tag{3.12}$$

Un modo compatto per rappresentare la verosimiglianza (3.12) consiste nel prendere in considerazione la matrice che raccoglie sulla diagonale gli elementi del vettore parametrico $\boldsymbol{\theta}_{h|qs}$:

$$diag(\boldsymbol{\theta}_{h|qs}) = \begin{bmatrix} \theta_{h1|qs} & 0 & \dots & 0 \\ 0 & \theta_{h2|qs} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \theta_{hk|qs} \end{bmatrix}. \tag{3.13}$$

Sia dunque $\mathbf{L}_{h|qs} = [b_{i,hj|qs}\theta_{hj|qs}]_{n_{qs}xk}$ la matrice le cui righe rappresentano, per ciascuna *i-esima* osservazione, gli elementi della sommatoria (3.10):

$$\mathbf{L}_{h|qs} = \mathbf{B}_{h|qs} \cdot diag(\boldsymbol{\theta}_{h|qs}) = \begin{bmatrix} b_{1,h1|qs}\theta_{h1|qs} & b_{1,h2|qs}\theta_{h2|qs} & \dots & b_{1,hk|qs}\theta_{hk|qs} \\ b_{2,h1|qs}\theta_{h1|qs} & b_{2,h2|qs}\theta_{h2|qs} & \dots & b_{2,hk|qs}\theta_{hk|qs} \\ \dots & \dots & \dots & \dots \\ b_{n_{qs},h1|qs}\theta_{h1|qs} & b_{n_{qs},h2|qs}\theta_{h2|qs} & \dots & b_{n_{qs},hk|qs}\theta_{hk|qs} \end{bmatrix}. \tag{3.14}$$

Dalla (3.12), la verosimiglianza $L(\boldsymbol{\theta}_{h|qs}; \mathbf{B}_{h|qs})$ risulta un polinomio i cui termini sono la risultante della seguente operazione: per ciascuna delle n_{qs} righe della matrice $\mathbf{L}_{h|qs}$ si prenda un singolo elemento e poi si applichi una moltiplicazione fra tutti questi elementi. Se si ripetesse tale procedura per tutte le possibili combinazioni dei singoli elementi di ciascuna riga otterremmo tutti

i termini del polinomio (3.12). Il numero totale di tali termini, $\bar{m}_{h|qs}$, equivale quindi ad una disposizione con ripetizione di *k* elementi presi a gruppi di n_{qs}:

$$\bar{m}_{h|qs} = k^{n_{qs}} . \tag{3.15}$$

Va comunque ricordato che tale valore corrisponde al numero massimo di termini se tutti i *believes* fossero positivi. Ma l'assenza di fiducia per alcuni stati, da parte dell'osservatore, implica la presenza di zeri nella matrice $\mathbf{L}_{h|qs}$ comportando una notevole riduzione dei termini nel polinomio (3.12). Per comprendere meglio questa riduzione si suddividano le n_{qs} righe della matrice $\mathbf{L}_{h|qs}$ in funzione del numero di elementi non nulli. Senza perdita di generalità e relativamente all'*i-esima* osservazione, si definisca il numero totale di evidenze fornite dall'osservatore, $\varepsilon_{i,h|qs} = card\left\{ j : b_{i,hj|qs} > 0 \right\}$, vale a dire il numero di elementi non nulli della *i-esima* riga della matrice $\mathbf{L}_{h|qs}$, con $1 \le \varepsilon_{i,h|qs} \le k$. Sia infine $\varepsilon^{(j)}_{h|qs} = card\left\{ i : \varepsilon_{i,h|qs} = j \right\}$ il numero di righe di $\mathbf{L}_{h|qs}$ costituite da *j* evidenze non nulle, sotto il vincolo $\sum_{j=1}^{k} \varepsilon^{(j)}_{h|qs} = n_{qs}$. Allora il numero di termini $\hat{m}_{h|qs}$ che derivano dalla produttoria (3.12), nel caso alcuni elementi della matrice $\mathbf{L}_{h|qs}$ fossero nulli, corrisponde al prodotto di disposizioni con ripetizione di j elementi, dove $1 \le j \le k$, presi a gruppi di $\varepsilon^{(j)}_{h|qs}$:

$$\hat{m}_{h|qs} = \prod_{j=1}^{k} j^{\varepsilon^{(j)}_{h|qs}} \le k^{n_{qs}} = \bar{m}_{h|qs} . \tag{3.16}$$

L'uguaglianza $\hat{m}_{h|qs} = \bar{m}_{h|qs}$ vale nel caso in cui tutte le *n* valutazioni dell'osservatore, per l'*h-esimo* attributo condizionatamente a *C* ed **S**, presentassero evidenza su tutti gli stati. Il che equivale a dire che $\varepsilon_{i,h|qs} = k$ per ciascuna osservazione e quindi si avrebbe un unico insieme di cardinalità $\varepsilon^{(j)}_{h|qs} = n$, da cui il risultato $k^{n_{qs}}$.

Come precedentemente descritto, gli $\hat{m}_{h|qs}$ termini del polinomio (3.12) si ottengono da tutte le possibili moltiplicazioni dei singoli elementi di ciascuna riga della matrice $\mathbf{L}_{h|qs}$. Sia $\mathbf{P}_{h|qs} = [p_{m,hj|qs}]_{\hat{m}_{h|qs} x k}$ la matrice le cui righe costituiscono il numero di elementi di ciascuna colonna *j* della matrice $\mathbf{L}_{h|qs}$ che sono stati moltiplicati fra di loro nel formare gli $\hat{m}_{h|qs}$ polinomi. Questo equivale a dire che la matrice $\mathbf{P}_{h|qs}$ contiene le potenze dei parametri $\theta_{hj|qs}$ che costituiscono le basi $\prod_{j=1}^{k} \theta_{hj|qs}^{p_{m,hj|qs}}$ degli $\hat{m}_{h|qs}$ polinomi.

E' inoltre del tutto possibile che fra gli $\hat{m}_{h|qs}$ termini della verosimiglianza $L(\boldsymbol{\theta}_{h|qs};\mathbf{B}_{h|qs})$, alcuni abbiano la medesima base $\prod_{j=1}^{k}\theta_{hj|qs}^{p_{m,hj|qs}}$. Allora se si sommano i termini con la stessa base, il loro numero si riduce ulteriormente al valore $\tilde{m}_{h|qs} \leq \hat{m}_{h|qs}$. Nel caso in cui tutti gli elementi della matrice $\mathbf{L}_{h|qs}$ fossero non nulli, il numero di termini con base $\prod_{j=1}^{k}\theta_{hj|qs}^{p_{m,hj|qs}}$ risulterebbe pari ad una permutazione con ripetizione di n_{qs} elementi ripetuti a gruppi di $p_{m,hj|qs}$, dove $\sum_{j=1}^{k} p_{m,hj|qs} = n_{qs}$. Tale numero di termini equivale al coefficiente di una variabile casuale multinomiale, vale a dire $\binom{n_{qs}}{p_{m,h1|qs},\ldots,p_{m,hk|qs}} = n_{qs}!\left(\prod_{j=1}^{k} p_{m,hj|qs}!\right)^{-1}$. Nella realtà, dove il numero di zeri per ciascuna riga è variabile, il calcolo di $\tilde{m}_{h|qs}$ risulta difficilmente esprimibile in modo sintetico. Infine, posto $\tilde{b}_{m,h|qs}$ il coefficiente derivante dalla somma dei coefficienti degli $\hat{m}_{h|qs}$ termini della (3.12) con le stesse basi, la $L(\boldsymbol{\theta}_{h|qs};\mathbf{B}_{h|qs})$ può essere così espressa:

$$L(\boldsymbol{\theta}_{h|qs};\mathbf{B}_{h|qs}) = \sum_{m=1}^{\tilde{m}_{h|qs}} \tilde{b}_{m,h|qs} \prod_{j=1}^{k} \theta_{hj|qs}^{p_{m,hj|qs}} . \qquad (3.17)$$

3.5 Distribuzione a posteriori dei parametri

Gli assunti introdotti per la costruzione del classificatore *Naive Bayes* modificato permettono di fare inferenza separatamente sui vettori parametrici $\boldsymbol{\theta}_{h|qs}$ piuttosto che sull'intera matrice $\boldsymbol{\Theta}$. Siano $\boldsymbol{\Theta}_{qs} = [\boldsymbol{\theta}_{1|qs},\ldots,\boldsymbol{\theta}_{N|qs}]$ e $\mathbf{B}_{qs} = [\mathbf{b}_{1|qs},\ldots,\mathbf{b}_{N|qs}]$ le matrici, rispettivamente, dei parametri e dei *believes* condizionate a C e ad $\mathbf{S}$. Allora, la densità di probabilità a posteriori della matrice dei parametri $\boldsymbol{\Theta}_{qs}$ si ricava attraverso la verosimiglianza $L(\boldsymbol{\Theta}_{qs};\mathbf{B}_{qs})$ e la densità di probabilità a priori $f(\boldsymbol{\Theta}_{qs})$:

$$f(\boldsymbol{\Theta}_{qs} \mid \mathbf{B}_{qs}) = \frac{L(\boldsymbol{\Theta}_{qs};\mathbf{B}_{qs})\, f(\boldsymbol{\Theta}_{qs})}{\int_{\boldsymbol{\Theta}_{qs}} L(\boldsymbol{\Theta}_{qs};\mathbf{B}_{qs})\, f(\boldsymbol{\Theta}_{qs})\, d\boldsymbol{\Theta}_{qs}} . \qquad (3.18)$$

La verosimiglianza $L(\boldsymbol{\Theta}_{qs};\mathbf{B}_{qs})$, che per la (3.4) e la (3.10) si fattorizza come:

$$L(\boldsymbol{\Theta}_{qs};\mathbf{B}_{qs})=\prod_{h=1}^{N}L(\boldsymbol{\theta}_{h|qs};\mathbf{B}_{h|qs}), \tag{3.19}$$

insieme all'ipotesi (3.9) permette la seguente fattorizzazione della (3.18):

$$f(\boldsymbol{\Theta}_{qs}\mid\mathbf{B}_{qs})=\frac{\prod_{h=1}^{N}L(\boldsymbol{\theta}_{h|qs};\mathbf{B}_{h|qs})\prod_{h=1}^{N}f(\boldsymbol{\theta}_{h|qs})}{\int_{\boldsymbol{\theta}_{1|qs}}\ldots\int_{\boldsymbol{\theta}_{N|qs}}\prod_{h=1}^{N}L(\boldsymbol{\theta}_{h|qs};\mathbf{B}_{h|qs})\prod_{h=1}^{N}f(\boldsymbol{\theta}_{h|qs})d\boldsymbol{\theta}_{1|qs}\ldots d\boldsymbol{\theta}_{N|qs}}=$$

$$=\prod_{h=1}^{N}\frac{L(\boldsymbol{\theta}_{h|qs};\mathbf{B}_{h|qs})f(\boldsymbol{\theta}_{h|qs})}{\int_{\boldsymbol{\theta}_{h|qs}}L(\boldsymbol{\theta}_{h|qs};\mathbf{B}_{h|qs})f(\boldsymbol{\theta}_{h|qs})d\boldsymbol{\theta}_{h|qs}}=$$

$$=\prod_{h=1}^{N}f(\boldsymbol{\theta}_{h|qs}\mid\mathbf{B}_{h|qs}). \tag{3.20}$$

La (3.20) conferma, dunque, la possibilità di fare inferenza sui singoli vettori parametrici condizionati $\boldsymbol{\theta}_{h|qs}$ e lavorare separatamente con le densità di probabilità a posteriori $f(\boldsymbol{\theta}_{h|qs}\mid\mathbf{B}_{h|qs})$. Poiché la (3.17) è una mistura di kernel multinomiali e la distribuzione a priori $f(\boldsymbol{\theta}_{h|qs})$ è una Dirichlet esplicitamente rappresentata da:

$$Dir_k(\boldsymbol{\theta}_{h|qs}\mid\boldsymbol{\alpha}_{h|qs})=\frac{\Gamma(\alpha_{0,h|qs})}{\prod_{j=1}^{k}\Gamma(\alpha_{hj|qs})}\prod_{j=1}^{k}\theta_{hj|qs}^{\alpha_{hj|qs}-1}, \tag{3.21}$$

dove $\alpha_{0,h|qs}=\sum_{j=1}^{k}\alpha_{hj|qs}$, allora la distribuzione a posteriori $f(\boldsymbol{\theta}_{h|qs}\mid\mathbf{B}_{h|qs})$ viene aggiornata in forma chiusa come una mistura di $\tilde{m}_{h|qs}$ Dirichlet. Infatti, se si considera:

$$f(\boldsymbol{\theta}_{h|qs}\mid\mathbf{B}_{h|qs})=\frac{L(\boldsymbol{\theta}_{h|qs};\mathbf{B}_{h|qs})f(\boldsymbol{\theta}_{h|qs})}{\int_{\boldsymbol{\theta}_{h|qs}}L(\boldsymbol{\theta}_{h|qs};\mathbf{B}_{h|qs})f(\boldsymbol{\theta}_{h|qs})d\boldsymbol{\theta}_{h|qs}}, \tag{3.22}$$

il numeratore diventa:

$$L(\boldsymbol{\theta}_{h|qs};\mathbf{B}_{h|qs})f(\boldsymbol{\theta}_{h|qs})=$$

$$= \sum_{m=1}^{\tilde{m}_{h|qs}} \tilde{b}_{m,h|qs} \prod_{j=1}^{k} \theta_{hj|qs}^{p_{m,hj|qs}} \frac{\Gamma(\alpha_{0,h|qs})}{\prod_{j=1}^{k} \Gamma(\alpha_{hj|qs})} \prod_{j=1}^{k} \theta_{hj|qs}^{\alpha_{hj|qs}-1} =$$

$$= \sum_{m=1}^{\tilde{m}_{h|qs}} \tilde{b}_{m,h|qs} \frac{\Gamma(\alpha_{0,h|qs})}{\prod_{j=1}^{k} \Gamma(\alpha_{hj|qs})} \prod_{j=1}^{k} \theta_{hj|qs}^{\alpha_{hj|qs}+p_{m,hj|qs}-1} \tag{3.23}$$

Inoltre, definendo $\tilde{m}_{h|qs}$ variabili casuali Dirichlet $Dir_{\tilde{m}_{h|qs}}(\boldsymbol{\theta}_{m,h|qs} \mid \boldsymbol{\alpha}_{h|qs} + \mathbf{p}_{m,h|qs})$, la (3.23) si può riscrivere come:

$$L(\boldsymbol{\theta}_{h|qs}; \mathbf{B}_{h|qs}) f(\boldsymbol{\theta}_{h|qs}) =$$

$$= \sum_{m=1}^{\tilde{m}_{h|qs}} \tilde{b}_{m,h|qs} \frac{\Gamma(\alpha_{0,h|qs})}{\prod_{j=1}^{k} \Gamma(\alpha_{hj|qs})} \frac{\prod_{j=1}^{k} \Gamma(\alpha_{hj|qs} + p_{m,hj|qs})}{\Gamma(\alpha_{0,h|qs} + n_{qs})} \frac{\Gamma(\alpha_{0,h|qs} + n_{qs})}{\prod_{j=1}^{k} \Gamma(\alpha_{hj|qs} + p_{m,hj|qs})} \prod_{j=1}^{k} \theta_{hj|qs}^{\alpha_{hj|qs}+p_{m,hj|qs}-1} =$$

$$= \sum_{m=1}^{\tilde{m}_{h|qs}} \tilde{b}_{m,h|qs} \frac{\Gamma(\alpha_{0,h|qs})}{\prod_{j=1}^{k} \Gamma(\alpha_{hj|qs})} \frac{\prod_{j=1}^{k} \Gamma(\alpha_{hj|qs} + p_{m,hj|qs})}{\Gamma(\alpha_{0,h|qs} + n_{qs})} Dir_{\tilde{m}_{h|qs}}(\boldsymbol{\theta}_{m,h|qs} \mid \boldsymbol{\alpha}_{h|qs} + \mathbf{p}_{m,h|qs}) =$$

$$= \sum_{m=1}^{\tilde{m}_{h|qs}} \omega_{m,h|qs} Dir_{\tilde{m}_{h|qs}}(\boldsymbol{\theta}_{m,h|qs} \mid \boldsymbol{\alpha}_{h|qs} + \mathbf{p}_{m,h|qs}), \tag{3.24}$$

dove

$$\omega_{m,h|qs} = \tilde{b}_{m,h|qs} \frac{\Gamma(\alpha_{0,h|qs})}{\prod_{j=1}^{k} \Gamma(\alpha_{hj|qs})} \frac{\prod_{j=1}^{k} \Gamma(\alpha_{hj|qs} + p_{m,hj|qs})}{\Gamma(\alpha_{0,h|qs} + n_{qs})}. \tag{3.25}$$

