Corso di Laurea Magistrale in Economia

Data Science

A.A. 2018/2019

Lez. 6 – Modelli di Classificazione

Definizione

- Si collocano tra i metodi di apprendimento supervisionato e mirano alla predizione di un attributo target categorico.
- Differiscono dai modelli di stima che trattano attributi numerici.
- A partire da un insieme di osservazioni riferite al passato per le quali è nota la classe di appartenenza, i modelli di classificazione si propongono di generare un insieme di regole che consentono di predire la classe di osservazioni future.
- Ruolo centrale nella teoria dell'apprendimento per le implicazioni teoriche le innumerevoli applicazioni:
  - Algoritmi dotati di capacità di apprendere dall'esperienza passata sono fondamentali per emulare le capacità di induzione che il cervello umano possiede
  - Alcuni esempi applicativi:
    - Selezione destinatari campagne di marketing
    - Identificazione di frodi
    - Riconoscimento immagini
    - Diagnosi precoce di patologie
    - Catalogazione testi

Problemi di classificazione

- In un problema di classificazione si dispone di un dataset *D* contenente *m* osservazioni costituite da *n* attributi esplicativi ed 1 attributo *target* categorico (*classe*).
- Gli attributi esplicativi (*predittori*) possono essere categorici e/o numerici
- Le osservazioni sono anche dette *istanze*
- La variabile *target* assume un numero finito di valori:
  - Solo 2 classi → classificazione *binaria*
  - Più di 2 classi → classificazione *multi-classe*
- I legami tra le variabili esplicative vengono tradotti in *regole di classificazione* che vengono impiegate per predire la classe di osservazioni delle quali è noto solo il valore degli attributi esplicativi

Esempi

Example 10.1 – Retention in the mobile phone industry. Example 5.2 on the analysis of customer loyalty in the mobile phone industry is a binary classification problem in which the target attribute takes the value 1 if a customer has discontinued service and 0 otherwise. The features of each customer, described in Table 5.3, represent the predictive attributes. The purpose of a classification model is to derive general rules from the examples contained in the dataset and to apply these rules in order to assign the class to new instances for which the target value is unknown. In this way, the classification model may prove useful in identifying those customers who are likely to discontinue service, and therefore to drive a retention marketing campaign.

Esempi

attribute	meaning
area	residence area
numin	number of calls received in period $t-2$
timein	duration in seconds of calls received in period $t-2$
numout	number of calls placed in the period $t-2$
Pothers	percentage of calls placed to other mobile telephone companies in period $t-2$
Pmob	percentage of calls placed to the same mobile telephone company in period $t-2$
Pland	percentage of calls placed to land numbers in period $t-2$
numsms	number of messages sent in period $t-2$
numserv	number of calls placed to special services in period $t-2$
numcall	number of calls placed to the call center in period $t-2$
diropt	binary variable indicating whether the customer corresponding to the record has subscribed to a special rate plan for calls placed to selected numbers
churner	binary variable indicating whether the customer corresponding to the record has left the service in period t

Esempi

area	numin	timein	numout	Pothers	Pmob	Pland	numsms	numserv	numcall	diropt	churner
3	32	8093	45	0.14	0.75	0.12	18	1	0	0	0
3	277	157842	450	0.26	0.35	0.38	9	3	0	1	0
1	17	15023	20	0.37	0.23	0.40	1	1	0	0	0
1	46	22459	69	0.10	0.39	0.51	33	1	0	0	0
1	19	8640	9	0.00	0.00	1.00	0	0	0	0	0
2	17	7652	66	0.16	0.42	0.43	1	3	0	1	0
3	47	17768	11	0.45	0.00	0.55	0	0	0	0	0
3	19	9492	42	0.18	0.34	0.48	3	1	0	0	1
1	1	84	9	0.09	0.54	0.37	0	0	0	0	1
2	119	87605	126	0.84	0.02	0.14	12	1	0	0	0
4	24	6902	47	0.25	0.26	0.48	4	1	0	0	0
1	32	28072	43	0.28	0.66	0.06	0	1	0	0	0
3	103	112120	24	0.61	0.28	0.11	24	2	0	0	0
3	45	21921	94	0.34	0.47	0.19	45	2	0	1	0
1	8	25117	89	0.02	0.89	0.09	189	1	3	0	0
3	4	945	16	0.00	0.00	1.00	0	0	0	0	1
2	83	44263	83	0.00	0.00	0.67	0	0	0	0	1
2	22	15979	59	0.05	0.53	0.41	5	2	0	1	1
2	0	0	57	0.00	1.00	0.00	15	1	1	0	1
4	162	114108	273	0.18	0.15	0.41	2	3	0	1	1
4	21	4141	70	0.14	0.58	0.28	0	1	0	1	1
4	33	10066	45	0.12	0.21	0.67	0	0	0	0	1
4	5	965	40	0.41	0.27	0.32	64	1	0	0	1

Esempi

Example 10.2 – Segmentation of customers phoning a call center. Many services and manufacturing companies nowadays have a call center that their customers may call to request information or report problems. In order to size the staff and the activities of a call center and to verify the quality of the services offered, it is useful to classify customers based on the number of calls made to the call center. The target attribute may be obtained through a proper discretization of the numerical variable indicating the number of calls, setting for example: class $0 \equiv$ no calls, class $1 \equiv 1$ call, class $2 \equiv$ from 2 to 4 calls, class $4 \equiv$ more than 4 calls. Again predictive attributes are provided by the features of the customers. Hence, the segmentation of the customers with respect to the number of calls made to the call center is a multicategory classification problem.

