Esercitazione #3- Data Mining

D'Andrea Giuseppe

Domanda 1

a)

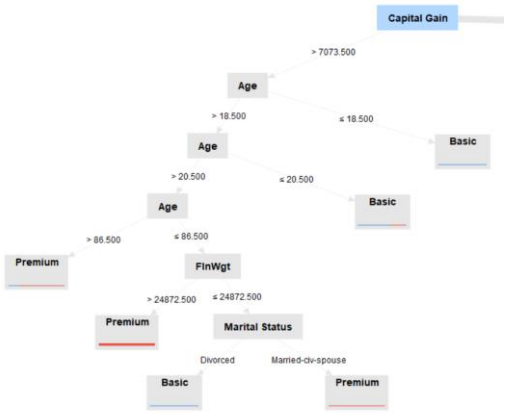
L'albero di decisione con l'algoritmo Decision Tree usando l'intero dataset (Utenti.xls) per il training, settando il "minimal gain" a 0.01 e mantenendo la configurazione di default per gli altri parametri, ha come attributo più selettivo "Capital Gain". I valori possibili di "Capital Gain" sono stati discretizzati ottenendo due range: "<= 7073.500" e "> 7073.500".

b)

L'altezza dell'albero di decisione dipende dalle dimensioni e dalle caratteristiche del dataset utilizzato e rappresenta il cammino massimo, espresso come numero di archi attraversati, tra la radice ed una foglia. Per il dataset preso in esame (Utenti.xls), con "minimal gain" a 0.01 e mantenendo la configurazione di default per gli altri parametri, l'altezza dell'albero di decisione con l'algoritmo Decision Tree risulta essere pari a 9.

c)

In generale, un partizionamento puro identifica idealmente delle partizioni pure. Una foglia è pura se ha tutti gli elementi etichettati come appartenenti alla stessa classe. Nell'albero di decisione preso in esame, ci sono dei partizionamenti puri (come quello mostrato in figura) oltre a molti partizionamenti quasi puri.



Domanda 2

La profondità dell'albero di decisione varia in base alle dimensioni e alle caratteristiche del dataset. Il parametro "maximal depth" può essere usato per limitare l'altezza massima dell'albero di decisione. Come visto in precedenza, settando il "minimal gain" a 0.01 e mantenendo la configurazione di default per gli altri parametri (maximal depth = 10), l'altezza dell'albero di decisione risulta essere pari a 9. Impostando un valore di "maximal depth" minore di 10 (in RapidMiner il parametro tiene conto dei nodi e non degli archi attraversati nel cammino massimo, e dunque è necessario incrementare il valore di un'unità rispetto al valore di altezza dell'albero) la ricorsione dell'algoritmo viene interrotta ad un livello superiore e si genera un albero potenzialmente meno dettagliato.

L'algoritmo Decision Tree con la configurazione di default utilizza il "Gain Ratio" come criterio per calcolare il guadagno, che viene calcolato in un nodo prima di partizionarlo. Il nodo viene diviso se il suo guadagno è maggiore del guadagno minimo, configurabile in RapidMiner con il parametro "minimal gain". Un valore più alto di "minimal gain" si traduce in meno partizionamenti e quindi in un albero più piccolo. Un valore troppo alto impedirà completamente la divisione e verrà generato un albero con un singolo nodo. Viceversa, mantenendo costante "maximal depth" e decrementando il "minimal gain" si nota come l'albero di decisione diventi sempre più fitto e dunque più dettagliato.

Domanda 3

a)

Utilizzando "Native Country" come attributo di classe, l'albero di decisione con l'algoritmo Decision Tree usando l'intero dataset (Utenti.xls) per il training, settando il "minimal gain" a 0.01 e mantenendo la configurazione di default per gli altri parametri, ha come attributo più selettivo "Education-Num". I valori possibili di " Education-Num" sono stati discretizzati ottenendo due range: "<= 3.500" e "> 3.500".

b)

L'altezza dell'albero di decisione dipende dalle dimensioni e dalle caratteristiche del dataset utilizzato e rappresenta il cammino massimo, espresso come numero di archi attraversati, tra la radice ed una foglia. Utilizzando "Native Country" come attributo di classe, per il dataset preso in esame (Utenti.xls), con "minimal gain" a 0.01 e mantenendo la configurazione di default per gli altri parametri, l'altezza dell'albero di decisione con l'algoritmo Decision Tree risulta essere pari a 9.

c)

In generale, un partizionamento puro identifica idealmente delle partizioni pure. Una foglia è pura se ha tutti gli elementi etichettati come appartenenti alla stessa classe. Nell'albero di decisione preso in esame, essendo l'attributo di classe non binario, non ci sono dei partizionamenti puri ma solo alcuni partizionamenti quasi puri.