Il denominatore della (3.22), invece, si riduce a:

$$\int_{\boldsymbol{\theta}_{h|qs}} L(\boldsymbol{\theta}_{h|qs}; \mathbf{B}_{h|qs}) f(\boldsymbol{\theta}_{h|qs}) d\boldsymbol{\theta}_{h|qs} =$$

$$= \int_{\boldsymbol{\theta}_{h|qs}} \sum_{m=1}^{\tilde{m}_{h|qs}} \omega_{m,h|qs} Dir_{\tilde{m}_{h|qs}}(\boldsymbol{\theta}_{m,h|qs} \mid \boldsymbol{\alpha}_{h|qs} + \mathbf{p}_{m,h|qs}) d\boldsymbol{\theta}_{h|qs} =$$

$$= \sum_{m=1}^{\tilde{m}_{h|qs}} \omega_{m,h|qs} \int_{\boldsymbol{\theta}_{h|qs}} Dir_{\tilde{m}_{h|qs}}(\boldsymbol{\theta}_{m,h|qs} \mid \boldsymbol{\alpha}_{h|qs} + \mathbf{p}_{m,h|qs}) d\boldsymbol{\theta}_{h|qs} =$$

$$= \sum_{m=1}^{\tilde{m}_{h|qs}} \omega_{m,h|qs}. \tag{3.26}$$

In conclusione, la densità di probabilità a posteriori (3.22) si può riscrivere nel seguente modo:

$$f(\boldsymbol{\theta}_{h|qs} \mid \boldsymbol{\alpha}_{h|qs}, \mathbf{B}_{h|qs}) =$$

$$= \frac{\sum_{m=1}^{\tilde{m}_{h|qs}} \omega_{m,h|qs} Dir_{\tilde{m}_{h|qs}}(\boldsymbol{\theta}_{m,h|qs} \mid \boldsymbol{\alpha}_{h|qs} + \mathbf{p}_{m,h|qs})}{\sum_{m=1}^{\tilde{m}_{h|qs}} \omega_{m,h|qs}} =$$

$$= \sum_{m=1}^{\tilde{m}_{h|qs}} q_{m,h|qs} Dir_{\tilde{m}_{h|qs}}(\boldsymbol{\theta}_{m,h|qs} \mid \boldsymbol{\alpha}_{h|qs} + \mathbf{p}_{m,h|qs}), \tag{3.27}$$

dove

$$q_{m,h|qs} = \frac{\omega_{m,h|qs}}{\sum_{m=1}^{\tilde{m}_{h|qs}} \omega_{m,h|qs}} =$$

$$= \tilde{b}_{m,h|qs} \frac{\Gamma(\alpha_{0,h|qs})}{\prod_{j=1}^{k} \Gamma(\alpha_{hj|qs})} \frac{\prod_{j=1}^{k} \Gamma(\alpha_{hj|qs} + p_{m,hj|qs})}{\Gamma(\alpha_{0,h|qs} + n_{qs})} \cdot \left(\sum_{m=1}^{\tilde{m}_{h|qs}} \tilde{b}_{m,h|qs} \frac{\Gamma(\alpha_{0,h|qs})}{\prod_{j=1}^{k} \Gamma(\alpha_{hj|qs})} \frac{\prod_{j=1}^{k} \Gamma(\alpha_{hj|qs} + p_{m,hj|qs})}{\Gamma(\alpha_{0,h|qs} + n_{qs})} \right)^{-1} =$$

$$= \frac{\tilde{b}_{m,h|qs} \prod_{j=1}^{k} \Gamma(\alpha_{hj|qs} + p_{m,hj|qs})}{\sum_{m=1}^{\tilde{m}_{h|qs}} \tilde{b}_{m,h|qs} \prod_{j=1}^{k} \Gamma(\alpha_{hj|qs} + p_{m,hj|qs})}. \tag{3.28}$$

La (3.27) esplicita la distribuzione a posteriori del vettore parametrico $\boldsymbol{\theta}_{h|qs}$ condizionatamente alla matrice dei corrispondenti *believes* $\mathbf{B}_{h|qs}$. Ne deriva, dunque, una mistura di $\tilde{m}_{h|qs}$ Dirichlet, $Dir_{\tilde{m}_{h|qs}}(\boldsymbol{\theta}_{m,h|qs} \mid \boldsymbol{\alpha}_{h|qs} + \mathbf{p}_{m,h|qs})$, i cui iperparametri corrispondono a quelli della distribuzione a priori, $\boldsymbol{\alpha}_{h|qs}$, aggiornati dalle rispettive righe $\mathbf{p}_{m,h|qs}$ della matrice $\mathbf{P}_{h|qs}$.

3.6 Predittiva e funzione di classificazione

L'osservazione sulle n unità del campione permette di fare inferenza sui parametri degli attributi $\boldsymbol{\theta}_{h|qs}$, condizionatamente a C e ad $\mathbf{S}$. Il modello di

classificazione così ottenuto viene impiegato per assegnare $l'(n+1)-esimo$ individuo ad una specifica classe della variabile di classe *C*. Si supponga, in prima istanza, che per tale individuo siano note $\mathbf{S}^{n+1}=\mathbf{s}^{n+1}$ e $\mathbf{X}^{n+1}=\mathbf{x}^{n+1}$ mentre la classe $C^{n+1}=c^{n+1}$ non sia osservata. Indicando per semplicità $\mathbf{x}^{n+1}$ al posto di $\mathbf{X}^{n+1}=\mathbf{x}^{n+1}$, e lo stesso dicasi per le altre variabili, la predittiva sulla generica classe c_q^{n+1} dell'$(n+1)-esimo$ individuo, condizionatamente a $\mathbf{x}^{n+1}$ e $\mathbf{s}^{n+1}$, si può esprimere come:

$$P(c_q^{n+1}\mid \mathbf{x}^{n+1},\mathbf{s}^{n+1})=\frac{P(\mathbf{x}^{n+1}\mid c_q^{n+1},\mathbf{s}^{n+1})P(c_q^{n+1}\mid \mathbf{s}^{n+1})}{P(\mathbf{x}^{n+1}\mid \mathbf{s}^{n+1})}. \tag{3.29}$$

La regola di classificazione (2.6), che assegna un individuo alla classe per la quale risulti massima la predittiva (3.29) non viene influenzata dal denominatore $P(\mathbf{x}^{n+1}\mid \mathbf{s}^{n+1})$. Questo permette di riscrivere la (3.29) nel seguente modo:

$$P(c_q^{n+1}\mid \mathbf{x}^{n+1},\mathbf{s}^{n+1})\propto P(\mathbf{x}^{n+1}\mid c_q^{n+1},\mathbf{s}^{n+1})P(c_q^{n+1}\mid \mathbf{s}^{n+1}), \tag{3.30}$$

che diventa, per l'assunto di indipendenza condizionata degli attributi (3.3):

$$P(c_q^{n+1}\mid \mathbf{x}^{n+1},\mathbf{s}^{n+1})\propto P(c_q^{n+1}\mid \mathbf{s}^{n+1})\prod_{h=1}^{N}P(\mathbf{x}_h^{n+1}\mid c_q^{n+1},\mathbf{s}^{n+1}). \tag{3.31}$$

La variabile $\mathbf{X}_h^{n+1}\mid c_q^{n+1},\mathbf{s}^{n+1}$ espressa sinteticamente nella notazione $\mathbf{X}_{h|qs}^{n+1}$ è distribuita, condizionatamente al set di parametri $\boldsymbol{\theta}_{h|qs}$, come una multinomiale $Mu_k\left(\mathbf{x}_{h|qs}^{n+1}\mid \boldsymbol{\theta}_{h|qs},1\right)$.

Poiché si è appreso sul vettore $\boldsymbol{\theta}_{h|qs}$ tramite la procedura di *learning* basata sui *believes* condizionati $\mathbf{B}_{qs}$ delle n_{qs} osservazioni del *training set*, si deriva la distribuzione della variabile casuale $\mathbf{X}_{h|qs}^{n+1}$:

$$\begin{aligned}
&P(\mathbf{X}_{h|qs}^{n+1}\mid \boldsymbol{\alpha}_{h|qs},\mathbf{B}_{qs})=\int_{\boldsymbol{\theta}_{h|qs}}P(\mathbf{X}_{h|qs}^{n+1}\mid \boldsymbol{\theta}_{h|qs})\,f(\boldsymbol{\theta}_{h|qs}\mid \boldsymbol{\alpha}_{h|qs},\mathbf{B}_{qs})d\boldsymbol{\theta}_{h|qs}=\\
&=\int_{\boldsymbol{\theta}_{h|qs}}Mu_k\left(\mathbf{x}_{h|qs}^{n+1}\mid \boldsymbol{\theta}_{h|qs},1\right)\cdot\left(\sum_{m=1}^{\tilde{m}_{h|qs}}q_{m,h|qs}Dir_{\tilde{m}_{h|qs}}(\boldsymbol{\theta}_{m,h|qs}\mid \boldsymbol{\alpha}_{h|qs}+\mathbf{p}_{m,h|qs})\right)d\boldsymbol{\theta}_{h|qs}=\\
&=\sum_{m=1}^{\tilde{m}_{h|qs}}q_{m,h|qs}\int_{\boldsymbol{\theta}_{h|qs}}Mu_k\left(\mathbf{x}_{h|qs}^{n+1}\mid \boldsymbol{\theta}_{h|qs},1\right)\cdot Dir_{\tilde{m}_{h|qs}}(\boldsymbol{\theta}_{m,h|qs}\mid \boldsymbol{\alpha}_{h|qs}+\mathbf{p}_{m,h|qs})d\boldsymbol{\theta}_{h|qs}=\\
&=\sum_{m=1}^{\tilde{m}_{h|qs}}q_{m,h|qs}Md_k(\mathbf{x}_{h|qs}^{n+1}\mid \boldsymbol{\alpha}_{h|qs}+\mathbf{p}_{m,h|qs},1)\ \text{(Guimaraes e Lindrooth, 2005).}
\end{aligned} \tag{3.32}$$

Quindi la variabile casuale $\mathbf{X}_{h|qs}^{n+1} \mid \boldsymbol{\alpha}_{h|qs}, \mathbf{B}_{qs}$ è una mistura di $\tilde{m}_{h|qs}$ Dirichlet Multinomiali su una singola osservazione $Md_k(\mathbf{x}_{h|qs}^{n+1} \mid \boldsymbol{\alpha}_{h|qs} + \mathbf{p}_{m,h|qs}, 1)$. Allora la (3.31) si può scrivere come:

$$P(c_q^{n+1} \mid \mathbf{x}^{n+1}, \mathbf{s}^{n+1}, \boldsymbol{\alpha}_{h|qs}, \mathbf{B}_{qs}) \propto P(c_q^{n+1} \mid \mathbf{s}^{n+1}) \prod_{h=1}^{N} P(\mathbf{x}_{h|qs}^{n+1} \mid \boldsymbol{\alpha}_{h|qs}, \mathbf{B}_{qs}), \tag{3.33}$$

dove la probabilità $P(\mathbf{x}_{h|qs}^{n+1} \mid \boldsymbol{\alpha}_{h|qs}, \mathbf{B}_{qs})$ per la (3.32) è:

$$P(\mathbf{x}_{h|qs}^{n+1} \mid \boldsymbol{\alpha}_{h|qs}, \mathbf{B}_{qs}) = \sum_{m=1}^{\tilde{m}_{h|qs}} q_{m,h|qs} Md_k(\mathbf{x}_{h|qs}^{n+1} \mid \boldsymbol{\alpha}_{h|qs} + \mathbf{p}_{m,h|qs}, 1) =$$

$$= \frac{\boldsymbol{\alpha}_{h|qs} + \sum_{m=1}^{\tilde{m}_{h|qs}} q_{m,h|qs} \mathbf{p}_{m,h|qs}}{\alpha_{0,h|qs} + n}. \tag{3.34}$$

Il vettore delle probabilità a priori condizionate $P(c_q^{n+1} \mid \mathbf{s}^{n+1})$ che compare nella (3.31) può essere stimato con la stima di massima verosimiglianza (2.9):

$$\hat{P}(c_q^{n+1} \mid \mathbf{s}^{n+1}) = \frac{n_{qs}}{n_s}, \tag{3.35}$$

mediante le osservazioni del *training set* che costituisce un campione rappresentativo della popolazione di riferimento e dove $n_s = \sum_q n_{qs}$.

Allora la stima della (3.31) diviene:

$$\hat{P}(c_q^{n+1} \mid \mathbf{x}^{n+1}, \mathbf{s}^{n+1}, \boldsymbol{\alpha}_{h|qs}, \mathbf{B}_{qs}) \propto n_{qs} \prod_{h=1}^{N} \frac{\boldsymbol{\alpha}_{h|qs} + \sum_{m=1}^{\tilde{m}_{h|qs}} q_{m,h|qs} \mathbf{p}_{m,h|qs}}{\alpha_{0,h|qs} + n}. \tag{3.36}$$

Viceversa, se l'attributo $\mathbf{X}_h$ non è osservabile si ricorre alla *soft evidence* ad esso associata utilizzando i corrispondenti *believes* forniti dall'osservatore sugli stati dell'*h-esimo* attributo. Siano quindi noti $\mathbf{s}^{n+1}$ e $\mathbf{b}_h^{n+1}$ e si voglia stimare la probabilità che l' $(n+1)-esimo$ individuo appartenga alla classe c^{n+1}. Nel caso siano indicati almeno due stati dall'osservatore si considera la combinazione lineare fra le probabilità di questi stati con coefficienti che corrispondono ai *believes*, $\sum_{j=1}^{k} b_{hj}^{n+1} P(x_{hj|qs}^{n+1} \mid \boldsymbol{\alpha}_{h|qs}, \mathbf{B}_{qs})$. Tale combinazione lineare include anche il caso *hard evidence*, in cui l'osservatore indichi un solo stato come possibile manifestazione dell'*h-esimo* attributo, che corrisponde alla (3.34).

Considerando la combinazione lineare fra gli stati ritenuti possibili realizzazioni, la stima della (3.33) si riscrive:

$$\hat{P}(c_q^{n+1} \mid \mathbf{b}^{n+1}, \mathbf{s}^{n+1}, \boldsymbol{\alpha}_{h|qs}, \mathbf{B}_{qs}) \propto n_{qs} \prod_{h=1}^{N} \sum_{j=1}^{k} b_{hj}^{n+1} \frac{\alpha_{hj|qs} + \sum_{m=1}^{\tilde{m}_{h|qs}} q_{m,hj|qs}\, p_{m,hj|qs}}{\alpha_{0,h|qs} + n}. \quad (3.37)$$

Per assegnare l'$(n+1)-esimo$ individuo ad una classe di età a partire dalla predittiva (3.37) si utilizza la regola decisionale (2.11):

$$\hat{c}^{n+1} = f(n+1) = \underset{q}{\operatorname{argmax}}\, P(c_q^{n+1} \mid \mathbf{b}^{n+1}, \mathbf{s}^{n+1}, \boldsymbol{\alpha}_{h|qs}, \mathbf{B}_{qs}) =$$

$$= \underset{q}{\operatorname{argmax}} \left(n_{qs} \prod_{h=1}^{N} \sum_{j=1}^{k} b_{hj}^{n+1} \frac{\alpha_{hj|qs} + \sum_{m=1}^{\tilde{m}_{h|qs}} q_{m,hj|qs}\, p_{m,hj|qs}}{\alpha_{0,h|qs} + n} \right). \quad (3.38)$$

La regola di classificazione alternativa (2.12) è legata all'introduzione di una soglia probabilistica π che garantisce l'assegnazione di un individuo ad una classe solo se viene superata tale soglia di accettabilità, la quale varierà a seconda del caso, del giudice e della legislazione.

In generale, indicando con P_q la probabilità $P(c_q^{n+1} \mid \mathbf{b}^{n+1}, \mathbf{s}^{n+1}, \boldsymbol{\alpha}_{h|qs}, \mathbf{B}_{qs})$, si avrà per $Q > 2$:

$$f_{\pi,Q}(n+1) = \begin{cases} c_{q^*} & \text{se } \underset{q}{\operatorname{argmax}}\, P_q = q^* \text{ e } P_{q^*} \geq \pi \\ \textit{non classificato} & \text{altrimenti} \end{cases}, \quad (3.39)$$

dove è contemplata la possibilità che non si possa classificare l'individuo. Nel caso invece la variabile di classe sia dicotomica, $Q = 2$, la (3.39) si può riscrivere:

$$f_{\pi,2}(n+1) = \begin{cases} c_2 & \text{se } \underset{q}{\operatorname{argmax}}\, P_q = 2 \text{ e } P_2 \geq \pi \\ c_1 & \text{altrimenti} \end{cases}, \quad (3.40)$$

dove l'assegnazione alla classe c_2 è ancora vincolata al superamento della soglia probabilistica π ma, a differenza della (3.39), con la (3.40) si può

sempre classificare l'individuo. Nel caso $\pi = 0,50$, la regola di classificazione (3.40) coincide con la (3.38).

3.7 Performance dell'esperto e del modello

Attraverso una suddivisione del campione nel *training* e *test data set* si può verificare la *performance* della classificazione ottenuta mediante l'impiego del modello stimato sulla base delle valutazioni fornite dall'osservatore.

Un diverso tipo di valutazione può essere intrapresa attraverso indici di variabilità intra-osservatore, comparando i *believes* forniti dallo stesso esperto sulle stesse unità in istanti temporali differenti.