Problemi di classificazione

From a mathematical viewpoint, in a classification problem m known examples are given, consisting of pairs $(\mathbf{x}_i, y_i)$, $i \in \mathcal{M}$, where $\mathbf{x}_i \in \mathbb{R}^n$ is the vector of the values taken by the n predictive attributes for the ith example and $y_i \in \mathcal{H} = \{v_1, v_2, \ldots, v_H\}$ denotes the corresponding target class. Each component x_{ij} of the vector $\mathbf{x}_i$ is regarded as a realization of the random variable X_j, $j \in \mathcal{N}$, which represents the attribute $\mathbf{a}_j$ in the dataset $\mathcal{D}$. In a binary classification problem one has $H = 2$, and the two classes may be denoted as $\mathcal{H} = \{0, 1\}$ or as $\mathcal{H} = \{-1, 1\}$, without loss of generality.

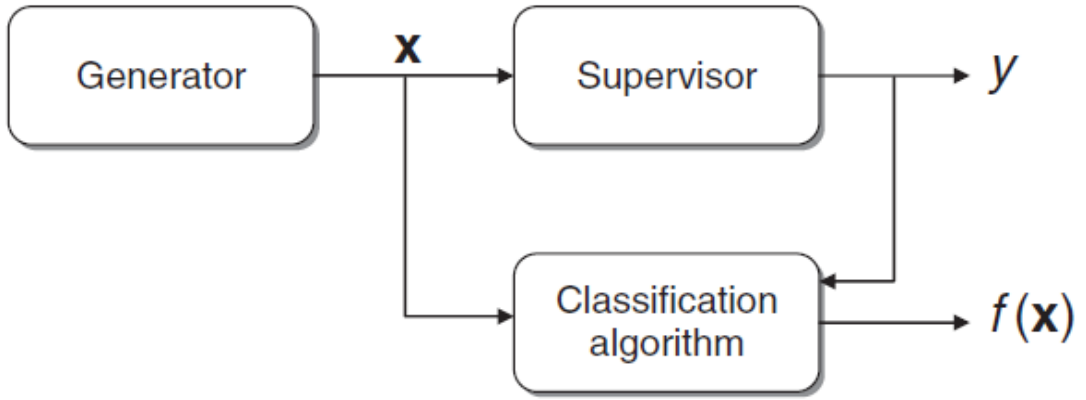
Let $\mathcal{F}$ be a class of functions $f(\mathbf{x}) : \mathbb{R}^n \mapsto \mathcal{H}$ called *hypotheses* that represent hypothetical relationships of dependence between y_i and $\mathbf{x}_i$. A *classification problem* consists of defining an appropriate hypothesis space $\mathcal{F}$ and an algorithm $A_{\mathcal{F}}$ that identifies a function $f^* \in \mathcal{F}$ that can optimally describe the relationship between the predictive attributes and the target class. The joint probability distribution $P_{\mathbf{x},y}(\mathbf{x}, y)$ of the examples in the dataset $\mathcal{D}$, defined over the space $\mathbb{R}^n \times \mathcal{H}$, is generally unknown and most classification models are nonparametric, in the sense that they do not make any prior assumption on the form of the distribution $P_{\mathbf{x},y}(\mathbf{x}, y)$.

Problemi di classificazione

Generator. The task of the generator is to extract random vectors $\mathbf{x}$ of examples according to an unknown probability distribution $P_{\mathbf{x}}(\mathbf{x})$.

Supervisor. The supervisor returns for each vector $\mathbf{x}$ of examples the value of the target class according to a conditional distribution $P_{y|\mathbf{x}}(y|\mathbf{x})$ which is also unknown.

Algorithm. A classification algorithm $A_{\mathcal{F}}$, also called a *classifier*, chooses a function $f^* \in \mathcal{F}$ in the hypothesis space so as to minimize a suitably defined loss function.