Domanda 4

L'accuratezza misura il numero di predizione corrette rispetto al numero totale di predizioni fatte. È un indice di valutazione che funziona bene se la distribuzione delle classi è abbastanza bilanciata.

L'accuratezza aumenta al diminuire del parametro "maximal depth" anche se il livello di dettaglio dell'albero di decisione sta diminuendo sempre più. Questo fenomeno, per cui il modello diventa troppo semplice per classificare nuovi record in modo corretto, prende il nome di "underfitting".

Si può anche verificare il fenomeno opposto, quello dell'"overfitting", quando l'albero di decisione è molto dettagliato e ramificato. Così facendo il modello generato risulta troppo incentrato sul dataset di training per poter classificare in modo corretto nuovi record. Questo fenomeno si verifica quando si aumenta o si mantiene costante il valore di "maximal depth" e si diminuisce progressivamente il valore di "minimal gain".

Domanda 5

L'accuratezza del classificatore K-Nearest Neighbor (K-NN) aumenta incrementando il valore k (1, 3, 5, 7, 10, …), ovvero il numero di vicini più vicini da considerare per stabilire l'etichetta di classe. Applicando un 10-fold Stratified Cross-Validation si raggiunge un massimo nell'accuratezza (79.90%) per k=20 per poi iniziare una progressiva attenuazione per valori di k maggiori di 20 (25, 30, …).

L'accuratezza ha questo andamento perché, per valori troppo piccoli di k, il classificatore è sensibile ai punti rumorosi, mentre, per valori troppo grandi di k, il classificatore include tra i vicini dei punti che appartengono ad un'altra classe.

Utilizzando il classificatore Naïve Bayes, applicando un 10-fold Stratified Cross-Validation, si ottiene un'accuratezza media (83.34%) superiore a quella del K-NN (79.90% con k=20). Questo risultato può essere dovuto al fatto che gli attributi del dataset preso in esame (Utenti.xls) sono abbastanza indipendenti tra loro, che è l'ipotesi necessaria del classificatore Naïve Bayes affinché il risultato sia di buona qualità.

Domanda 6

La **Error! Reference source not found.**figura sottostante mostra la matrice di correlazione ottenuta dal dataset analizzato (Utenti.xls). Essa riporta la correlazione mutua (e simmetrica) tra coppie di attributi. La tabella mostra molti "?" in quanto ha senso calcolare la correlazione solo per valori numerici, mentre il dataset è caratterizzato anche da attributi nominali. In RapidMiner è presente un blocco denominato "Nominal to Numerical" che permette di trasformare un attributo da nominale a numerico. Ha però delle limitazioni che potrebbero portare a risultati completamente errati.

Per i valori presenti nella matrice di correlazione si evince una bassa correlazione tra gli attributi numerici. Per questo, in aggiunta ai risultati precedentemente ottenuti (il classificatore Naïve Bayes ha un'accuratezza superiore al K-NN) si può ipotizzare che gli attributi siano indipendenti tra loro.

Attribut...	Age	Workcl...	FlnWgt	Educati...	Educati...	Marital ...	Occupa...	Relation...	Race	Sex	Capital ...	Capital l...	Hours-p...	Native C...
Age	1	?	-0.077	?	0.037	?	?	?	?	-0.089	0.078	0.058	0.069	?
Workclass	?	1	?	?	?	?	?	?	?	?	?	?	?	?
FlnWgt	-0.077	?	1	?	-0.043	?	?	?	?	-0.027	0.000	-0.010	-0.019	?
Education	?	?	?	1	?	?	?	?	?	?	?	?	?	?
Educatio...	0.037	?	-0.043	?	1	?	?	?	?	-0.012	0.123	0.080	0.148	?
Marital St...	?	?	?	?	?	1	?	?	?	?	?	?	?	?
Occupati...	?	?	?	?	?	?	1	?	?	?	?	?	?	?
Relation...	?	?	?	?	?	?	?	1	?	?	?	?	?	?
Race	?	?	?	?	?	?	?	?	1	?	?	?	?	?
Sex	-0.089	?	-0.027	?	-0.012	?	?	?	?	1	-0.048	-0.046	-0.229	?
Capital ...	0.078	?	0.000	?	0.123	?	?	?	?	-0.048	1	-0.032	0.078	?
Capital l...	0.058	?	-0.010	?	0.080	?	?	?	?	-0.046	-0.032	1	0.054	?
Hours-p...	0.069	?	-0.019	?	0.148	?	?	?	?	-0.229	0.078	0.054	1	?
Native C...	?	?	?	?	?	?	?	?	?	?	?	?	?	1