Inoltre si potrebbero paragonare due differenti esperti misurando la variabilità inter-osservatore confrontando i *believes* associati alle stesse unità. Mediante le variabilità intra e inter-osservatore si ricavano poi, rispettivamente, l'indice di ripetibilità e l'indice di riproducibilità introdotti nel *Cap. 2*.

Per valutare invece l'intero modello che considera congiuntamente il metodo di classificazione dentale di Demirijan, la valutazione dell'esperto, la predittiva e la regola classificatoria, stimato sulle valutazioni che l'esperto fornisce sul *training set*, si utilizzano i risultati ottenuti tramite un *test data set*. Ripetendo per *R* repliche la procedura di estrazione casuale da un campione complessivo di un *training set* necessario alla stima del modello e di un *test set* per la verifica della sua *performance*, si possono ottenere indici medi che attestano il grado di corretta classificazione.

3.7.1 Errori di classificazione

Sia $C = \{c_1, c_2, ..., c_Q\}$ la variabile di classe osservata nel *test data set* e $\hat{C} = \{\hat{c}_1, \hat{c}_2, ..., \hat{c}_Q\}$ la variabile che rappresenta la classe di attribuzione mediante la regola decisionale (3.38) o (3.39). Per ciascuna unità di ciascun *test set* si ottiene, quindi, l'assegnazione ad una classe $\hat{c}$, la cui correttezza potrà essere successivamente verificata confrontandola con la classe osservata c.

Si consideri il generico evento condizionato $e_{ij} \equiv \hat{c}_j \mid c_i$ vale a dire l'evento che corrisponde all'assegnazione alla classe $\hat{c}_j$ di un'unità che appartiene di fatto alla classe c_i. Dunque se $i = j$ significa che l'unità è stata correttamente assegnata altrimenti, nel caso $i \neq j$, ci si trova di fronte ad un errore di classificazione che sarà tanto maggiore quanto più grande è la differenza $|i - j|$. Il caso peggiore corrisponde a $|i - j| = Q - 1$, che si verifica

quando un'unità appartenente ad una delle due classi estreme c_1 o c_Q viene classificata, rispettivamente, nella classe $\hat{c}_Q$ o $\hat{c}_1$, producendo i corrispondenti errori e_{1Q} o e_{Q1}. Ovviamente ci si riferisce alle unità per le quali è stata effettuata una classificazione – si ricordi che la regola classificatoria (3.39) può produrre anche casi non classificabili – e quindi per tali unità non si può parlare né di corretta né di scorretta classificazione.

Sia n_{ij} la frequenza congiunta della realizzazione della variabile doppia $(C,\hat{C}) = (c_i, \hat{c}_j)$ con corrispondente tabella di contingenza:

Tabella 3.1. *Distribuzione doppia di frequenza di* $(C,\hat{C})$

$C \setminus \hat{C}$	$\hat{c}_1$	...	$\hat{c}_Q$
c_1	n_{11}	...	n_{1Q}
...	...	...	...
c_Q	n_{Q1}	...	n_{QQ}

dove il totale per riga, $n_{i.} = \sum_{j=1}^{Q} n_{ij}$, corrisponde alla frequenza della classe osservata c_i, se tutte le unità sono state classificate. La percentuale totale di corretta classificazione prodotta, equivale semplicemente alla somma delle frequenze relative poste sulla diagonale principale, $\sum_{i=1}^{Q} n_{ii} \Big/ n$.

Qualora si adottasse la regola classificatoria (3.38), la probabilità del generico errore e_{ij} sarà:

$$P(e_{ij}) = \frac{n_{ij}}{n_{i.}}. \tag{3.41}$$

Nel caso invece si utilizzi la (3.39) bisogna tener conto della possibilità che non tutte le unità vengono classificate. Se nc_i rappresenta il numero di unità non classificate appartenenti alla classe c_i, con $n_{i.} = \sum_{j=1}^{Q} n_{ij} + nc_i$, si ha:

$$P(e_{ij}) = \frac{n_{ij}}{n_{i.} - nc_i}. \tag{3.42}$$

Si consideri ora il caso in cui la variabile di classe sia dicotomica, $C = \{c_0, c_1\}$, e la regola classificatoria sia la (3.40), garantendo così la classificazione per tutte le unità del *test set*. Inoltre, la variabile di classe *C* sia osservata su ciascuna delle *n* unità, come accade per tutte le unità di un *test data set*.

Una scelta opportuna del vincolo probabilistico π può essere effettuata al fine di ottenere migliori prestazioni classificatorie. Sia dunque la corrispondente tavola 2x2:

Tabella 3.2. *Distribuzione doppia di frequenza di* $(C, \hat{C})$ *nel caso dicotomico*

$C \setminus \hat{C}$	$\hat{c}_0$	$\hat{c}_1$
c_0	n_{00}	n_{01}
c_1	n_{10}	n_{11}

In letteratura medica si ricorre spesso ad una situazione come quella rappresentata dalla *Tab. 3.2*, in cui si voglia ad esempio individuare quali soggetti sono malati, c_1, e quali sani, c_0 (Agresti, 2002). Mediante opportuni strumenti di valutazione, come diagnosi, radiografie oppure *test*, si procede quindi alla classificazione degli individui in positivi, $\hat{c}_1$, o negativi $\hat{c}_0$.

Per valutare le prestazioni di classificazione dello strumento utilizzato, Agresti (2002) definisce i seguenti indici:

$$sensibilità = P(e_{11}) = \frac{n_{11}}{n_{1.}} \tag{3.43}$$

$$specificità = P(e_{00}) = \frac{n_{00}}{n_{0.}} \tag{3.44}$$

Facendo variare la soglia probabilistica π si possono quindi ottenere valori di sensibilità o specificità opportuni tenendo conto del trade-off che sussiste fra tali indici. Ad esempio, all'aumentare della soglia π cresce il numero n_{10} dei soggetti malati erroneamente classificati, diminuendo conseguentemente la sensibilità, ma diminuisce anche il numero n_{01} degli individui sani erroneamente classificati, facendo così aumentare la specificità.

Per la determinazione di una soglia probabilistica π ottimale, a cui corrispondono valori di sensibilità e specificità ricercati, viene impiegato il grafico della curva ROC (*Receiver Operating Characteristic*) che mette in relazione la sensibilità con $(1 - \text{specificità})$, per i vari livelli probabilistici π.

In un'applicazione reale in cui si voglia assegnare l' $(n+1)-esima$ unità ad una delle Q classi della variabile C, si utilizza la valutazione fornita dall'osservatore facendo uso del modello classificatorio proposto e stimato sulle n osservazioni del *training set*. Nel caso in cui la classificazione produca il risultato $\hat{c}_Q^{n+1}$, o equivalentemente $\hat{c}_1^{n+1}$, è lecito chiedersi quale sia la probabilità che la vera classe di appartenenza sia c_1^{n+1}, oppure c_Q^{n+1}, vale a dire aver commesso l'errore peggiore $c_1^{n+1} \mid \hat{c}_Q^{n+1}$, o l'equivalente $c_Q^{n+1} \mid \hat{c}_1^{n+1}$. Allora per il teorema della probabilità condizionata si ha che:

$$P\left(c_1^{n+1} \mid \hat{c}_Q^{n+1}\right) = \frac{P\left(\hat{c}_Q^{n+1} \mid c_1^{n+1}\right) P(c_1^{n+1})}{P\left(\hat{c}_Q^{n+1}\right)} = \frac{P\left(e_{1Q}^{n+1}\right) P(c_1^{n+1})}{\sum_{i=1}^{Q} P\left(e_{iQ}^{n+1}\right) P(c_i^{n+1})}, \tag{3.45}$$

ed allo stesso modo dicasi per $P\left(c_Q^{n+1} \mid \hat{c}_1^{n+1}\right)$. Le probabilità degli errori $P\left(e_{iQ}^{n+1}\right)$ possono essere stimate attraverso il valor medio, effettuato sugli R *test data set*s a disposizione, della (3.41) o (3.42) a secondo che la regola classificatoria, rispettivamente, sia in grado di classificare tutte le unità oppure no. Dunque, indicato con $e_{r,1Q}$ l'errore commesso nell'*r-esimo test data set* nel classificare in $\hat{c}_{R,Q}$ unità appartenenti alla classe $c_{R,1}$, la stima della probabilità $P\left(e_{1Q}^{n+1}\right)$ risulta:

$$\hat{P}\left(e_{1Q}^{n+1}\right) = \bar{P}\left(e_{1Q}\right) = \frac{\sum_{r=1}^{R} P\left(e_{r,1Q}\right)}{R}. \tag{3.46}$$

La stima della probabilità $P(c_i^{n+1})$ può essere ricavata dalla (2.9) come proporzione misurata sull'intero campione derivante dall'unione del *training* e *test data set*.

La stima $\hat{P}\left(c_1^{n+1} \mid \hat{c}_Q^{n+1}\right)$ fornisce quindi una misura della probabilità che il modello produca la "peggiore" classificazione qualora l'unità sia assegnata alla classe $\hat{c}_Q^{n+1}$. Tale probabilità tiene conto anche del fatto che non tutte le unità possono essere classificate, se viene utilizzata la regola di classificazione (3.39), riproporzionando gli errori sul totale delle sole unità classificate tramite la (3.42).

A tal proposito è utile conoscere anche la probabilità che il modello non fornisca alcuna classificazione, $P(NC_\pi)$, a seconda del vincolo probabilistico π impiegato nella (3.39). Sia dunque $nc_{r,\pi}$ il numero delle unità non classificate per l'*r-esimo test data set* in funzione della soglia probabilistica π,

allora un indice della *capacità classificatoria* CC_π di un modello potrebbe essere:

$$CC_\pi = 1 - P\left(NC_\pi\right) = 1 - \frac{\sum_{r=1}^{R} nc_{r,\pi}}{n \cdot R}. \tag{3.47}$$

Le misure di *performance* (3.46) e la (3.47) potrebbero quindi essere impiegate come indicatori delle prestazioni che si andrebbero a produrre con il modello di classificazione proposto.

3.7.2 Riproducibilità e ripetibilità

Nel caso in cui ci possa essere incertezza nella lettura di una unità di osservazione è possibile che la valutazione fornita da un osservatore possa essere diversa da quella che egli riproporrebbe a distanza di tempo, oppure da quella di un altro oservatore, sempre sulla stessa unità. Nel caso ci sia la possibilità di utilizzare la *soft evidence* indicando più stati possibili con relativi *believes*, significa che le valutazioni fornite sulle medesime unità dallo stesso osservatore in istanti temporali differenti, oppure da diversi osservatori, possono essere soggette a discrepanze.

Si consideri il vettore dei *believes* per l'*h-esimo* attributo specificato per l'*i-esima* unità, $\mathbf{b}_{i,h} = (b_{i,h1},...,b_{i,hk})$. Si indichi ulteriormente con $b_{i,hj}^{t_1}$ e $b_{i,hj}^{t_2}$ la fiducia riposta dall'osservatore nel *j-esimo* stato dell'*h-esimo* attributo al tempo t_1 e t_2. Allora:

$$S_{i,h}^{t_1,t_2} = \sum_{j=1}^{k} \left| b_{i,hj}^{t_1} - b_{i,hj}^{t_2} \right| \tag{3.48}$$

è una misura delle divergenze delle due distribuzioni dei *believes* che quantifica le differenze fra le due "letture" fornite dallo stesso osservatore sulla stessa unità ma in tempi differenti. Si prova facilmente che $0 \le S_{i,h}^{t_1,t_2} \le 2$ dove gli estremi 0 e 2 sono ottenuti, rispettivamente, nel caso di perfetta concordanza (l'osservatore assegna i medesimi *believes* agli stati nelle due rilevazioni) e discordanza massima (l'osservatore fornisce evidenze su stati diversi).

Disponendo di un campione di n elementi e una doppia osservazione in t_1 e t_2, si può ricavare un indice che misura la *variabilità intra-osservatore locale*, dove il termine "locale" si riferisce all'*h-esimo* attributo:

$$S_h^{t_1,t_2} = \frac{\sum_{i=1}^{n} S_{i,h}^{t_1,t_2}}{2n} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k}\left|b_{i,hj}^{t_1} - b_{i,hj}^{t_2}\right|}{2n}, \tag{3.49}$$

dove $0 \le S_h^{t_1,t_2} \le 1$. Ottenuta una misura di variabilità intra-osservatore locale se ne può ricavare l'*indice di ripetibilità locale* $\mu_h^{t_1,t_2}$:

$$\mu_h^{t_1,t_2} = 1 - S_h^{t_1,t_2} = 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{k}\left|b_{i,hj}^{t_1} - b_{i,hj}^{t_2}\right|}{2n}, \tag{3.50}$$

con $0 \le \mu_h^{t_1,t_2} \le 1$, che misura la capacità dell'osservatore di ripetere la stessa valutazione fornita in precedenza sulla medesima unità.

Un indice di ripetibilità globale μ^{t_1,t_2} si ricava come media degli indici di ripetibilità locali:

$$\mu^{t_1,t_2} = \frac{\sum_{h=1}^{N} \mu_h^{t_1,t_2}}{N} = 1 - \frac{\sum_{h=1}^{N}\sum_{i=1}^{n}\sum_{j=1}^{k}\left|b_{i,hj}^{t_1} - b_{i,hj}^{t_2}\right|}{2n \cdot N}. \tag{3.51}$$

Infine, gli indici t_1 e t_2 possono anche essere considerati come le etichette attribuite ai due differenti osservatori. In questo caso la (3.49) diverrebbe una misura della *variabilità inter-osservatori locale* fra due diversi osservatori che forniscono le proprie valutazione sulle medesime unità. Equivalentemente, gli indici (3.50) e (3.51) misureranno, rispettivamente, misurerà l'*indice di riproducibilità locale* e *globale*, come complementare della variabilità inter-osservatore, vale a dire una misura della capacità dei due osservatori di riprodurre le stesse valutazioni sulle medesime unità.

Per valutare le discordanze fra due osservatori bisogna tener conto del fatto che alcuni degli accordi osservati potrebbero essere puramente casuali. Se si considera il caso di due osservatori chiamati a fornire le valutazioni sulle medesime unità, potrebbe capitare, ad esempio, che per alcune unità uno dei due tiri ad indovinare oppure tenda sistematicamente a dare lo stesso giudizio.

Allora si introduce un indice di accordo che sia depurato dall'effetto del caso. Spitzer et al. (1967) affermano che nel caso emerga una corrispondenza fra le valutazioni di due esperti, allora si può parlare di un'associazione fra i due pareri ma non necessariamente accordo. Non è infatti irrilevante la componente di accordo casuale che aumenta al diminuire del numero di classi della variabile di classe esaminata (Dhanjal et al, 2006). Per queste ragioni Cohen (1960) introdusse un indice, ampiamente utilizzato, denominato *Kappa di Cohen*:

$$K = \frac{po - pc}{1 - pc}, \tag{3.52}$$

dove *po* è la proporzione di volte in cui i due osservatori concordano circa lo stato della variabile osservata e *pc* è la proporzione di volte in cui ci si attende che essi siano in accordo per puro caso. Applicando la (3.52) all'indice di riproducibilità locale (3.50) si ottiene:

$$\tilde{\mu}_h^{t_1,t_2} = \frac{\mu_h^{t_1,t_2} - {}_r\mu_h^{t_1,t_2}}{1 - {}_r\mu_h^{t_1,t_2}}, \tag{3.53}$$

dove ${}_r\mu_h^{t_1,t_2}$ corrisponde ad una misura di *riproducibilità casuale locale*. Per il calcolo di ${}_r\mu_h^{t_1,t_2}$ è stato proposto un indice che sfrutta l'indipendenza stocastica delle tabelle di contingenza doppie dove, al posto delle frequenze marginali si prendono in esame i totali dei *believes* forniti su tutte le *n* osservazioni.