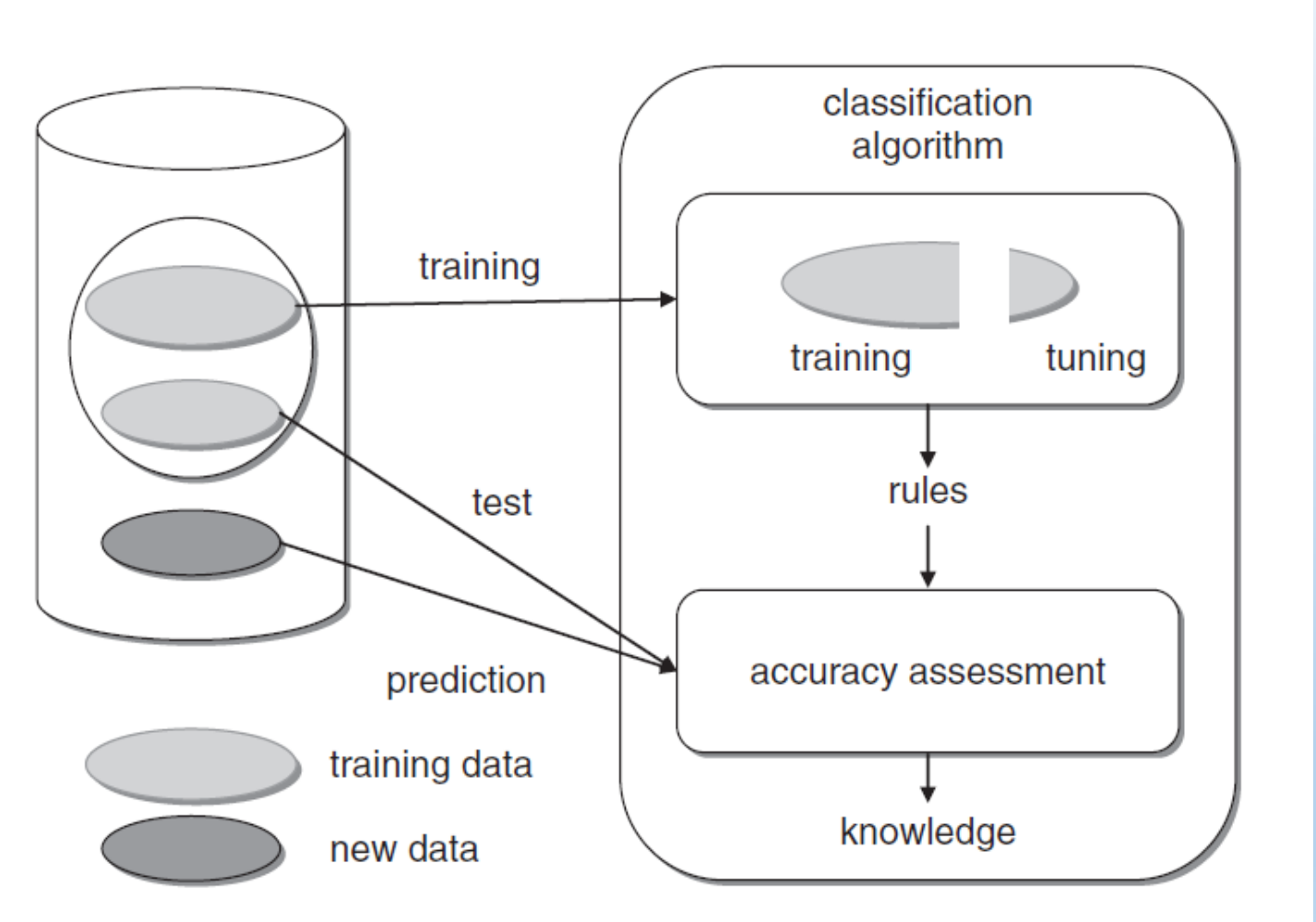
Utilizzo dei modelli di classificazione

- I modelli di classificazione, così come i modelli di stima, sono orientati sia all'interpretazione che alla predizione
- Le tipologie di modelli più semplici conducono a regole di classificazione intuitive che si prestano facilmente all'interpretazione
- I modelli più evoluti producono regole meno intelligibili ma garantiscono una maggiore accuratezza predittiva
- Una parte degli esempi nel dataset D viene utilizzate per il *training* di un modello di classificazione, ovvero per ricavare il legame funzionale tra la variabile target e le variabili esplicative, espresso mediante l'ipotesi $f^* \in F$.
- La restante parte dei dati disponibili viene successivamente impiegata per valutare l'accuratezza del modello generato e per scegliere il modello migliore tra quelli sviluppati

Sviluppo di un modello di classificazione

- **Training**
  - L'algoritmo di classificazione viene applicato agli esempi appartenenti al sottoinsieme *T* di *D* (*training set*), allo scopo di ricavare le regole di classificazione che consentono di attribuire a ciascuna osservazione **x** la corrispondente classe target *y*
- **Test**
  - Le regole prodotte nella fase di training vengono impiegate per classificare le osservazione di *D* non incluse nel training set, per le quali è noto il valore della classe target
  - Per valutare l'accuratezza del modello la classe di appartenenza di ciascuna istanza del *test set* viene confrontata con la classe predetta dal modello
  - È necessario che training set e test set siano disgiunti (rischio di sovrastima)
- **Predizione**
  - Effettivo utilizzo del modello di classificazione per assegnare la classe target alle nuove osservazioni
  - La predizione si ottiene applicando le regole generate in fase di training alle variabili esplicative che descrivono la nuova istanza.

Sviluppo di un modello di classificazione



Tassonomia dei modelli di classificazione

- Modelli euristici
  - Utilizzano procedure di classificazione basate su schemi semplici e intuitivi
  - Metodi *nearest neighbor*: si basano sulla nozione di distanza tra le osservazioni
  - Alberi di classificazione: utilizzano schemi di tipo *divide-et-impera* per creare raggruppamenti di osservazioni quanto più possibile omogenee rispetto alla classe target

Tassonomia dei modelli di classificazione

- Modelli di separazione
  - Ricavano nello spazio R^n degli attributi una collezione H di regioni disgiunte $\{S_1, S_2, \ldots, S_H\}$ che permettono di separare le osservazioni sulla base della classe target di appartenenza.
  - Le osservazioni $\mathbf{x_i}$ contenute nella regione S_h vengono assegnate alla classe $y_i = v_h$.
  - Ciascuna regione può essere costituita da un insieme composito ottenuto mediante operazioni di unione e intersezione di regioni di forma elementare
  - Le regioni non devono essere troppo complesse per non compromettere la capacità di generalizzazione del modello
  - Analisi discriminante
  - Modello *Perceptron*
  - Reti neurali
  - *Support Vector Machines*

Tassonomia dei modelli di classificazione

- Modelli di regressione
  - Si ipotizza in modo esplicito una forma funzionale per le probabilità condizionate che corrispondono al processo di assegnazione della classe target da parte del supervisore
  - Regressione lineare: ipotesi di esistenza di un legame lineare tra la variabile dipendente e i predittori e si determina il valore dei coefficienti di regressione
  - Regressione logistica: riconduce i problemi di classificazione binaria alla regressione lineare mediante una trasformazione

Tassonomia dei modelli di classificazione

- Modelli probabilistici
  - Si formula un'ipotesi circa la forma funzionale delle probabilità condizionate dalle osservazioni dara la classe target di appartenenza, indicate come probabilità condizionate alla classe
  - Successivamente, sulla base di una stima delle probabilità *a priori* si ricavano le probabilità *a posteriori* (mediante il teorema di Bayes)
  - Classificatori bayesiani
  - Reti bayesiane

Funzione di punteggio

- La maggior parte dei classificatori permette di ricavare una funzione di punteggio $g(\mathbf{x})$: $R^n \rightarrow R$ che associa a ciascuna osservazione **x** un numero reale
- A seguito di una standardizzazione, la funzione punteggio può essere interpretata come stima della probabilità che la classe predetta per l'osservazione **x** sia corretta
- Evidente per i modelli probabilistici
- Per i modelli di separazione il punteggio corrisponde alla distanza di un'osservazione dalla frontiera della regione corrispondente alla classe target
- Negli alberi di classificazione il punteggio corrisponde alla densità di osservazioni della classe target presenti nel nodo foglia cui l'osservazione è assegnata

Valutazione dei modelli di classificazione

- **Accuratezza**
  - È un indicatore della capacità di predizione di un modello a fronte di osservazioni future
  - Consente il confronto tra modelli diversi per la scelta del classificatore migliore per il sistema in esame
  - T training set (t osservazioni), V test set (v osservazioni)
  - $D = T \cup V$ e $m=t+v$
  - Indicatore più naturale: percentuale di osservazioni di V classificate correttamente
  - y_i classe di $\mathbf{x_i} \in V$, $f(\mathbf{x_i})$ classe predetta mediante la funzione f identificata dall'algoritmo di apprendimento A, si definisce la **funzione di perdita**

$$L(y_i, f(\mathbf{x}_i)) = \begin{cases} 0, & \text{if } y_i = f(\mathbf{x}_i), \\ 1, & \text{if } y_i \neq f(\mathbf{x}_i). \end{cases}$$

  - Pertanto l'accuratezza sarà:

$$\text{acc}_A(\mathcal{V}) = \text{acc}_{A_{\mathcal{F}}}(\mathcal{V}) = 1 - \frac{1}{v} \sum_{i=1}^{v} L(y_i, f(\mathbf{x}_i)).$$

Valutazione dei modelli di classificazione

- **Velocità**
  - I metodi più «veloci» consentono di trattare problemi di grandi dimensioni
  - I metodi di classificazione caratterizzati da tempi di elaborazione più elevati possono essere applicati a un training set di modeste dimensioni, con osservazioni prese a campione
- **Robustezza**
  - Un metodo è robusto se le regole da esso generate e la relativa accuratezza non variano in modo significativo al variare della scelta del training set e del test set, e se è capace di gestire dati mancanti e *outliers*
- **Scalabilità**
  - Possibilità di apprendere da dataset di grandi dimensioni
- **Interpretabilità**
  - Se l'analisi è rivolta all'interpretazione, oltre che alla predizione, è necessario che le regole generate siano semplici e comprensibili per i decisori

Metodo *holdout*

- Serve a suddividere le m osservazioni disponibili tra training set (T) e test set (V), per poi valutare l'accuratezza mediante $acc_A(V)$
- In genere, T viene formato mediante campionamento semplice, V per differenza
- La cardinalità di T varia tra metà e i due terzi delle osservazioni disponibili
- L'accuratezza calcolata con il metodo holdout dipende dalla scelta di V, e può quindi costituire una sovrastima o una sottostima dell'accuratezza effettiva

Campionamenti casuali ripetuti

- Si replica *r* volte il metodo holdout
- Per ogni iterazione si estrae un campione casuale indipendente T_k (con *t* osservazioni) e si calcola $acc_A(V_k)$, dove $V_k = D - T_k$
- Al termine si stima l'accuratezza del classificatore mediante la media campionaria

$$\mathrm{acc}_A = \mathrm{acc}_{A_{\mathcal{F}}} = \frac{1}{r}\sum_{k=1}^{r} \mathrm{acc}_{A_{\mathcal{F}}}(\mathcal{V}_k).$$

- La stima così ottenuta è più attendibile

Alberi di classificazione

- Uno dei metodi più noti ed utilizzati
  - Semplicità concettuale
  - Facilità d'uso
  - Velocità di elaborazione
  - Robustezza rispetto a valori mancanti e *outliers*
  - Facile interpretabilità
- Lo sviluppo corrisponde alla fase di training ed è eseguito mediante una procedura ricorsiva di tipo euristico, basata su uno schema di tipo *divide-et-impera*, indicata come *Top Down Induction of Decision Tree* (TDIDT)

Alberi di classificazione

- Le osservazioni del training set, inizialmente presenti nel nodo radice, vengono ripartite in sottoinsiemi disgiunti che confluiscono in 2 o più nodi discendenti (*branching*)
- In corrispondenza di ognuno dei nodi così generati si verifica se sono soddisfatte le condizioni di arresto
- Se almeno una di queste condizioni è soddisfatta, il nodo non è soggetto a partizione (nodo foglia)
- In caso contrario, si procede ad una ulteriore suddivisione delle osservazioni contenute nel nodo in esame
- Quando nessun nodo può essere ulteriormente ripartito, ciascun nodo foglia viene etichettato con il valore della classe alla quale appartiene la maggioranza delle osservazioni
- La divisione degli esempi in ogni nodo viene effettuata mediante una regola di separazione selezionata in base ad una funzione di valutazione
- Al variare della regola di separazione si ottengono differenti alberi
- L'insieme delle regola di separazione lungo il cammino dalla radice ad una foglia costituisce una regola di classificazione

Schema TDIDT

1) In fase di inizializzazione, ciascuna osservazione viene inclusa nel nodo radice dell'albero. La radice viene inserita nella lista *L* dei nodi attivi
2) Se la lista *L* è vuota la procedura si arresta. Altrimenti si selezione un nodo *J* appartenente a *L*, lo si rimuove dalla lista e lo utilizza come nodo di analisi
3) Si determina la regola ottimale di separazione delle osservazioni presenti in *J*, sulla base di un opportuno criterio prestabilito. Si applica la regola di separazione generata e si costruiscono i nodi discendenti suddividendo le osservazioni presenti in *J*. Per ogni nodo discendente, si verificano le condizioni per arrestare la suddivisione. Se sono soddisfatte, il nodo *J* costituisce una foglia cui viene assegnata la classe target determinata dalla maggioranza delle osservazioni presenti in *J*. Altrimenti il nodo discendente viene aggiunto a *L*. Si ripete il passo 2.