Siano dunque $\mathbf{B}_h^{t_1} = [b_{i,hj}^{t_1}]_{nxk}$ e $\mathbf{B}_h^{t_2} = [b_{i,hj}^{t_2}]_{nxk}$ le matrice dei *believes* dell'*h-esimo* attributo per tutte le *n* osservazioni del *training data set*, dove t_1 e t_2 indicano due instanti temporali diversi oppure differenti osservatori:

$$\mathbf{B}_h^{t_1} = \begin{bmatrix} b_{1,h1}^{t_1} & b_{1,h2}^{t_1} & \dots & b_{1,hk}^{t_1} \\ b_{2,h1}^{t_1} & b_{2,h2}^{t_1} & \dots & b_{2,hk}^{t_1} \\ \dots & \dots & \dots & \dots \\ b_{n,h1}^{t_1} & b_{n,h2}^{t_1} & \dots & b_{n,hk}^{t_1} \end{bmatrix}, \tag{3.54}$$

$$\mathbf{B}_h^{t_2} = \begin{bmatrix} b_{1,h1}^{t_2} & b_{1,h2}^{t_2} & \dots & b_{1,hk}^{t_2} \\ b_{2,h1}^{t_2} & b_{2,h2}^{t_2} & \dots & b_{2,hk}^{t_2} \\ \dots & \dots & \dots & \dots \\ b_{n,h1}^{t_2} & b_{n,h2}^{t_2} & \dots & b_{n,hk}^{t_2} \end{bmatrix}. \tag{3.55}$$

I totali di colonna delle matrici (3.54) e (3.55) corrispondono alla fiducia globale che un osservatore ha riposto in un determinato stato *j* dell'attributo *h* per tutte le *n* osservazioni, relativamente a t_1 e t_2:

$$b_{\cdot hj}^{t_1} = \sum_{i=1}^{n} b_{i,hj}^{t_1}, \tag{3.56}$$

$$b_{\cdot hj}^{t_2} = \sum_{i=1}^{n} b_{i,hj}^{t_2}. \tag{3.57}$$

Per ciascun *h-esimo* attributo si costruisca una tabella a doppia entrata delle misurazioni relative a t_1 e t_2 effettuate sulle medesime unità, riportando i totali dei *believes*, (3.56) e (3.57), corrispondenti a ciascun *j-esimo* stato:

Tabella 3.3. *Tabella a doppia entrata relativamente alle valutazioni in* t_1 *e* t_2 *dell'h-esimo attributo*

$h: t_1 \setminus t_2$	1	...	k	
1	...	...	...	$b_{.h1}^{t_1}$
...	...	...	...	...
k	...	...	...	$b_{.hk}^{t_1}$
	$b_{.h1}^{t_2}$	...	$b_{.hk}^{t_2}$	

Ad esempio si consideri una tabella di contingenza doppia – corrispondente alla *Tab. 3.3* ma con le frequenze al posto dei *believes* – che contempli esclusivamente valutazioni di tipo *hard evidence*. Se si calcolano le frequenze teoriche nel caso di indipendenza stocastica si ottengono, sulla diagonale principale, il numero di unità che ci si aspetta siano state valutate allo stesso modo per puro caso. Allora la somma delle frequenze teoriche relative presenti in diagonale è una possibile misura del grado di accordo casuale fra le due diverse misurazioni.

L'indice di riproducibilità casuale locale proposto utilizza la medesima procedura applicata ai *believes* caratterizzanti valutazioni di tipo *soft evidence*. Poiché le frequenze congiunte relative, nel caso di indipendenza stocastica, equivalgono al prodotto delle corrispondenti frequenze marginali relative, allora una misura della riproducibilità casuale locale può essere:

$$r\,\mu_h^{t_1,t_2} = \sum_{j=1}^{k} \frac{b_{.hj}^{t_1}}{n} \cdot \frac{b_{.hj}^{t_2}}{n} = \sum_{j=1}^{k} \frac{b_{.hj}^{t_1} b_{.hj}^{t_2}}{n^2} \tag{3.58}$$

L'indice (3.52) viene utilizzato per variabili nominali mentre nel caso di variabili ordinali sarebbe opportuno differenziare il grado di disaccordo che, in ambiti quali quello medico e forense, comporterebbe conseguenze non banali. Spitzer et al. (1967) propongono l'indice *Kappa di Cohen ponderato* per differenziare l'entità del disaccordo con opportuni pesi a seconda della "distanza" fra gli stati:

$$K_W = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}\, po_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}\, pc_{ij}}, \tag{3.59}$$

dove po_{ij} e pc_{ij} rappresentano le proporzioni, rispettivamente, osservate e casuali, di unità per le quali sono stati indicati l'*i-esimo* stato con la misurazione t_1 ed il *j-esimo* stato con la misurazione t_2 mentre w_{ij} è il corrispondente peso del disaccordo che è proporzionale alla differenza $|i-j|$ e all'entità della valutazione stessa. Per quanto riguarda le celle sulla diagonale si ha quindi $w_{ij} = 0$. Comunque, se non si ritenesse necessario pesare le differenze fra gli stati, l'indice (3.53) fornisce un'idonea misura di riproducibilità locale depurata dall'effetto del caso.

3.8 Inclusione della variabilità intra-osservatore nella predittiva

L'osservazione incerta degli attributi, caratterizzata dall'uso della *soft evidence*, consente di introdurre nella predittiva (3.37) un ulteriore fattore di variabilità connesso direttamente alla variabilità intra-osservatore (3.49).

L'idea è quella di considerare un modello, simile ai *modelli di mutazione* utilizzati in ambito genetico (Evett e Weir, 1998), che possa distribuire i *believes* agli stati adiacenti rispetto a quelli su cui è stata fornita *soft evidence*. Questo equivale a considerare, nel caso l'osservatore fornisca *soft evidence* relativamente al *j-esimo* stato dell'*h-esimo* attributo per l'$(n+1)-esimo$ individuo, la possibilità che anche gli stati adiacenti $j-1$ e $j+1$ possano essere indicati in occasioni successive come possibili realizzazioni. Si assume dunque che il grado di incertezza dell'osservazione sia descritta da un modello di *transizione one-step*.

Il modello di transizione può essere impiegato per modificare i *believes* necessari al calcolo della predittiva (3.37) mediante la quale si effettua la classificazione. La distribuzione dei *believes* agli stati adiacenti avverrà in funzione della capacità che un osservatore ha nel ripetere le stesse valutazioni sulle medesime unità a distanza di tempo. Maggiore è la ripetibilità dell'osservatore più piccola sarà la "transizione" della valutazione verso i due stati adiacenti.

Sia dunque definita la *probabilità di transizione* $\phi_{h,ji}$ dallo stato j allo stato i dell'*h-esimo* attributo:

$$\phi_{h,ji} = \begin{cases} (1+\mu_h^{t_1,t_2})/2 & \text{se } i = j \text{ con } j \in \{1,k\} \\ \mu_h^{t_1,t_2} & \text{se } i = j \text{ con } j \notin \{1,k\} \\ (1-\mu_h^{t_1,t_2})/2 & \text{se } |i-j| = 1 \\ 0 & \text{se } |i-j| \geq 2 \end{cases}, \tag{3.60}$$

dove $\phi_{h,jj}$ corrisponde alla probabilità di permanenza nel *j-esimo* stato indicato. Tale modello di transizione comporta la distribuzione dei *believes* agli stati adiacenti per una quantità pari a $(1-\mu_h^{t_1,t_2})/2$ compresi gli stati estremi $j=1$ e $j=k$, dove non è stato posto, come era naturale pensare, una probabilità di transizione allo stato adiacente pari a $1-\mu_h^{t_1,t_2}$. Questo perché, nel caso di dati mancanti con *believes* uniformemente distribuiti e pari a $1/k$, si avrebbero probabilità maggiori per i due stati $j=2$ e $j=k-1$. Con il modello di transizione proposto dalla (3.60), invece, si garantisce che anche in presenza di dati mancanti la transizione lascia inalterati i *believes*.

Tanto minore è l'indice di ripetibilità locale di un osservatore tanto maggiore sarà la variazione dei *believes*. Per ciascuno stato si potrà così calcolare il *belief* modificato b_{hj}^*:

$$b_{hj}^* = \sum_{i=1}^{k} \phi_{h,ij} b_{hi} = \sum_{i=j-1}^{j+1} \phi_{h,ij} b_{hi} . \tag{3.61}$$

Nel caso di ripetibilità massima, $\mu_h^{t_1,t_2} = 1$, l'osservatore è in grado di ripetere perfettamente le stesse valutazioni per l'*h-esimo* attributo. La probabilità di permanenza $\phi_{h,jj} = 1$ significa che ciascun *belief* elicitato non viene modificato, vale a dire che $b_{hj}^* = b_{hj} \quad \forall j$. Questo perché l'osservatore ha una capacità di ripetersi perfetta ed i *believes* che fornisce sono praticamente "certi".

Capitolo 4
Applicazione e risultati

4.1 Campione e variabili

Il campione esaminato è composto da $n = 559$ soggetti per i quali si dispongono delle radiografie ortopanoramiche (OPT) di tutti i terzi molari. Ciascuna OPT è stata osservata da due osservatori che, in modo indipendente dall'altro, hanno fornito la propria valutazione circa lo *Stato di Mineralizzazione* raggiunto da ciascun terzo molare secondo la classificazione di Demirijan. Al termine dell'osservazione è emerso che lo stato iniziale A della scala di Demirijan non è mai stato attribuito e sono stati individuati pochissimi denti nello stato B o C. Si sono quindi raggruppati i primi quattro stati (A, B, C e D) in uno solo, facendo diminuire le 8 categorie della classificazione di Demirijan a sole $k = 5$. Ciascun esperto, per ogni OPT e senza conoscere l'età del soggetto, ha fornito le valutazioni sugli sviluppi dentali dei quattro terzi molari. Nel caso di incertezza fra due stati adiacenti l'osservatore ha fornito il grado di fiducia riposto in ciascuno stato come possibile manifestazione, il cosiddetto *belief*.

La variabile di classificazione *Età*, osservata su tutti i soggetti in anni compiuti, è compresa fra 16 e 23 per il 98,4% degli individui ed è stata ricodificata in base alla classificazione dicotomica o tricotomica, mentre l'insieme delle covariate $\mathbf{S}$ contiene la sola variabile *Genere* che ripartisce il campione in 307 maschi (54,9%) e 252 femmine (45,1%). Non è presa in considerazione la variabile *Gruppo Etnico* di appartenenza in quanto sono stati esaminati tutti soggetti di etnia caucasica che vivono in Italia.

Inoltre le OPT sono state classificate, secondo la *Tecnologia* che le ha prodotte, in 449 OPT analogiche (80,3%) e 110 digitali (19,7%). Le OPT

analogiche sono ottenute da un'esposizione diretta di una pellicola fotografica ai raggi X e scannerizzata in un file jpg a 200 dpi da uno scanner professionale *EPSON Expression 1680 Pro*. Le OPT digitali sono ottenute invece da un metodo elettronico di acquisizione CDD (*Capability Development Document*) che esporta i files direttamente in formato jpg da un sistema radiografico. Tutti i files sulle OPT sono stati messi in un sito web (http://www.proofweb.eu) per permettere la valutazione da parte degli esperti.

Sebbene sia una delle variabili osservate del campione a disposizione, la tecnologia non rientra nella procedura di *learning* parametrico per stimare i modelli di classificazione. La tecnologia è dunque utilizzata esclusivamente per articolare i risultati e le *performance* sia dei modelli che degli osservatori sui quali i modelli sono stimati.

4.2 Osservatori, soft evidence e missing data

Chiamiamo con *Esperto A* ed *Esperto B* i due osservatori che hanno fornito le valutazioni sugli stati di mineralizzazione dei terzi molari in ciascuna OPT. Non tutte le OPT hanno però prodotto le quattro valutazioni, una per terzo molare. Infatti, la presenza di dati mancanti è una circostanza usuale in qualunque insieme di informazioni. Nel caso in esame si sono verificate due cause che li hanno prodotti: l'assenza fisica del dente (NA) e l'impossibilità di classificarli (NC) neanche tramite la *soft evidence* su due stati adiacenti. Quindi tali casi, costituenti i *missing data*, sono trattati come casi particolari di *soft evidence* mediante assegnazione a tutti gli stati dentali di un *belief* pari a $1/5$.

In alcune OPT mancano fisicamente uno o più terzi molari la cui assenza, a detta degli esperti, non costituisce alcuna informazione sull'età del soggetto. Per quanto siano di diversa natura, anche i casi non classificabili sono trattati come *missing*: la mancata classificazione avviene per l'impossibilità di leggere la radiografia la cui resa dipende anche dalla tecnica con cui è realizzata, vale a dire se si tratta di una scansione digitale o di una fotografia analogica.

Di seguito sono presentate, per ciascun esperto, le percentuali medie[1] del numero di OPT che risultano non classificabili, suddivise per tecnologia e per arcata mascellare (D_{12} [2]) o mandibolare (D_{34}).

[1] Per ciascuno dei quattro terzi molari viene calcolata la percentuale di OPT nelle quali il dente risulta non classificabile. Infine si calcola la media di tali percentuali per arcata.

[2] Con le notazioni D_1, D_2, D_3 e D_4 sono stato indicati i terzi molari, rispettivamente, superiore sinistro, superiore destro, inferiore destro ed inferiore sinistro. Per indicare la coppia dentale ij si è usato D_{ij} mentre D_{1234} specificherà tutti e quattro i terzi molari.

Tabella 4.1. *Percentuale del numero di OPT dichiarate non classificabili per arcata e tecnologia - Esperto A*

% OPT NC	D_{12}	D_{34}
Analogica	11,14	2,45
Digitale	3,64	1,82
Totale	9,66	2,33

Tabella 4.2. *Percentuale del numero di OPT dichiarate non classificabili per arcata e tecnologia - Esperto B*

% OPT NC	D_{12}	D_{34}
Analogica	4,01	4,01
Digitale	2,73	1,82
Totale	3,76	3,58

I risultati mostrano come l'Esperto B abbia una capacità di classificazione più alta dell'Esperto A per la coppia dentale superiore (D_{12}) mentre avviene il contrario per l'arcata inferiore (D_{34}), la quale risulta anche essere più facilmente classificabile per entrambi gli esperti. Da notare come la tecnologia digitale faciliti l'osservazione ad entrambi gli esperti.

Da un esame preliminare è emerso come l'utilizzo più o meno intenso che ciascun esperto fa della *soft evidence* trovi corrispondenza con il percorso professionale dei due esperti: "forense" per l'Esperto A, che ha una maggiore esperienza nelle perizie legali riguardanti valutazioni dentali, e "clinico" per l'Esperto B, che invece ha un *background* basato sulle decisioni da affrontare per curare pazienti affetti da patologie odontostomatologiche.

Nella *Tab. 4.*3 e *Tab. 4.4.* sono mostrate le percentuali di OPT[3] sui casi osservati che presentano *hard* o *soft evidence*, suddivise per tecnologia ed esperto.

[3] Una OPT osservata, in cui tutti i terzi molari sono classificati, è considerata hard *evidence* se tutte le quattro valutazioni dentali presentano *hard evidence*, *soft evidence* in caso contrario.

Tabella 4.3. *Percentuale di hard e soft evidence sui dati osservati per tecnologia - Esperto A*

% evidence sui dati osservati	Hard evidence	Soft evidence
Analogica	64,56	35,44
Digitale	55,81	44,19
Totale	62,77	37,23

Tabella 4.4. *Percentuale di hard e soft evidence sui dati osservati per tecnologia - Esperto B*

% evidence sui dati osservati	Hard evidence	Soft evidence
Analogica	87,83	12,17
Digitale	56,67	43,33
Totale	81,26	18,74

Dall'esame delle tabelle emerge che entrambi gli osservatori forniscono valutazioni più "decise", di tipo *hard evidence*, se le OPT osservate utilizzano la tecnologia analogica, anche se questo non significa necessariamente una migliore *performance* dell'attività di classificazione rispetto alla variabile età.

Il primo osservatore, l'Esperto A, avendo una maggiore esperienza in ambito forense, ricorre spesso all'impiego della *soft evidence* (nel 37,23% dei casi totali osservati, vale a dire indipendentemente dalla tecnologia ed escludendo i casi mancanti) sottolineando una linea più "cauta", forse legata alla piena percezione della responsabilità che la perizia richiesta assume in sede legale. Il secondo osservatore, l'Esperto B, il cui percorso professionale è di tipo "clinico", appare più deciso e impiega maggiormente le *hard evidence* adottando le *soft evidence* solamente nel 18,74 % dei casii totali osservati.

4.3 Riproducibilità: i due esperti a confronto

Un aspetto peculiare del giudizio degli osservatori è la differente valutazione che questi possono effettuare a distanza di tempo sulle medesime OPT o quella intercorrente fra diversi osservatori. Questo fa riflettere sull'assunzione di

scambiabilità degli osservatori che hanno provveduto a leggere le OPT ovvero dell'irrilevanza, dal punto di vista inferenziale, di coloro che hanno eseguito le osservazioni.

Il confronto fra i due esperti avviene mediante gli indici di ripetibilità e riproducibilità che chiameremo, in accordo con la letteratura corrente (Maber et al., 2006, Dhanjal et al., 2006; Cameriere et al., 2008), rispettivamente, riproducibilità intra-osservatore e riproducibilità inter-osservatori. Dai dati a disposizione non si sono riscontrate rilevanti differenze fra gli stati indicati dai due esperti, o da uno solo di essi a distanza di tempo, sulle stesse OPT. Ciò ha portato alla scelta di non ponderare i *believes* per il calcolo dell'accordo casuale e quindi nella (3.52) si è utilizzata la (3.58).

Le *Tab*. *4*.5 e *Tab*. *4*.6 mostrano i valori della riproducibilità inter-osservatori, rispettivamente, non depurata e depurata dal caso, suddivisa per terzo molare e tecnologia.

Tabella 4.5. *Riproducibilità inter-osservatori per terzo molare e tecnologia*

Riproducibilità inter-osservatori	D_1	D_2	D_3	D_4
Analogica	0,659	0,673	0,688	0,738
Digitale	0,622	0,65	0,668	0,648
Totale	0,651	0,668	0,685	0,722

Tabella 4.6. *Riproducibilità inter-osservatori depurata dal caso per terzo molare e tecnologia*

Riproducibilità inter-osservatori depurata	D_1	D_2	D_3	D_4
Analogica	0,531	0,552	0,588	0,651
Digitale	0,488	0,525	0,558	0,538
Totale	0,522	0,546	0,583	0,63

Sebbene i valori della riproducibilità inter-osservatori presenti nelle *Tab*. *4*.5 e *Tab*. *4*.6 possano sembrare non soddisfacenti, secondo Landis e Koch (1977) sono "sostanziali" per quanto riguarda il terzo molare inferiore sinistro (D_4) e "moderati" per i rimanenti terzi molari in base alla loro classificazione.