Alberi di classificazione

- In fase di predizione, per assegnare la classe target a una nuova osservazionesi attraversa l'albero dal nodo radice a un nodo foglia, raggiunto seguendo il cammino determinato dalla sequenza di regole soddisfatte dai valori degli attributi della nuova osservazione.
- La classe target predetta coincide con la classe mediante cui è stato etichettato in fase di sviluppo il nodo foglia così raggiunto
- A partire da un training set è possibile costruire un numero esponenziale di alberi di classificazione distinti. Il problema di determinare l'albero ottimale è *NP-hard* (computazionalmente molto complesso)
- La procedura TDIDT richiede la specifica di alcuni aspetti cruciali

Fasi cruciali

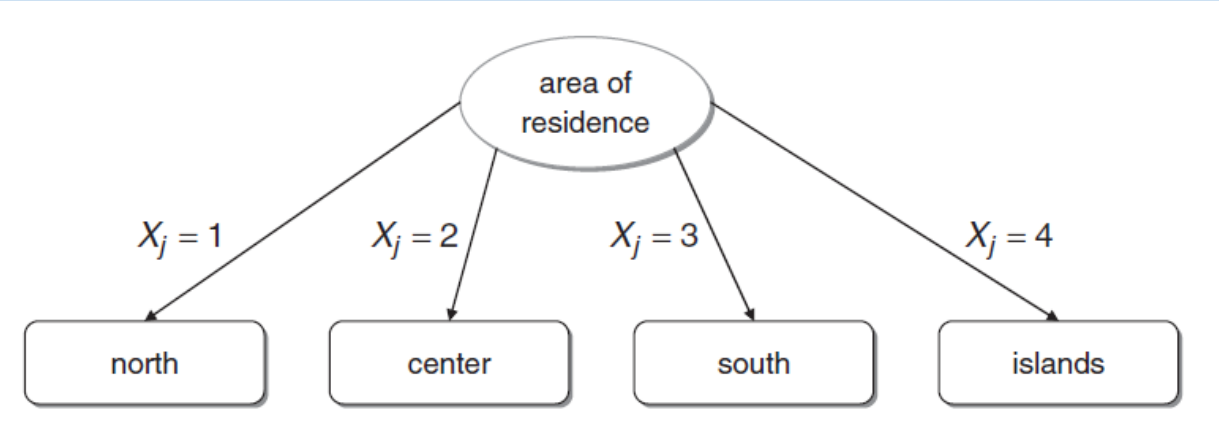
- Regole di separazione
  - Per ogni nodo dell'albero è necessario indicare i criteri per identificare la regola ottimale di separazione delle osservazioni e di creazione dei nodi discendenti
- Criteri di arresto
  - In ogni nodo vengono applicati diversi criteri di arresto per stabilire se è necessario proseguire ricorsivamente lo sviluppo oppure se il nodo debba costituire una foglia
- Criteri di *pruning*
  - Occorre applicare criteri di *potatura* per evitare una crescita eccessiva dell'albero nella fase di sviluppo, e successivamente per ridurre il numero di nodi

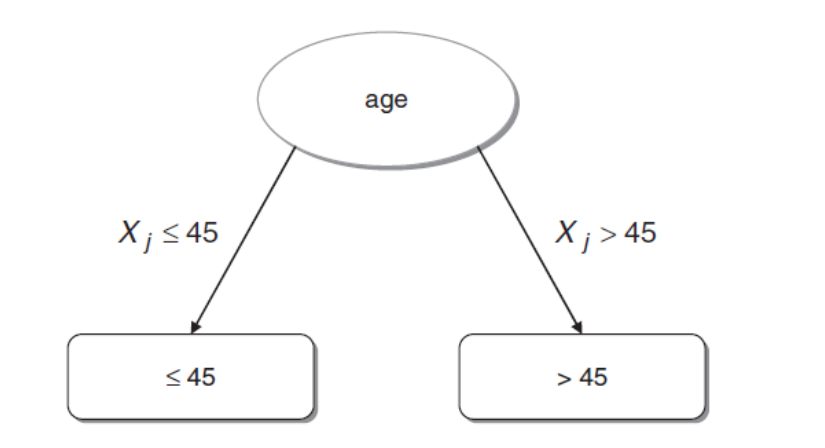
Regole di separazione – numero di ramificazioni

- Alberi binari
  - Da ogni nodo si dipartono al più due ramificazioni
  - Rappresentazione naturale della suddivisione delle osservazioni sulla base del valore di un attributo esplicativo binario
  - Nel caso di attributi categorici con più di due classi, gli alberi binari devono formare 2 gruppi di categorie per la separazione. Ad es., clienti delle zone 1 e 2 nel ramo destro, quelli delle zone 3 e 4 nel ramo sinistro
  - Gli attributi numerici possono essere separati in basi ad un valore soglia. Ad es., i clienti di età inferiore a 45 anni nel ramo destro, quelli di età superiore nel ramo sinistro
- Alberi generali
  - Da ogni nodo si diparte un numero arbitrario di ramificazioni
  - Consentono di trattare più agevolmente attributi esplicativi categorici a più valori
  - Per gli attributi numerici è comunque necessario procedere ad un raggruppamento di valori contigui (discretizzazione dinamica)

Regole di separazione – modalità di separazione

- Alberi univariati
  - La regola di separazione si basa sul valore di un unico attributo X_j
  - Se l'attributo è categorico, le osservazioni vengono ripartite mediante relazioni del tipo $X_j \in B_k$, dove la collezione $\{B_k\}$ è formata da sottoinsiemi disgiunti ed esaustivi dell'insieme dei valori assunti da Xj.





Regole di separazione – modalità di separazione

- Alberi multivariati
  - La regola di separazione si basa sul valore assunto da una funzione degli attributi $\varphi(x_1, x_2, \ldots, x_n)$ e conduce ad una regola nella forma $\varphi(x) \leq b$ oppure $\varphi(x) > b$
  - Se la funzione è costituita da una combinazione lineare delle variabili esplicative

$$\sum_{j=1}^{n} w_j x_j \leq b,$$

dove il valore soglia *b* e i coefficienti *w1, w2, wn* devono essere determinati, ad esempio risolvendo un problema di ottimizzazione per ogni nodo

Criteri di separazione univariati

- Alberi univariati più diffusi per la semplicità delle regole generate
- Individuazione attributo esplicativo migliore e selezione della partizione più efficace mediante il calcolo di una *funzione di valutazione*
- Questa fornisce una misura di «disomogeneità» nei valori della classe target tra le osservazioni appartenenti al nodo padre e quelle dei discendenti
- La massimizzazione della funzione conduce all'individuazione della partizione che genera nodi più omogenei al loro interno di quanto non sia il nodo padre
- Siano p_h, $h \in H$, la % di osservazioni di classe target v_h contenute nel nodo q e Q il numero totale di osservazioni in q
- Deve valere la relazione

$$\sum_{h=1}^{H} p_h = 1.$$

Criteri di separazione univariati

- Indice di eterogeneità, $I(q)$, è funzione delle frequenze p_h
- Deve soddisfare 3 requisiti:
  - Assumere valore massimo quando le osservazioni del nodo sono distribuite in modo omogeneo su tutte le classi
  - Assumere valore minimo quando tutte le istanze del nodo appartengono alla stessa classe
  - Rappresentare una funzine simmetrica rispetto alla frequenze relative p_h
- Diversi indicatori (anche detti misure di impurità o di disomogenetà)

Criteri di separazione univariati

- **Indice di misclassificazione**, misura la % di punti classificati in modo errato, nell'ipotesi di assegnare tutte le istanze del nodo q alla classe cui appartiene la maggioranza di queste, secondo il criterio del *majority voting*

$$\text{Miscl}(q) = 1 - \max_h p_h,$$

- **Indice di entropia**

$$\text{Entropy}(q) = -\sum_{h=1}^{H} p_h \log_2 p_h;$$

con la convenzione $0 \log_2 0 = 0$

- **Indice di Gini**

$$\text{Gini}(q) = 1 - \sum_{h=1}^{H} p_h^2.$$

Criteri di separazione univariati

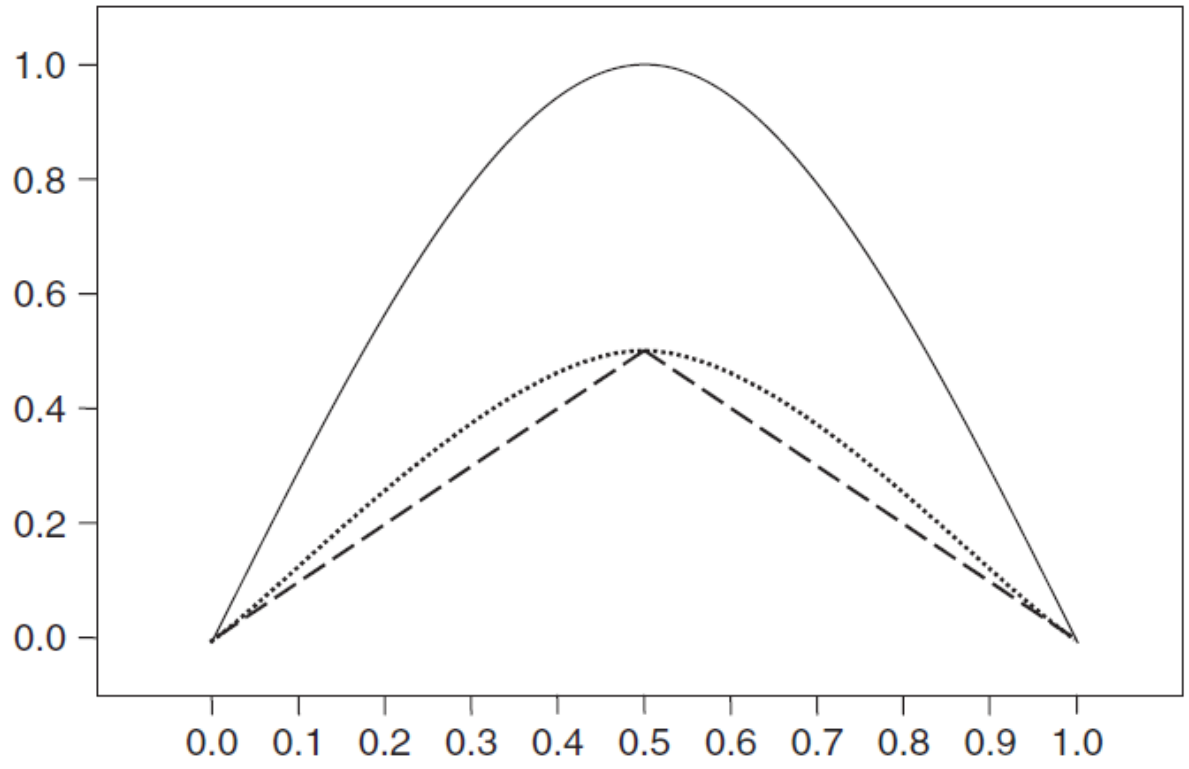


Figure 10.14 Graph of the misclassification index (dashed line), the Gini index (dotted line) and the entropy (full line) for a binary target attribute as the frequency of the examples in one class varies