Figura 4.1. *Classificazione dei valori della Kappa di Cohen (Landis e Koch, 1977, p.165)*

Kappa Statistic	*Strength of Agreement*
<0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost Perfect

Inoltre non va dimenticato come la *soft evidence* induca una riduzione della riproducibilità rispetto al caso in cui sia richiesto agli osservatori di fornire esclusivamente *hard evidence*. Basterebbe pensare, ad esempio, il caso in cui su di una OPT i due osservatori indichino gli stessi due stati adiacenti ma con *believes* differenti: 0,80 e 0,20 per il primo osservatore e 0,70 e 0,30 per il secondo. Questo significa che la divergenza inter-osservatori per tale OPT, misurata con la (3.48), è $S = 0.2$ a differenza di $S = 0$ nel caso *hard* in cui l'attribuzione sarebbe ricaduta sullo stato con maggior fiducia.

I terzi molari dell'arcata inferiore (D_{34}) presentano un grado di riproducibilità inter-osservatore più alto rispetto a quelli dell'arcata superiore (D_{12}) e inoltre risulta una maggiore corrispondenza fra le valutazioni dei due esperti se la tecnologia impiegata è di tipo analogico.

I risultati riportati nelle *Tab. 4.*5 e *Tab. 4.*6 riguardanti la riproducibilità inter-osservatori suggeriscono una rilevante presenza dell'incertezza nel fornire una valutazione sullo stato di maturazione dentale secondo la classificazione di Demirijan da parte degli esperti. Ciò rafforza la convinzione che è necessario permettere ad un osservatore di fare uso della *soft evidence* per poter meglio formalizzare la propria valutazione sui singoli terzi molari.

Per quanto riguarda la riproducibilità intra-osservatore sono stati estratti casualmente, per ciascun esperto, due sottoinsiemi di 33 OPT digitali e 44 analogiche. Tali casi sono stati sottoposti agli esperti per una seconda valutazione a distanza di un mese e le risultanti valutazioni sono state confrontate con quelle inizialmente fornite sulle stesse OPT. Anche in questo caso sono state escluse dal calcolo degli indici di riproducibilità intra-osservatore le OPT NA ed NC comuni alle due misurazioni messe a confronto.

Le tabelle dalla *Tab. 4.*7 alla *Tab. 4.*10 mostrano i valori dell'indice di riproducibilità non depurato e depurato dal caso, suddiviso per terzo molare, tecnologia ed esperto.

Tabella 4.7. *Riproducibilità intra-osservatore per terzo molare e tecnologia - Esperto A*

Riproducibilità intra-osservatore	D_1	D_2	D_3	D_4
Analogica	0,844	0,814	0,842	0,857
Digitale	0,776	0,79	0,867	0,865
Totale	0,817	0,804	0,851	0,86

Tabella 4.8. *Riproducibilità intra-osservatore depurata dal caso per terzo molare e tecnologia – Esperto A*

Riproducibilità intra-osservatore depurata	D_1	D_2	D_3	D_4
Analogica	0,770	0,719	0,759	0,787
Digitale	0,698	0,723	0,809	0,788
Totale	0,741	0,721	0,778	0,788

Tabella 4.9. *Riproducibilità intra-osservatore per terzo molare e tecnologia - Esperto B*

Riproducibilità intra-osservatore	D_1	D_2	D_3	D_4
Analogica	0,689	0,719	0,714	0,829
Digitale	0,754	0,673	0,759	0,791
Totale	0,714	0,700	0,731	0,815

Tabella 4.10. *Riproducibilità intra-osservatore depurata dal caso per terzo molare e tecnologia - Esperto B*

Riproducibilità intra-osservatore depurata	D_1	D_2	D_3	D_4
Analogica	0,627	0,631	0,621	0,772
Digitale	0,668	0,578	0,658	0,666
Totale	0,643	0,609	0,635	0,732

La riproducibilità intra-osservatore risulta in generale più marcata per l'Esperto A, indipendentemente dalla tecnologia utilizzata e dal terzo molare osservato. Poiché l'Esperto A fa un maggiore uso della *soft evidence* si potrebbe pensare che l'utilizzo della stessa migliori la possibilità che un osservatore ripeta la medesima valutazione nel tempo e per la stessa OPT, contrariamente a quanto attteso. Comunque entrambi gli esperti riescono più facilmente a riprodurre la propria valutazione qualora si considerino i terzi molari dell'arcata inferiore (D_{34}) ed una tecnologia digitale.

Un'altra osservazione che riguarda sia la riproducibilità inter che intra-osservatore, sia nel caso digitale che analogico, consiste nel fatto che l'indice stesso presenti, per i casi analizzati, valori crescenti a partire dal dente sinistro dell'arcata superiore (D_1) per poi proseguire in senso orario fino a quello sinistro dell'arcata inferiore (D_4). Un'interpretazione di tale risultato potrebbe ricercarsi nel fatto che una OPT viene generalmente letta in questo stesso ordine, ovvero in senso orario dal dente sinistro dell'arcata superiore (D_1). Dunque, l'osservatore sembrerebbe prendere confidenza con l'OPT, vale a dire le indecisioni di valutazione sulla medesima OPT si riducono man mano che l'esperto procede verso il dente sinistro dell'arcata inferiore (D_4). Tale confidenza è evidentemente legata alla qualità e tecnologia radiografica impiegata. Infatti, bisogna ricordare che la tecnologia delle OPT è stata suddivisa per semplicità in analogica e digitale, ma in realtà arriva ad un livello di differenziazione più dettagliato.

4.4 Classificazione dell'età

4.4.1 La procedura di learning parametrico: risultati

Ogni esperto effettua la propria valutazione relativamente ai quattro terzi molari di ciascuna OPT del campione. La variabile di classificazione età è stata suddivisa prima in due e poi in tre classi opportunamente costruite in funzione delle soglie di età di interesse. Dal campione a disposizione, per ogni esperto, si è estratto casualmente una *training data set* di $n = 447$ unità (l'80% delle osservazioni complessive) stratificato per genere, età ed e*vidence* (*hard*, *soft*, *missing*). Successivamente si utilizza, separatamente per ciascun terzo molare, la procedura di *learning* parametrico descritta nel *Cap. 3*. Per ogni terzo molare, condizionatamente a ciascuna classe di età e a ciascuna categoria del genere, si ricava la verosimiglianza polinomiale (3.17) che si ottiene a partire dalla matrice dei *believes* condizionati (3.11) del *training data set*.

La maggiore complessità computazionale affrontata in questa ricerca è stata la costruzione di questo polinomio per l'elevata numerosità dei termini che lo compongono.

Si consideri, ad esempio, il caso in cui l'età sia stata dicotomizzata e si utilizzi come matrice dei *believes* condizionati (3.11) quella fornita dall'Esperto A relativamente al terzo molare D_4, e le valutazioni sono relative ad individui maschi ($s=1$) maggiorenni ($q=1$). Tale matrice è composta da $n_{11}=190$ righe e quindi per la (3.15) il numero dei polinomi della verosimiglianza (3.12) possono arrivare ad un massimo di $\bar{m}_{4|11}=5^{190}=6.372\cdot 10^{132}$. Nel caso specifico fra le $n_{11}=190$ osservazioni 149 sono di tipo *hard evidence*, 17 *soft evidence* e 24 *missing data* allora il numero effettivo di polinomi che derivano dalla produttoria (3.12) è $\hat{m}_{4|11}=1^{149}\cdot 2^{17}\cdot 5^{24}=7.81\cdot 10^{21}$.

In realtà, poiché la verosimiglianza su di una singola osservazione missing per la (3.10) corrisponde a $1/k$, il cui valore si semplifica nella (3.22), allora si può ottenere la verosimiglianza polinomiale a meno dei *missing data* in quanto ininfluenti. A partire, dunque, da $\hat{m}_{4|11}=2^{17}=131.072$ polinomi, sommando fra loro quelli con la stessa base si ottiene la matrice delle potenze delle diverse basi del polinomio (3.17), $\mathbf{P}_{4|11}$ a meno dei casi mancanti. Nell'esempio citato il numero di polinomi con diversa base che caratterizzano la verosimiglianza (3.17) si è drasticamente ridotto a $\tilde{m}_{4|11}=122$.

Una volta che per ciascun esperto e terzo molare è stata ricavata la verosimiglianza a struttura polinomiale condizionatamente alla classe di età e al genere, si ottiene la predittiva (3.37) assumendo che i parametri che caratterizzano la distribuzione degli attributi condizionati, $\boldsymbol{\theta}_{h|qs}$, seguono una distribuzione a priori Dirichlet non informativa, e più specificatamente con iperparametri $\boldsymbol{\alpha}_{h|qs}=(1,1,1,1,1)\ \ \forall h,q,s$. Le unità che non sono state estratte nel *training set* costituiranno il *test data set* di $n=112$ unità. Per ciascuna unità del *test set* si applicherà la predittiva (3.37), appresa dal *training set*, per effettuare la classificazione in base alla regola decisionale (3.38) o (3.39) e confrontare l'assegnazione ottenuta con la classe di età osservata per tale unità. L'intera procedura è stata replicata estraendo nuovamente un *training data set* per $R=1.000$ replicazioni. Gli indici di *performance* (3.46) e (3.47) e le percentuali di classificazione presentate nei successivi paragrafi vanno dunque intesi come valori medi calcolati sulle 1.000 replicazioni.

4.4.2 Caso dicotomico

Nel caso ricorrente in cui si voglia classificare un soggetto per età come maggiorenne o minorenne si ponga la soglia $\tau_1=18$ e sia la variabile di classe

$C = \{c_0, c_1\}$ con $c_0 = \{t : t < 18\}$ e $c_1 = \{t : t \geq 18\}$. La regola decisionale utilizzata è la (3.40), per cui un soggetto viene assegnato alla classe dei maggiorenni solo se la probabilità predittiva della classe di età di appartenenza non è inferiore alla soglia probabilistica di classificazione π.

Le *Tab. 4.11* e *Tab. 4.12* mostrano la percentuale di individui correttamente classificati per esperto, essendo l'evidenza dentale fornita sui singoli terzi molari (D_1, D_2, D_3, D_4), sulle coppie ($D_{12}, D_{13}, D_{14}, D_{23}, D_{24}, D_{34}$) e su tutti e quattro i denti (D_{1234}) avendo assunto varie soglie probabilistiche nel *range* $0,50 \leq \pi \leq 0,99$.

Tabella 4.11. *Percentuale di individui correttamente classificati per combinazione dentale e alcune soglie probabilistiche π - Esperto A*

Soglia	D_1	D_2	D_3	D_4	D_{12}	D_{34}	D_{14}	D_{23}	D_{13}	D_{24}	D_{1234}
0,50	80,8	80,0	80,9	81,8	78,9	80,6	82,1	81,5	82,3	81,5	80,8
0,70	79,1	78,3	76,4	77,4	78,7	79,1	80,2	79,1	80,0	79,5	79,8
0,80	67,8	64,9	66,9	65,1	73,4	75,1	76,7	73,6	76,4	74,7	78,2
0,90	59,6	61,7	58,3	57,9	68,3	64,7	69,0	68,4	68,9	69,5	74,3
0,95	31,6	47,5	44,0	48,5	62,4	60,7	59,0	61,3	59,4	61,9	71,1
0,99	27,7	27,7	27,7	27,7	44,3	49,2	41,2	50,3	39,7	48,8	63,9

Tabella 4.12. *Percentuale di individui correttamente classificati per combinazione dentale e alcune soglie probabilistiche π - Esperto B*

Soglia	D_1	D_2	D_3	D_4	D_{12}	D_{34}	D_{14}	D_{23}	D_{13}	D_{24}	D_{1234}
0,50	77,9	78,8	79,2	81,1	78,1	79,8	81,1	79,0	79,3	80,1	79,3
0,70	74,3	76,5	74,4	77,7	75,7	78,5	78,1	77,5	76,9	77,9	77,5
0,80	69,2	67,9	68,4	71,9	73,6	76,3	75,5	75,3	74,5	75,7	76,9
0,90	58,8	59,1	52,0	42,8	65,9	68,9	68,9	68,6	70,0	67,7	75,4
0,95	28,9	30,3	29,7	33,4	57,3	54,7	56,8	57,0	57,0	57,9	72,1
0,99	27,7	27,7	27,7	27,7	30,9	30,5	29,8	29,4	29,4	30,8	60,0

Confrontando le *Tab. 4.11* e *Tab. 4.12* si osserva come il modello di classificazione stimato sulle letture fornite dall'Esperto A conduca ad una maggiore percentuale di corretta classificazione rispetto al modello stimato sulle valutazioni fornite dall'Esperto B. Inoltre, si osserva, come è naturale aspettarsi, che al crescere della soglia probabilistica di classificazione π il numero di individui correttamente classificati diminuisce. Questo potrebbe indurre a ritenere che la soglia ottimale, tale da massimizzare la corretta classificazione, sia 0,50. Al contrario, come specificato nel *par. 3.7.1*, bisogna tener conto anche degli errori di classificazione commessi e calcolare, del caso la variabile di classificazione sia dicotomica, gli indici di sensibilità e specificità.

Sotto questo riguardo le *Tab. 4.13* e *Tab. 4.14* presentano le percentuali di minorenni erroneamente classificati per esperto, soglia probabilistica ed evidenza fornita sulle varie combinazioni dentali.

Tabella 4.13. *Percentuale dei minorenni erroneamente classificati per combinazione dentale e alcune soglie probabilistiche π - Esperto A*

Soglia	D_1	D_2	D_3	D_4	D_{12}	D_{34}	D_{14}	D_{23}	D_{13}	D_{24}	D_{1234}
0,50	44,6	44,4	51,6	46,3	32,0	35,5	30,6	32,1	33,2	32,2	25,8
0,70	27,0	26,4	30,4	28,4	24,5	26,2	24,6	23,3	23,5	23,1	22,5
0,80	16,3	8,3	13,4	13,1	18,3	20,7	20,2	15,3	19,1	16,0	20,4
0,90	8,7	6,7	5,3	5,8	12,4	8,9	12,2	10,0	10,9	11,0	14,3
0,95	0,6	1,8	2,1	1,8	8,2	6,0	6,4	6,7	6,1	6,0	11,4
0,99	0	0	0	0	1,5	2,3	1,4	1,8	1,9	1,8	7,5

Tabella 4.14. *Percentuale dei minorenni erroneamente classificati per combinazione dentale e alcune soglie probabilistiche π - Esperto B*

Soglia	D_1	D_2	D_3	D_4	D_{12}	D_{34}	D_{14}	D_{23}	D_{13}	D_{24}	D_{1234}
0,50	55,8	50,4	49,1	44,0	34,6	32,1	32,9	34,3	36,2	34,3	28,1
0,70	25,3	28,8	26,5	28,7	26,9	27,6	26,9	23,0	24,7	26,2	24,5
0,80	16,3	17,5	17,4	23,1	19,6	23,0	23,8	19,6	19,1	23,7	22,8
0,90	8,5	7,8	7,6	3,4	12,5	17,9	14,,0	13,8	12,5	14,8	19,3
0,95	0,5	0,8	0,6	0,9	7,9	8,8	8,2	7,4	7,4	7,7	14,2
0,99	0	0	0	0	1	0,6	0,3	0,9	0,6	0,5	8,2

Dalle *Tab. 4.13* e *Tab. 4.14* risulta che una soglia probabilistica $\pi = 0,50$ conduce ad una percentuale dei minorenni erroneamente classificati non accettabile e che all'aumentare della soglia probabilistica π tale percentuale diminuisce. Anche da questo punto di vista il modello classificatoria stimato sulle valutazioni dell'Esperto A risulta migliore di quello dell'Esperto B.

Il complemento a 1 dei valori presentati nelle *Tab. 4.13* e *Tab. 4.14* corrisponde alla specificità di ciascun modello classificatorio, vale a dire la proporzione di minorenni correttamente classificati. E' perciò evidente che per una soglia probabilistica $\pi = 0,99$ tutti i minorenni sono correttamente classificati. Tale risultato si può anche leggere nelle *Tab. 4.11* e *Tab. 4.12* in cui i modelli che utilizzano le evidenze derivanti dai singoli terzi molari, per una soglia probabilistica $\pi = 0,99$, classificano tutti i soggetti come minorenni. La percentuale di corretta classificazione corrispondente è dunque pari alla percentuale dei minorenni presenti nel *test data set*, pari a 27,7% in quanto stratificato per età.

La sensitività, invece, rappresenta la capacità del modello di classificare correttamente un maggiorenne, vale a dire la percentuale di maggiorenni correttamente classificati. Quindi, la scelta della soglia probabilistica π e dell'evidenza dentale da impiegare sarà da ricercarsi rispetto alla massimizzazione congiunta della specificità e della sensitività.

Purtroppo il trade-off esistente fra i due indici rende non semplice la soluzione del problema. Si potrebbero mettere a confronto le sensitività e specificità degli 11 modelli classificatori proposti, a seconda dell'evidenza dentale utilizzata, mediante l'impiego delle curve ROC accennate nel *Cap. 3*.

L'indice $1 - specificità$ utilizzato nelle curve ROC e che corrisponde ai valori presenti nelle *Tab. 4.13* e *Tab. 4.14*, rappresenta la proporzione dei *falsi maggiorenni*, i quali, in ambito legale, hanno un peso non irrilevante.

Siano, dunque, presi in considerazione i 3 modelli che utilizzano l'evidenza D_4, D_{34} e D_{1234} poiché ritenuti, in molti lavori, migliori ai fini classificatori (Olze et al., 2004; Pinchi et al., 2005; Dhanjal et al., 2006; Cameriere et al., 2008).

Le *Tab. 4.15* e *Tab. 4.16* presentano, per ciascun esperto, i valori della sensibilità e 1-specificità dei 3 modelli al variare delle soglie probabilistiche π, rappresentati nelle corrispondenti *Fig. 4.2* e *Fig. 4.3*.

Tabella 4.15. *Sensibilità e 1-specificità per alcune soglie probabilistiche π in 3 modelli classificatori - Esperto A*

Soglia	sensibilità			1-specificità		
	D_4	D_{34}	D_{1234}	D_4	D_{34}	D_{1234}
0,50	92,6	86,8	83,3	46,3	35,5	25,8
0,70	79,6	81,1	80,7	28,4	26,2	22,5
0,80	56,8	73,5	77,7	13,1	20,7	20,4
0,90	44,0	54,6	69,9	5,8	8,9	14,3
0,95	29,5	48,0	64,4	1,8	6,0	11,4
0,99	0	30,6	53,0	0	2,3	7,5

Figura 4.2. *Rappresentazione grafica della sensibilità e 1-specificità per soglia probabilistica π in 3 modelli classificatori - Esperto A*

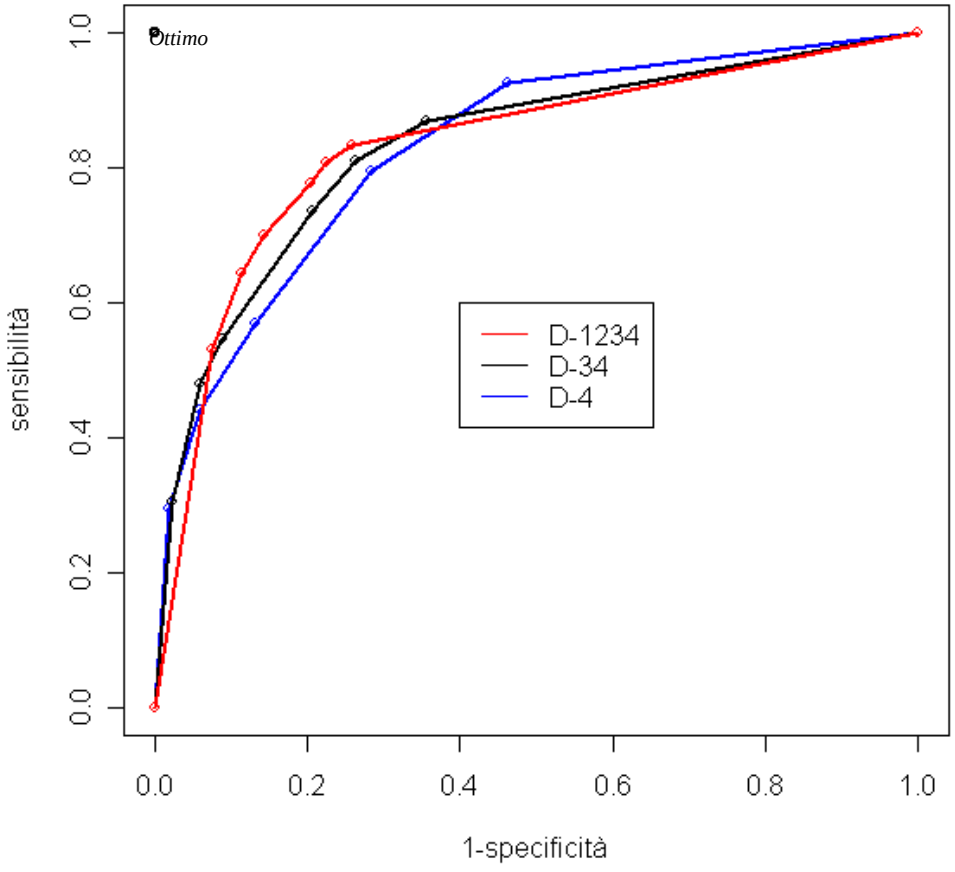
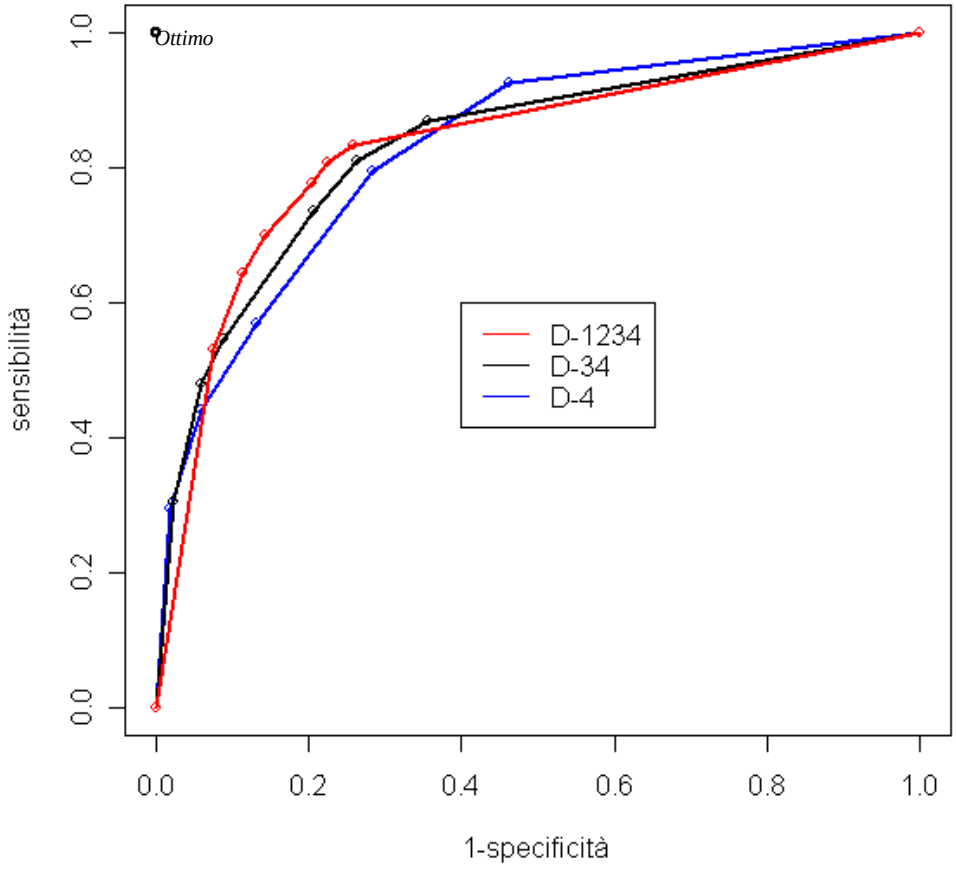


Tabella 4.16. *Sensibilità e 1-specificità per alcune soglie probabilistiche π in 3 modelli classificatori - Esperto B*

Soglia (B)	sensibilità			1-specificità		
	D_4	D_{34}	D_{1234}	D_4	D_{34}	D_{1234}
0,50	90,7	84,4	82,1	44	32,1	28,1
0,70	80,1	80,8	78,3	28,7	27,6	24,5
0,80	70	76	76,8	23,1	23	22,8
0,90	22,2	63,8	73,4	3,4	17,9	19,3
0,95	8,3	44,5	66,9	0,9	8,8	14,2
0,99	0	4,1	47,8	0	0,6	8,2

Figura 4.3. *Rappresentazione grafica della sensibilità e 1-specificità per soglia probabilistica π in 3 modelli classificatori - Esperto B*



Come già detto in precedenza il modello stimato sull'Esperto A produce migliori risultati si in termini di sensibilità che di specificità rispetto al modello stimato sulle valutazioni dell'Esperto B. Nelle *Fig. 4.2* e *Fig. 4.3* sono stati aggiunti per maggior chiarezza i punti estremi (0;0) e (1;1) fatta eccezione, per entrambi gli esperti, del caso (0;0) presente nel modello che utilizza l'evidenza derivante dal terzo molare inferiore sinistro (D_4) per una soglia probabilistica $\pi = 0,99$. Inoltre la situazione ottimale, vale a dire sensibilità e specificità massime, è rappresentata dal punto (1;0).

I grafici mostrano come il modello classificatorio basato sull'evidenza derivante da tutti e quattro i terzi molari (D_{1234}) produca risultati migliori rispetto al modello che utilizza solo le evidenze derivanti dai terzi molari dell'arcata inferiori (D_{34}) che a sua volta produce una migliore sensibilità e specificità rispetto al modello classificatorio che impiega la sola evidenza del terzo molare inferiore sinistro (D_4). La soglia probabilistica di classificazione π cresce a partire dal punto (1;1) fino al punto (0;0): i valori a cui sembra corrispondere una migliore *performance* classificatoria, relativamente al punto di ottimo, sembrano variare fra $\pi = 0,70$ e $\pi = 0,80$.

I risultati finora presentati, nel caso l'età sia dicotomica, mostrano come le percentuali di corretta classificazione e di errore dipendano fortemente dalla soglia probabilistica di classificazione π. Poiché i risultati ottenuti non sono del tutto soddisfacenti, si è interessati ad esplorare i dati introducendo un'ulteriore classe per la variabile età.

4.4.3 Caso tricotomico

Il cambio di *status* da minorenni a maggiorenni ovviamente non corrisponde ad un cambio altrettanto repentino nello sviluppo dentale dei terzi molari.

Tale considerazione mette in dubbio l'efficacia della classificazione per età utilizzando solo due classi a causa della difficoltà derivante dall'attribuzione di una classe di età per individui di età prossime alla soglia dei 18 anni.

Una possibile soluzione consiste nell'introduzione di una terza classe centrale di "non attribuzione dell'età" a cavallo di tale soglia. Questo significa che gli individui che il modello classificatorio attribuisce alla classe centrale, in realtà, richiedono un'analisi più approfondita per poter esprimere un giudizio sulla loro età. In questo modo ci si aspetta che alcuni dei soggetti erroneamente classificati nel caso dicotomico possano ricadere nella classe di "non attribuzione dell'età", migliorando conseguentemente le *performance* nelle due classi esterne.

Siano dunque le soglie di età $\tau_1 = 17$ e $\tau_2 = 19$, la variabile di classe $C = \{c_0, c_1, c_2\}$ con $c_0 = \{t : t < 17\}$, $c_1 = \{t : 17 \le t < 19\}$ e $c_2 = \{t : t \ge 19\}$ e la regola decisionale adottata la (3.38).

Le *Tab. 4.17* e *Tab. 4.18* mostrano le percentuali di classificazione dei soggetti appartenenti alla classe di età $c_0 = \{t : t < 17\}$, per esperto e tecnologia.

Tabella 4.17. *Percentuali di classificazione per individui di età* $t < 17$ *- Esperto A*

Individui di età $t < 17$	Classificati in		
	c_0 (correttamente)	c_1 (da valutare)	c_2 (erroneamente)
Digitale	72,1	22,7	5,3
Analogica	81,3	16,9	1,7
Totale	79,1	18,4	2,5

Tabella 4.18. *Percentuali di classificazione per individui di età* $t < 17$ *- Esperto B*

Individui di età $t < 17$	Classificati in		
	c_0 (correttamente)	c_1 (da valutare)	c_2 (erroneamente)
Digitale	76,1	19,4	4,5
Analogica	74,7	17,8	7,5
Totale	75,0	18,1	6,9

Le *Tab. 4.17* e *Tab. 4.18* mostrano che l'Esperto A è in grado di classificare correttamente una percentuale maggiore, rispetto all'Esperto B, di soggetti minorenni (con età $t < 17$) specie se la tecnologia radiografica è di tipo analogico, mentre l'Esperto B ottiene migliori risultati con la tecnologia digitale. I valori nelle ultime colonne corrispondono alle percentuali di errata classificazioni in cui un soggetto di età $t < 17$ viene attribuito alla classe di età $t \ge 19$, vale a dire la proporzione di falsi maggiorenni prodotta da ciascun modello classificatorio. Nel caso dell'Esperto A, qualora venisse utilizzata una tecnologia analogica, si produrrebbe una percentuale di falsi maggiorenni pari

al 2,5% mentre per l'Esperto B si raggiunge il suo migliore risultato (4,5%) con un tecnologia digitale.

Inoltre, entrambi gli esperti producono una "non attribuzione di età" in circa il 18% degli individui appartenenti alla classe di età $c_0 = \{t : t < 17\}$, indipendentemente dalla tecnologia.

Le *Tab. 4.19* e *Tab. 4.20* presentano le percentuali di classificazione dei soggetti appartenenti alla classe di età $c_1 = \{t : 17 \le t < 19\}$, per esperto e tecnologia.

Tabella 4.19. *Percentuali di classificazione per individui di età* $17 \le t < 19$ *- Esperto A*

Individui di età $17 \le t < 19$	Classificati in		
	c_0 (erroneamente)	c_1 (correttamente)	c_2 (erroneamente)
Digitale	31,0	37,8	31,2
Analogica	26,4	35,8	37,8
Totale	27,2	36,0	36,9

Tabella 4.20. *Percentuali di classificazione per individui di età* $17 \le t < 19$ *- Esperto B*

Individui di età $17 \le t < 19$	Classificati in		
	c_0 (erroneamente)	c_1 (correttamente)	c_2 (erroneamente)
Digitale	41,4	22,8	35,8
Analogica	26,9	29,2	43,9
Totale	29,0	28,3	42,7

La classificazione degli individui di età $17 \le t < 19$ comporta una maggiore variabilità nell'attribuzione di una classe d'età e tale difficoltà discriminatoria giustifica così l'introduzione della terza classe centrale. L'Esperto A sembrerebbe non ottenere migliori risultati con una tecnologia o l'altra a differenza dell'Esperto B, il cui modello classificatorio produce una *performance* migliore se l'OPT è analogica. Fra i due modelli classificatori, ancora una volta quello stimato sull'Esperto A risulta migliore. Infatti questo

modello classifica correttamente, per quanto tali soggetti debbano essere sottoposti ad un'ulteriore valutazione, una percentuale più alta rispetto a quella derivante dal modello stimato sull'Esperto B ed è minore anche la percentuale di individui erroneamente attribuita alle due classi di età esterne.

Infine si valutino le percentuali di classificazione dei soggetti appartenenti alla classe di età $c_2 = \{t : t \geq 19\}$. Le *Tab. 4.21* e *Tab. 4.22* presentano tali percentuali per esperto e tecnologia.

Tabella 4.21. *Percentuali di classificazione per individui di età* $t \geq 19$ *- Esperto A*

Individui di età $t \geq 19$	Classificati in		
	c_0 (erroneamente)	c_1 (da valutare)	c_2 (correttamente)
Digitale	5,9	13,9	80,2
Analogica	7,6	19,9	72,5
Totale	7,2	18,5	74,3

Tabella 4.22. *Percentuali di classificazione per individui di età* $t \geq 19$ *- Esperto B*

Individui di età $t \geq 19$	Classificati in		
	c_0 (erroneamente)	c_1 (da valutare)	c_2 (correttamente)
Digitale	8,0	13,5	78,6
Analogica	7,9	18,9	73,2
Totale	7,9	17,7	74,5

Anche nel caso di individui di età $t \geq 19$ il modello di classificazione basato sull'Esperto A produce risultati migliori rispetto a quello relativo all'Esperto B ed entrambi gli esperti ottengono migliori *performance* con la tecnologia digitale. Per simmetria con le *Tab. 4.17* e *Tab. 4.18*, la prima colonna delle *Tab. 4.21* e *Tab. 4.22* corrisponde alle percentuali di individui di età $t \geq 19$ classificati nella classe $c_0 = \{t : t < 17\}$, vale a dire la proporzione di falsi minorenni.

Le percentuali di falsi maggiorenni e falsi minorenni prodotte da entrambi gli esperti sono piuttosto contenute e tali da rendere attraente l'idea di utilizzare una terza classe di età a cavallo della soglia di interesse e l'evidenza derivante da tutti e quattro i terzi molari per l'attribuzione di individui viventi non adulti a classi di età.