Criteri di separazione univariati

- I criteri di separazione univriati basati sul guadagno di informazione confrontano uno degli indicatori di impurità calcolato per il nodo padre con il medesimo indicatore calcolato per il complesso dei discendenti per poi scegliere l'attributo e la relativa aprtizione che rendono massima la differenza
- Sia $I(\bullet)$ uno degli indicatori definiti e una regola di partizione suddivida le osservazioni del nodo q in K nodi discendenti $\{q_1, q_2, \ldots, q_K\}$, ciascuno con Q_k osservazioni
- Se la suddivisione è originata dall'attributo X_j che assume H_j valori distinti, l'insieme delle osservazioni in q può essere ripartito in $K = H_j$ sottoinsiemi.
- Ogni discendente q_j ha tutte le osservazioni nelle quali Xj assume valore vj.
- Se $H_j > 2$, albero generale
- Se albero binario, occorre dividere gli H_j in 2 gruppi non vuoti e calcolare gli indici per le 2^{Hj-1} possibili partizioni
- Se X_j è numerico si possono suddividere le osservazioni per intervallo di valori, al variare dei valori soglia

Criteri di separazione univariati

- L'impurità dei discendenti, e quindi della regola di partizione, è data dalla somma pesata delle impurità dei singoli discendenti, dove i pesi sono rappresentati dalla % di osservazioni del nodo padre che si collocano nel nodo discendente

$$I(q_1, q_2, \ldots, q_K) = \sum_{k=1}^{K} \frac{Q_k}{Q} I(q_k).$$

- Gli algoritmi scelgono per ciascun nodo la regola e il corrispondente attributo che determinano il valore minimo dell'espressione precedente, che equivale a massimizzare il guadagno di informazione:

$$\begin{aligned}\Delta(q, q_1, q_2, \ldots, q_K) &= I(q) - I(q_1, q_2, \ldots, q_K) \\ &= I(q) - \sum_{k=1}^{K} \frac{Q_k}{Q} I(q_k).\end{aligned}$$

Esempio – classificazione abbandono (churner)

Table 10.4 Discretized input data for Example 5.2

area	numin	timein	numout	Pothers	Pmob	Pland	numsms	numserv	numcall	diropt	churner
2	1	1	2	1	4	1	3	2	2	0	1
1	1	3	3	2	4	1	4	2	3	0	0
3	2	1	2	2	4	1	3	2	1	0	0
1	2	3	2	3	4	1	1	2	1	0	0
2	3	4	4	4	1	1	3	2	1	0	0
3	3	4	1	4	2	1	4	3	1	0	0
3	3	3	4	4	3	1	4	3	1	1	0
1	1	1	1	1	3	2	1	1	1	0	1
2	2	2	2	1	3	2	2	3	1	1	1
4	2	1	3	2	3	2	1	2	1	1	1
3	1	1	2	2	2	2	2	2	1	0	1
4	3	4	4	2	1	2	2	4	1	1	1
2	1	1	3	2	3	2	2	4	1	1	0
4	2	1	2	3	2	2	2	2	1	0	0
3	3	4	4	3	2	2	2	4	1	1	0
1	1	2	1	4	2	2	2	2	1	0	0
4	1	1	2	4	2	2	4	2	1	0	1
1	1	1	1	1	1	3	1	1	1	0	0
3	1	1	1	1	1	3	1	1	1	0	1
2	3	4	3	1	1	3	1	1	1	0	1
1	3	3	3	1	2	3	4	2	1	0	0
4	2	2	2	2	2	3	1	1	1	0	1
3	3	2	1	4	1	3	1	1	1	0	0

Indice di entropia

- Per il nodo radice

$$I_E(q) = \text{Entropy}(q) = -\frac{13}{23}\log_2\frac{13}{23} - \frac{10}{23}\log_2\frac{10}{23} = 0.988.$$

- Sulla base dell'attributo «Zona», la radice verrebbe ripartita in 4 nodi, cui corrispondono le seguenti % di appartenenza alle classi target {0, 1}

$$p_0(q_1) = 5/6, \quad p_0(q_2) = 2/5, \quad p_0(q_3) = 5/7, \quad p_0(q_4) = 1/5,$$
$$p_1(q_1) = 1/6, \quad p_1(q_2) = 3/5, \quad p_1(q_3) = 2/7, \quad p_1(q_4) = 4/5.$$

- L'indice di entropia dei discendenti è:

$$I_E(q_1, q_2, q_3, q_4) = \frac{6}{23}I_E(q_1) + \frac{5}{23}I_E(q_2) + \frac{7}{23}I_E(q_3) + \frac{5}{23}I_E(q_4)$$
$$= \frac{6}{23}0.650 + \frac{5}{23}0.971 + \frac{7}{23}0.863 + \frac{5}{23}0.722 = 0.8.$$