.Nel caso dicotomico l'alternativa ai modelli di classificazione finora proposti è legata alla regola decisionale (3.39) che fa uso della soglia probabilistica di classificazione π. Poiché tale vincolo comporta la possibilità che alcuni soggetti non siano attribuiti ad alcuna classe di età, allora all'aumentare della soglia probabilistica di classificazione π diminuirà la capacità classificatoria CC_π (3.47) del modello ed aumenterà, conseguentemente, la probabilità che il modello non fornisca alcuna classificazione, $P(NC_\pi)$.

A tal proposito siano messi a confronto i due modelli stimati sugli esperti: nella *Tab. 4.23* sono presentate, per varie soglie probabilistiche π, le probabilità di non classificazione e le probabilità di produrre falsi maggiorenni $P(FM_\pi)$ nel caso l'individuo superi la soglia. A seguire le corrispondenti *Fig. 4.4* e *Fig. 4.5* che rappresentano l'andamento, rispettivamente, della probabilità $P(NC_\pi)$ e $P(FM_\pi)$ in funzione delle soglie probabilistiche π.

Tabella 4.23. *Valori delle probabilità di non classificazione* $P(NC_\pi)$ *e di produrre falsi minorenni* $P(FM_\pi)$ *per esperto e soglia probabilistica* π

Soglia π	Esperto A		Esperto B	
	$P(NC_\pi)$	$P(FM_\pi)$	$P(NC_\pi)$	$P(FM_\pi)$
0,50	0,0296	0,0048	0,0512	0,0108
0,55	0,0877	0,0037	0,1513	0,0111
0,60	0,1496	0,0026	0,2415	0,0107
0,65	0,2117	0,0023	0,3226	0,0085
0,70	0,2707	0,0024	0,4011	0,0068
0,75	0,3317	0,0026	0,4798	0,0059
0,80	0,3932	0,0029	0,5399	0,0050
0,85	0,4678	0,0033	0,5978	0,0045
0,90	0,5626	0,0040	0,6872	0,0014
0,95	0,7304	0,0004	0,8147	0,0000
0,99	0,9669	0,0000	0,9861	0,0000

Figura 4.4. *Rappresentazione grafica della probabilità di non classificazione* $P(NC_\pi)$ *per esperto e soglia probabilistica* π

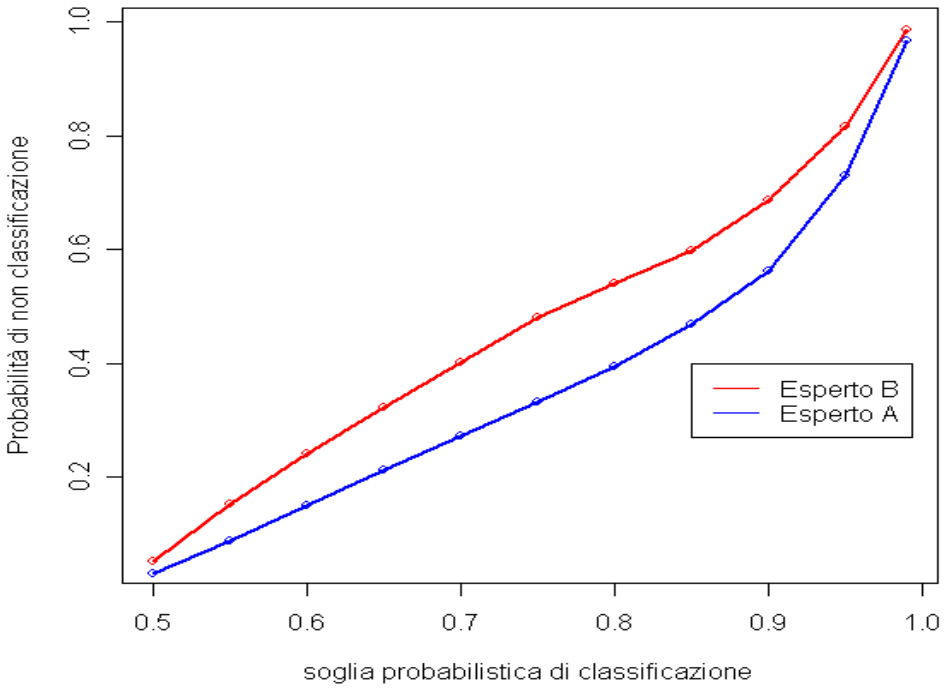
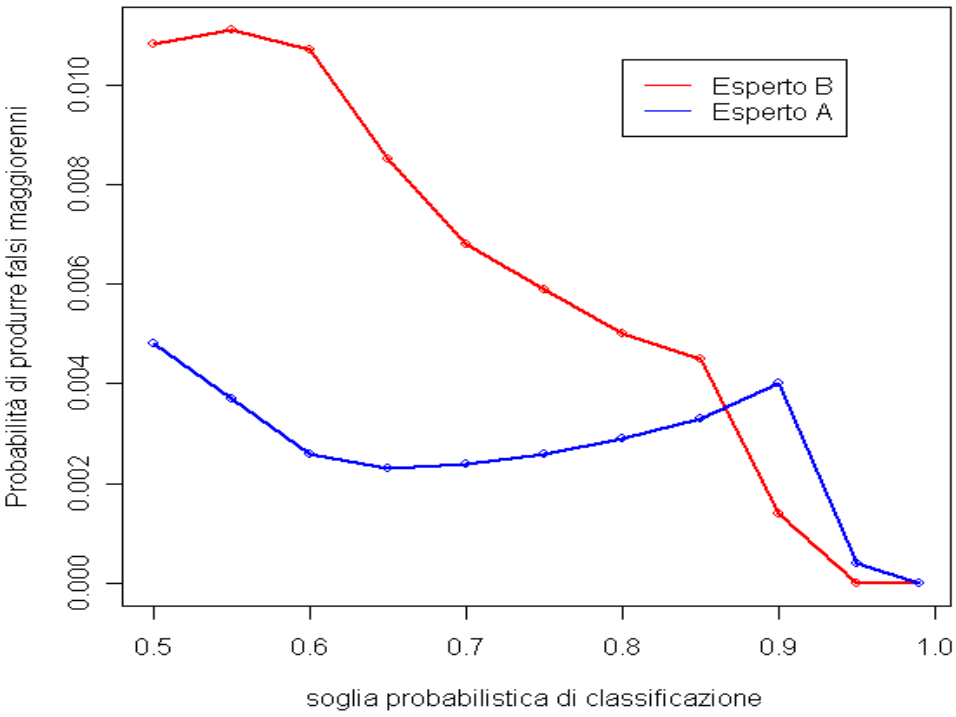


Figura 4.5. *Rappresentazione grafica della probabilità di produrre falsi maggiorenni* $P(FM_\pi)$ *per esperto e soglia probabilistica* π



Nella *Fig. 4.4* si osserva, come ci si poteva aspettare, un legame diretto approssimativamente lineare fra la soglia probabilistica di classificazione π e la probabilità di non classificazione $P(NC_\pi)$, in quanto, al crescere del vincolo probabilistico è sempre più difficile determinare una predittiva che lo soddisfi.

Anche nella *Fig. 4.5* non sorprende l'andamento negativo fra la soglia probabilistica π e la probabilità di produrre un falso negativo qualora l'individuo venisse classificato, proprio in merito alle considerazioni fatte sulle *Tab. 4.13* e *Tab. 4.14*.

Infine, è interessante l'osservazione congiunta della *Fig. 4.4* e *Fig. 4.5* per confrontare i due modelli. A parità di soglia probabilistica, il modello di classificazione stimato sulle valutazione fornite dall'Esperto A produce *performance* nettamente migliori rispetto al modello relative all'Esperto B. Per quanto riguarda la probabilità di produrre falsi negativi secondo l'Esperto A, che varia fra il 2‰ e 5‰, essa sembrerebbe essere quasi indipendente dalla soglia probabilistica a differenza di quanto accade per il modello secondo l'Esperto B. Questo significa che nel modello stimato sull'Esperto A si può scegliere una soglia probabilistica $\pi = 0,50$ a cui corrisponde una piccola percentuale di soggetti non classificati (2,96%). Il modello stimato sulle valutazioni dell'Esperto B, per riuscire ad ottenere gli stessi risultati, vale a dire una probabilità di produrre falsi maggiorenni intorno al 5‰, dovrebbe utilizzare una soglia probabilistica $\pi = 0,85$ la quale implica, però, un alta percentuale di soggetti non classificati prossima al 60%.

4.5 Caso reale con modello di transizione

A titolo esemplificativo si prenda un caso da esaminare, l'$(n+1)-esimo$ soggetto. Si supponga di dover fornire un giudizio sull'età di un individuo di genere femminile che non disponga di un regolare documento di riconoscimento. Supponiamo che sia interpellato l'Esperto A il quale, per meglio valutare lo stato di maturazione dei terzi molari, ricorra ad una OPT digitale. Si supponga, inoltre, che l'esperto A indichi lo stato *H* per il terzo molare superiore sinistro (D_1), gli stati *E* ed *F* per quello superiore destro (D_2), nessuno stato per il terzo molare inferiore destro (D_3) che si supponga essere assente ed infine lo stato G per quello inferiore sinistro (D_4). I *believes* forniti dall'Esperto A per ciascuno dei 4 terzi molari siano dunque:

$$\mathbf{b}_1^{n+1} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix},\ \mathbf{b}_2^{n+1} = \begin{bmatrix} 0 \\ 0,6 \\ 0,4 \\ 0 \\ 0 \end{bmatrix},\ \mathbf{b}_3^{n+1} = \begin{bmatrix} 0,2 \\ 0,2 \\ 0,2 \\ 0,2 \\ 0,2 \end{bmatrix},\ \mathbf{b}_4^{n+1} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}\ .$$

Si applichi, adesso, il modello di transizione dei *believe*s, per l'Esperto A, caratterizzato dalla probabilità di transizione (3.60). Utilizzando le riproducibilità intra-osservatore locali e depurate dal caso presenti nella *Tab. 4.8* in corrispondenza della tecnologia digitale, si ricavano i *believes* modificati (3.61):

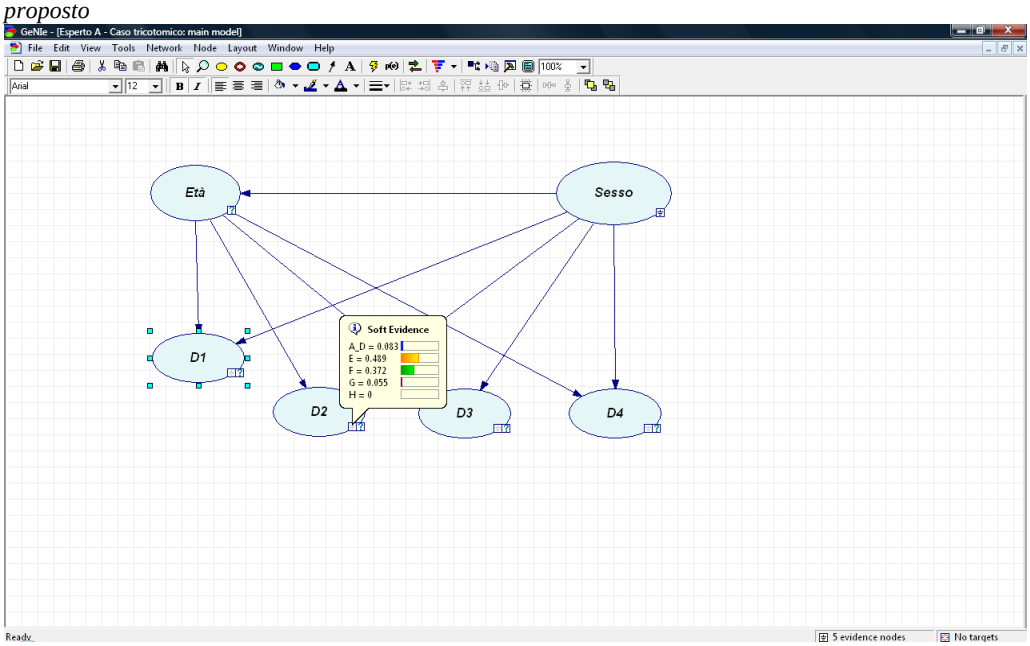
$$\mathbf{b}_1^{(n+1)^*} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0,151 \\ 0,849 \end{bmatrix},\ \mathbf{b}_2^{(n+1)^*} = \begin{bmatrix} 0,0831 \\ 0,4892 \\ 0,3723 \\ 0,0554 \\ 0 \end{bmatrix},\ \mathbf{b}_3^{(n+1)^*} = \begin{bmatrix} 0,2 \\ 0,2 \\ 0,2 \\ 0,2 \\ 0,2 \end{bmatrix},\ \mathbf{b}_4^{(n+1)^*} = \begin{bmatrix} 0 \\ 0 \\ 0,106 \\ 0,788 \\ 0,106 \end{bmatrix}\ .$$

Si supponga che l'Esperto A possa disporre di un *software* apposito che "propaghi" l'evidenza, racchiusa nei *believes* fornita sugli stati dei quattro terzi molari e sul genere, alla variabile di classe età non osservata. Tale software potrebbe essere *GeNIe 2.0 Graphical Network Interface*), interfaccia grafica di SMILE (Structural Modeling, Inference, and Learning Engine), che utilizza anche la propagazione della *soft evidence* nelle Reti Bayesiane.

Sia dunque la struttura di indipendenza del *Naive Bayes* modificato quella mostrata nella *Fig. 3.1*. Si inizializzino le CPT dei nodi attributo $\mathbf{X}_h$, condizionate alla variabile di classe C ed alla coviarata S, e la CPT della variabile di classe C, condizionata ad S. Per le inizializzazione della CPT si utilizzano i risultati della (3.27) e (3.35) derivanti della procedura di *learning* parametrico del *Cap. 3*.

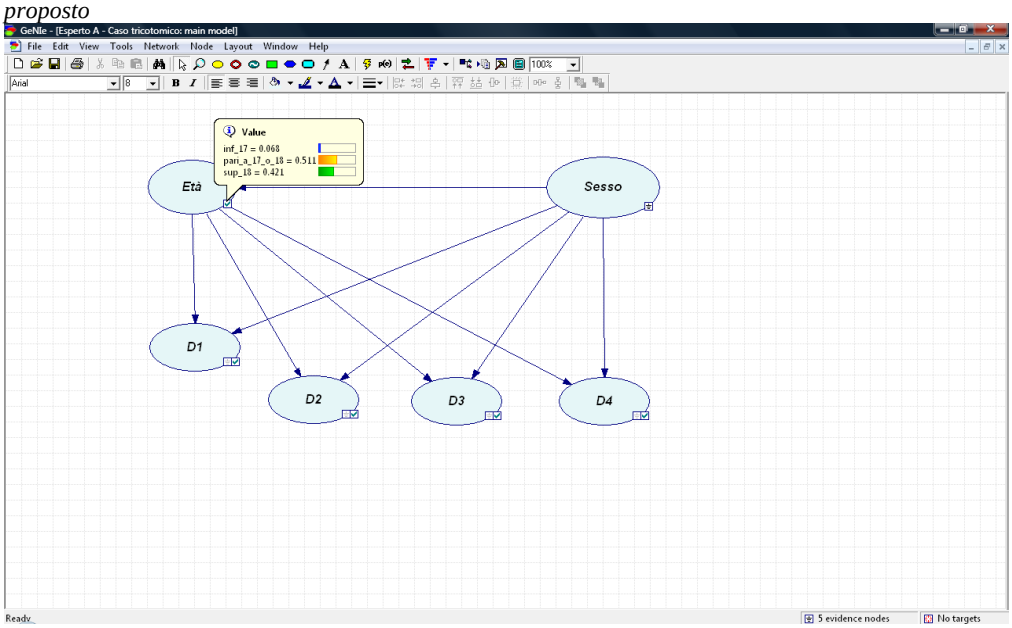
Dunque, l'Esperto A non deve far altro che immettere i *believes* modificati come *soft evidence* sugli attributi dei corrispondenti terzi molari ed il genere del soggetto, in questo caso femminile. L'interfaccia di *GeNIe 2.0* permette, con estrema semplicità, l'utilizzo dello stesso anche a persone meno esperte in ambito di reti Bayesiane, come potrebbero essere gli esperti odontologi chiamati a fornire una valutazione sulla maturazione dentale. Sia quindi immessa l'evidenza fornita dall'osservatore sui terzi molari ed il genere del soggetto esaminato come mostrato dalla nella *Fig. 4.6*.

Figura 4.6. *Interfaccia grafica di GeNIe 2.0 durante l'immissione della soft evidence per l'esempio proposto*



Una volta che l'Esperto A ha immesso nella rete tutta l'evidenza, *soft* o *hard*, ed eventuali casi *missing*, si propaga l'informazione fino al nodo variabile di classe *C*, ottenendo la predittiva (3.37) come si osserva nella *Fig. 4.7.*

Figura 4.7. *Interfaccia grafica di GeNIe 2.0 sulla predittiva della variabile di classe C per l'esempio proposto*



La predittiva così ottenuta è:

$$P(C = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix}) = \begin{bmatrix} 0,068 \\ 0,511 \\ 0,421 \end{bmatrix}.$$

Il modello di classificazione stimato sull'Esperto A procederà all'attribuzione della classe d'età per l'$(n+1)-esimo$ soggetto in funzione della regola decisionale impiegata, (3.38) o (3.39). I valori della predittiva sottolineano, comunque, un certo grado di incertezza nello stabilire la classe di appartenenza del soggetto esaminato. Per il modello classificatorio adoperato è improbabile che il soggetto abbia un'età inferiore ai 17 anni ma non è da escludersi che appartenga alla classe centrale, la quale suggerisce un'analisi ulteriore che possa discriminare maggiormente le possibili classi di età del soggetto.

Nel caso non si utilizzi il modello di transizione per modificare i *believes*, questi si potrebbero immettere direttamente nella rete sotto forma di due *hard evidence*, una *soft evidence* ed un *missing data*, per ottenere una predittiva praticamente identica a quella precedente:

$$P(C = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix}) = \begin{bmatrix} 0,069 \\ 0,528 \\ 0,403 \end{bmatrix}.$$

Conclusioni

In questa ricerca è stata presentata una metodologia per trattare l'assegnazione di un individuo vivente non adulto ad una determinata classe d'età mediante evidenza dentale. La soglia d'età di interesse è quella dei 18 anni ed il modello classificatorio proposto è stato applicato prima al caso dicotomico e poi a quello tricotomico includendo un classe di non-decisione a cavallo della soglia stessa.

Secondo la letteratura inerente al tema dell'assegnazione dell'età in funzione dello sviluppo dentale, sono stati valutati i terzi molari mediante la scala di classificazione dentale di Demirijan. L'attribuzione del dente ad uno degli stati dentali secondo Demirjian è stata ampliata introducendo la possibilità di indicare due stati adiacenti mediante la *soft evidence*.

I risultati ottenuti sulla riproducibilità intra-osservatore confermano l'efficacia nell'utilizzo della *soft evidence* la quale, se attribuita all'esperto A che ne fa maggior uso rispetto all'Esperto B, porta a migliori risultati di "coerenza" valutativa e soprattutto di *performance* classificatoria.

La tecnologia radiografica ha permesso di articolare i risultati ottenuto evidenziando come gli esperti chiamati a fornire le proprie valutazioni potrebbero utilizzare la tecnologia che sono più abituati a "leggere" e che conduce, quindi, ad una riproducibilità intra-osservatore più alta.

La misura di riproducibilità inter-osservatori, invece, suggerisce una stima dei modelli di classificazione *ad personam*, vale a dire il modello classificatorio deve essere "tarato" sulle valutazioni che l'esperto ha fornito precedentemente sulle unità del *training data set*.

Si conferma, inoltre, come i terzi molari inferiori conducano a risultati migliori rispetto a quelli dell'arcata superiore sia in termini di concordanza intra-osservatore e inter-osservatori che in termini di indici di *perfomance* classificatoria. L'evidenza derivante dall'utilizzo dell'informazione congiunta di tutti e quattro i terzi molari conduce a risultati più soddisfacenti rispetto a quella derivante da un loro sottoinsieme.

L'introduzione della terza classe d'età a cavallo dei 18 anni è stata conseguenza di due fatti. Da un lato, la capacità discriminatoria poco soddisfacente che un modello di classificazione fornisce nel caso l'età sia dicotomica. Infatti, gli individui che hanno un'età prossima alla soglia dei 18 anni, sia minorenni che maggiorenni, potrebbero essere facilmente classificati in modo errato, visto che in tale intorno non si producono significative differenze nella maturazione dentale di tali soggetti. Dall'altro, i risultati poco soddisfacenti del modello classificatorio nel caso dell'età dicotomica, specialmente se osservati in termini dell'indice che misura la percentuale di falsi maggiorenni prodotta. Tali individui, difatti, vengono erroneamente classificati e le conseguenze nell'applicazione delle leggi e normative vigenti potrebbero essere particolarmente gravi e difficilmente rimediabili: dunque è normale porre particolare attenzione a questo aspetto.

L'introduzione della classe centrale permette, quindi, la costruzione di un modello classificatorio basato su tre classi di età, riducendo notevolmente la percentuale di falsi maggiorenni. Essendo l'età osservata in anni compiuti, tale classe centrale ha un ampiezza di due anni e quindi una percentuale non indifferente di individui ricadrebbe all'interno di essa. Questo comporta una penalizzazione dal punto di vista classificatorio in quanto non si è in grado, per gli individui classificati in questa classe, di esprimere un'attribuzione del tipo maggiorenni-minorenni. Al contrario, un individuo che ricade nelle due classi esterne è fortemente discriminato, perlomeno dagli individui assegnati alla classe opposta. Questo suggerisce una prossima direzione di lavoro che intenda sfruttare misurazioni dell'età nel continuo per poter così modulare la classe centrale e poter così trovare una soluzione di classificazione ottimale: ridotta percentuale di falsi maggiorenni e ridotta percentuale di individui non classificabili.

L'Esperto A, la cui esperienza professionale è di tipo "forense" produce una *performance* classificatoria sia in termini di ridotta percentuale di falsi minorenni che ridotta di percentuale di individui non classificati. Questo potrebbe suggerire la possibilità si operare un'analisi classificatoria considerando anche la variabile *Esperto*. Non solo suddividere l'osservatore nelle due categorie, *clinico* e *forense*, caratterizzate dall'uso della *soft evidence* ma a più categorie in funzione del *background* professionale. Si potrebbe, ad esempio, operare una *Cluster analysis* raggruppando gli osservatori in gruppi di esperti che forniscano valutazioni fra loro molto simili, magari verificate su di un campione *test*. Si potrebbe utilizzare come misura di distanza fra i *clusters* la riproducibilità inter-osservatori. In questo modo si potrebbero raggruppare fra loro esperti che abbiamo valori di riproducibilità inter-osservatori alti e utilizzare per tutti gli osservatori del *cluster* un unico modello stimato in precedenza sulla base delle valutazioni di uno dei rappresentanti del *cluster*.

Questo significherebbe che attraverso un ridotto campione di prova ogni esperto verrebbe valutato in funzione della riproducibilità inter-osservatori ed

assegnato ad un *cluster*. Quindi ciascun nuovo osservatore potrebbe impiegare un modello di classificazione già esistente.
Inoltre questa procedura da "pre-test" avrebbe anche una seconda finalità. In sede legale ed in tutti quegli ambiti in cui un perito è chiamato a dare un giudizio in funzione del quale verranno poi prese delle decisioni spesso critiche, si tende a richiedere un esperto in funzione della disponibilità ed esperienza, ma senza che sia effettivamente misurata. Questo non significa dover fare un esame preliminare agli esperti bensì misurare in qualche modo la loro capacità valutativa per confrontarla con quella di altri e decidere quale esperto, e quindi modello classificatorio, sia più idoneo al compito richiesto.

Una metodologia come quella proposta suggerisce la possibilità di valutare l'esperto stesso prima ancora che esprima il proprio parere e quindi valutarne le abilità in merito alla scala di Demirijan, in questo caso, ed alla tecnologia radiografica utilizzata.

Inoltre, la soglia probabilistica della regola decisionale (3.38) può esser fatta variare in funzione del tema trattato e dell'accettabilità, da parte di un giudice, della valutazione probabilistica che il modello fornisce nell'assegnare un individuo ad una particolare classe d'età. Un caso penale, ad esempio, potrebbe richiedere una soglia di accettabilità probabilistica del 90% a differenza di un caso civile per il quale potrebbe bastarne più bassa. In funzione di tale soglia si potrebbe conoscere, prima ancora della selezione dell'esperto, la probabilità che questi produca un caso non classificabile o addirittura un falso maggiorenne.

In conclusione, tutto questo suggerisce un'attenzione particolare alla scelta dell'esperto che andrebbe testato e poi successivamente selezionato fra vari candidati, in funzione dei risultati ottenuti e della soglia di accettabilità fissata.

Indice delle tabelle

Indice delle figure

Bibliografia

Agresti A. (2002),
Categorical data analysis,
John Wiley & Sons Ltd, second edition, New York.

Bernardo J. M., Smith A.F.M. (1994),
Bayesian Theory,
John Wiley & Sons Ltd, Chichester.

Bilmes J. (2004),
On Soft Evidence in Bayesian Networks,
Tech. Rep. UWEETR-2004-0016, University of Washington, Dept. of Electr. Engineering.

Braga J., Heuze Y., Chabadel O., Sonan N.K. (2005),
Non-adult dental age assessment: correspondence analysis and linea regression versus Bayesian predictions,
Int. J. Legal Med. 119:260-274.

Brause R., Langsdorf T., Hepp M. (1999),
Neural Data Mining for credit card fraud detection,
Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence:103–106.

Cameriere R., De Angelis D., Ferrante L., Scarpino F., Cingolani M. (2007),
Age estimation in children by measurement of open apices in teeth: a European formula,
Int. J. Legal Med. 121:449-453.

Cameriere R., Ferrante L., Cingolani M. (2006),
Age estimation in children by measurement of open apices in teeth,
Int. J. Legal Med. 120:49-52.

Cameriere R., Ferrante L., De Angelis D., Scarpino F., Galli F. (2008),
The comparison between measurement of open apices of third molars and Demirjian stages to test chronological age over 18 years in living subjects,
Int. J. Legal Med. 122:493-497.

Cohen J. (1960),
A coefficient of agreemen for nominal scales,
Educational and Psychological Measurement 20:37-46.

Cowell R. G., Dawid A. P., Lauritzen S.L., Spiegelhalter D.J. (1999),
Probabilistic Networks and Expert Systems,
Springer-Verlag, New York.

Demirjian A., Goldstein H., Tanner J.M. (1973),
A new system of dental age assessment,
Hum. Biol. 45:221-227.

Dhanjal K.S., Bhardwaj M.K., Liversidge H.M. (2006),
Reproducibility of radiographic stage assessment of third molars,
Forensic Sci. Int. 159S:74-77.

Domingos P., Pazzani M. (1996),
Beyond independence: Conditions for the optimality of the simple Bayesian classifier,
Proceedings of the Thirteenth International Conference on Machine Learning:105–112.
L. Saitta (Ed.), San Francisco, CA: Morgan Kaufmann

Domingos P., Pazzani M. (1997),
On the optimality of the simple Bayesian classifier under zero-one loss,
Machine Learning, 29:103-130.

Duda R. O., Hart E. (1973),
Pattern classification and scene analysis,
Wiley, New York.

Duda O. Hart E., Stork D.G. (2000),
Pattern Classification,
Second edition, John Wiley & Sons, New York.

Everitt B.S. (2002),
The Cambridge dictionary of statistics,
Cambridge university press, USA.

Evett I.W., Weir B.S. (1998),
Interpreting DNA evidence. Statistical Genetics for forensic scientists,
Sinaur Associates Inc, Sunderland, Massachusetts.

Fisher R.A. (1936),
The use of multiple measurements in taxonomic problems,
Annals of Eugenics, 7:179–188.

Friedman J. (1997),
On bias, variance, 0/1 - loss, and the curse-of-dimensionality,
Data Mining and Knowledge Discovery, 1:55–77.

Friedman N., Geiger D., Goldszmidt M. (1997),
Bayesian network classifiers,
Machine Learning, 29:131-163.

Friedman N., Goldszmidt M. (1996),
Building classifiers using Bayesian networks,
Proceedings of the National Conference on Artificial Intelligence:1277–1284. Menlo Park, CA: AAAI Press

Gelman A., Carlin J. B., Stern H. S., Rubin D. B. (2004),
Bayesian Data Analysis,
Chapman and Hall, second edition, New York.

Geenen P.L., Van Der Gaag L.C., Loeffen W.L.A. (2004),
Building Naive Bayesian Classifiers from literature. A case study in classical Swine Fever,
Mimeo, Institute of Information and Computing Sciences, Utrecht University.

Guimaraes P., Richard Lindrooth R. (2005),
Dirichlet-Multinomial Regression,
Econometrics 0509001, EconWPA.

Hellerstein J., Thathachar J., Rish I. (2000),
Recognizing end-user transactions in performance management,
Proceedings of AAA1-2001:596-602, Austin, texas.

Hu Z., Chin W., Takeichi M. (2000),
Calculating a new data mining algorithm for market basket analysis,
Proceedings of the Second International Workshop on Practical Aspects of Declarative Languages (PADL'00):169-184, Springer Verlag, Massachusetts

Kim C., Hwang K.B.(2008),
Naive bayes classifier learning with feature selection for spam detection in social bookmarking,
Proc. Europ. Conf. on Machine Learning and Principles and Practice of Knowledge D. in D. Discovery in Databases (ECML/PKDD).

Kim S.B., Seo H.C., Rim H.C. (2003),
Poisson Naive Bayes for text classification with feature weighting,
Proceedings of the sixth international workshop on Information retrieval with Asian languages 11:33-40.

Kotz S., Johnson N.L. (1985),
Encyclopedia of Statistical Science,
5:378, John Wiley & Sons Ltd, New York.

Jensen F.V. (2007),
Bayesian Networks and Decision Graphs,
Springer-Verlag, second edition, New York.

Langley P., Iba W., Thompson K. (1992),
An analysis of Bayesian classifiers,
Proceedings, Tenth National Conference on Artificial Intelligence:223–228. San Jose: AAAI Press.

Landis J.R., Koch G.C. (1977),
The measurement of observer agreement for categorical data,
Biometrics 33:159-174.

Little R.J.A., Rubin D.B. (2002),
Statistical analysis with missing data,
John Wiley & Sons Ltd, second edition, New York.

Lucy D., Aykroydand R.G., Pollard A,M. (2002),
Nonparametric calibration for age estimation,
J. Royal Statistical Soc. 51:183-196.

Maber M., Liversidge H.M., Hector M.P. (2006),
Accuracy of age estimation of radiographic methods using developing teeth,
Forensic Sci. Int. 159S:68-73.

Meinl A., Tangl S., Huber C., Maurer B., Watzek G. (2007),
The chronology of third molar mineralization in the Austrian population. A contribution to forensic age estimation,
Forensic Sci. Int. 169:161-167.

Mitchell T. (1997),
Machine Learning,
McGraw-Hill, New York.

Olze A., Bilang D., Schmidt S., Wernecke K.D., Geserick G., Schmeling A. (2005),
Validation of common classification systems for assessing the mineralization of third morals,
Int. J. Legal Med. 119:22-26.

Olze A., Schmeling A., Taniguchi M., Maeda H., Van Niekerk P., Wernecke K.D., Geserick G. (2004),
Forensic age estimation in living subjects: the ethnic factor in wisdom tooth mineralization,
Int. J. Legal Med. 118:170-173.

Ouchtati S., Bedda M., Lachouri A. (2007),
Segmentation and recognition of handwritten numeric chains,
Journal of Computer Science 3 (4):242-248.

Pazzani M.J. (1995),
Searching for dependencies in Bayesian classifiers,
Proceedings of the fifth International Workshop on artificial Intelligence and Statistics,
D. Fisher e H. Lenz (Eds.), Ft. Lauderdale, FL.

Pearl J. (1988),
Probabilistic reasoning in intelligent system: Networks of plausible inference,
Morgan Kaufmann Publishers Inc., United States of America.

Pinchi V., Manetti G., Franchi E. (2005),
Identificazione dell'età e mineralizzazione dell'ottavo: verifica del metodo in una casistica italiana,
Arch. Med. Leg. Soc. Cri. 78°(23 S4)4:483-498.

Pomerlau D.A. (1989),
ALVINN: An Autonomous Land Vehicle in a Neural Network.
Advances in Neural Information Processing Systems I:305-313, Morgan Kaufmann.

Rish I. (2001),
An empirical study of the naive Bayes classifier,
IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence

Rojas R. (1996),
Neural Networks - A Systematic Introduction,
Springer-Verlag, Berlin, New-York.

Rubin D.B. (1976),
Inference and missing data,
Biometrika, 63(3):581-592.

Russell S., Norvig P. (2005),
Intelligenza artificiale, un approccio moderno,
Pearson Education Italia srl, Vol. 2, seconda edizione, Milano.

Samuel A.L. (1959),
Some studies in Machine Learning using the game of checkers,
IBM Journal 3(3):210–229.

Sardanelli F., Di Leo G. (2008),
Biostatistica in Radiologia. Progettare, realizzare e scrivere un lavoro scientifico radiologico,
Springer, Milano.

Saudi Arabian Standards Organization (2000),
Guide to the expression of uncertainty in measurements,
mimeo 13/2000, http://www.temperatures.ru/pdf/GUM.pdf.

Shapiro M.D. (1977),
The evaluation of clinical predictions. A method and initial application,
N Eng J Med 296(26):1509-1514.

Spitzer R.L., Cohen J., Fleiss J.L. (1967),
Quantification of agreement in psychiatric diagnosis: a new approach,
Arch. Gen. Psychiatry 17:83-89.

Swinburne, R. (2002),
Bayes' Theorem,
Oxford University Press, Oxford.

Tesauro G. (1995),
Temporal difference learning and TD-gammon,
Communications of the ACM, 38(3):5848.

Whittaker J. (1990),
Graphical models in applied multivariate statistics,
John Wiley & Sons Ltd, New York.

Zhang H. (2004),
The optimality of naive Bayes,
Proceedings of the 17th International FLAIRS conference (FLAIRS2004), AAAI Press.