Indice di entropia

- Pertanto il guadagno sarebbe il nodo radice

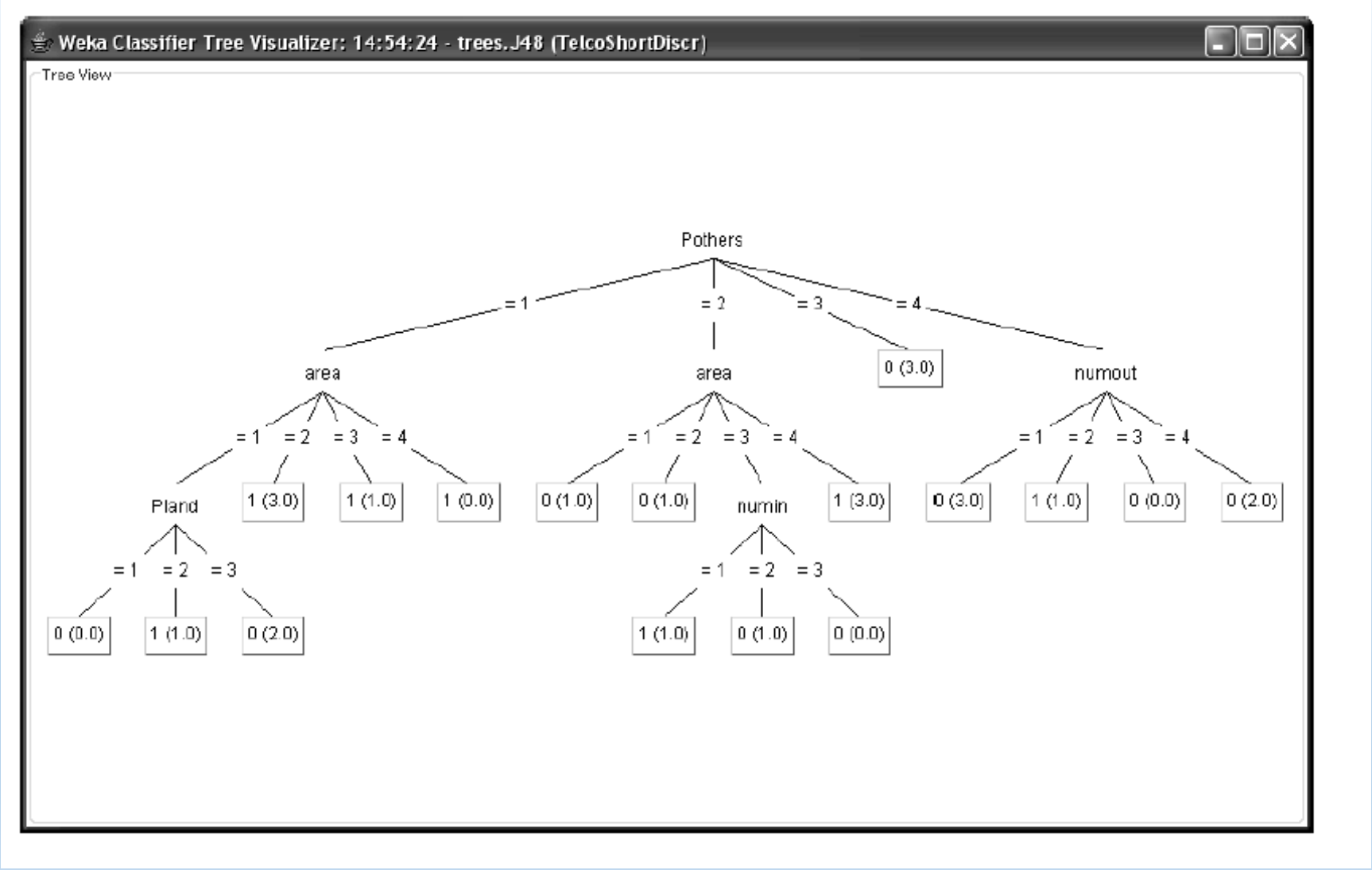
$$\Delta_E(area) = \Delta_E(q, q_1, q_2, q_3, q_4) = I_E(q) - I_E(q_1, q_2, q_3, q_4)$$
$$= 0.988 - 0.8 = 0.188.$$

- In modo analogo si possono calcolare i guadagno associati agli altri attributi:

$$\begin{aligned}
\Delta_E(numin) &= 0.057, & \Delta_E(Pland) &= 0.125,\\
\Delta_E(timein) &= 0.181, & \Delta_E(numsms) &= 0.080,\\
\Delta_E(numout) &= 0.065, & \Delta_E(numserv) &= 0.057,\\
\Delta_E(Pothers) &= 0.256, & \Delta_E(numcall) &= 0.089,\\
\Delta_E(Pmob) &= 0.043, & \Delta_E(dirop t) &= 0.005.
\end{aligned}$$

- Il guadagno massimo si ha per l'attributo *P others* che indica la % di chiamate verso altri operatori.
- Proseguendo in maniera analoga nelle successive iterazioni si ottiene l'albero seguente

Esempio – classificazione abbandono (churner)



Criteri di arresto

- Esistono 2 ragioni principali che inducono a limitare la crescita di un albero di classificazione
  - Un albero troppo ramificato presenta di solito un'accuratezza maggiore nei confronti del training set ma determina errori maggiori quando viene applicato al test set. Ciò perché un albero troppo ramificato «aderisce» troppo alle istanze del training set, quindi possiede una minore capacità di generalizzazione
  - Un albero più ramificato comporta una proliferazione di nodi foglia e quindi determina regole di classificazione molto profonde (unione di molte regole di separazione)
- In teoria la partizione di un nodo potrebbe avere termine quando in esso è presente una sola istanza.
- I criteri di arresto impiegati impediscono che si verifichi questa situazione, introducendo delle limitazioni sul numero minimo di osservazioni che un nodo deve contenere per essere ripartito, oppure sulla soglia minima di uniformità della classe target

Criteri di arresto

- Un nodo si trasforma in una foglia dell'albero quando si verifica almeno una delle seguenti condizioni:
  - **Numerosità**, ossia quando il nodo contiene un numero di osservazioni inferiore a una certa soglia
  - **Purezza**, quando la % delle osservazioni presenti nel nodo appartenenti alla medesima classe è superiore ad un determinato valore (accuratezza desiderata)
  - **Miglioramento**, quando l'eventuale suddivisione del nodo determina un guadagno inferiore a una certa soglia

Tecniche di *pruning*

- Dopo aver terminato lo sviluppo dell'albero, si utilizzano adottate tecniche per ridurre il numero di ramificazioni senza peggiorare l'accuratezza del modello
- Per valutare la convenienza associata a ogni fase iterativa di pruning si impiega l'albero per classificare le osservazioni del tuning set
- Al termine della valutazione si sceglie l'albero cui è associato l'errore di predizione minimo.