Quaderno #2- Data Mining

D'Andrea Giuseppe

Domanda 1

a)

L'albero di decisione con l'algoritmo Decision Tree usando l'intero dataset (breast.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 "node-caps". I valori possibili di "node-caps" sono "yes", "no" e "?" e questo significa che l'attributo viene gestito dall'algoritmo come un attributo categorico a tre vie, di cui una ("?") è rappresentativa della situazione "valore mancante". La presenza di quest'ultima partizione ("?") non dovrebbe impattare sui risultati dell'algoritmo in quanto solo una frazione abbastanza piccola del dataset è affetta da questo problema.

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 (breast.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 6.

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, non ci sono dei partizionamenti puri ma sono presenti alcuni partizionamenti quasi puri, come quello mostrato in Figura 1: la foglia associata al ramo "yes" di "irradiat" è una foglia pura in quanto tutti i record sono etichettati come "no-recurrence-events"; la foglia associata al ramo "no" di "irradiat" è una foglia quasi pura poiché la maggior parte dei record hanno "recurrence-events" come etichetta di classe.

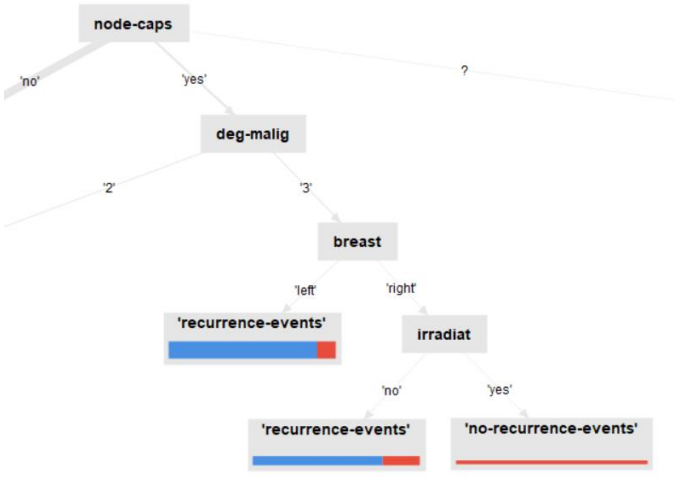

Figura 1

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 6. Impostando un valore di "maximal depth" minore di 7 (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. Questo fenomeno si evince dalle Figura 4 (configurazione di partenza), Figura 7 e Figura 8 in cui si è mantenuto costante il parametro "minimal gain" e si è man mano decrementato il parametro "maximal depth".

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 (Figura 2). 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 (Figura 2, Figura 3, Figura 4 e Figura 5).

Infine, in Figura 6 è rappresentato un albero di decisione in cui sono è stato diminuito il parametro "maximal depth" ed aumentato leggermente il parametro "minimal gain", generando un albero più basso e con un livello di dettaglio inferiore rispetto a quello di partenza.

Figura 2 - Decision Tree - Maximal depth = 10; Minimal gain = 0.1

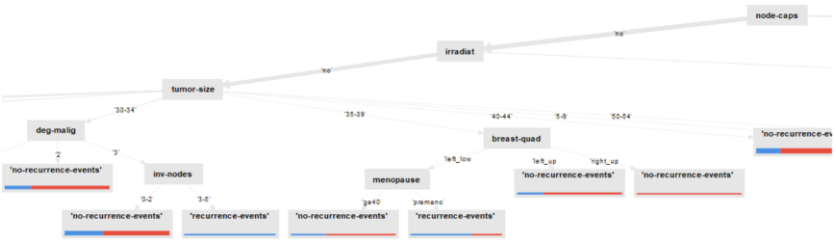

Figura 3 - Decision Tree - Maximal depth = 10; Minimal gain = 0.05

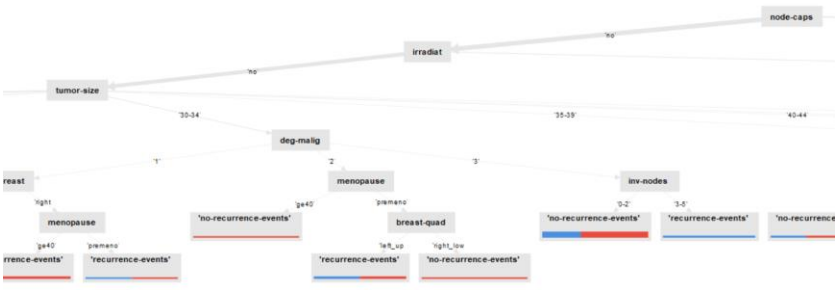

Figura 4 - Decision Tree - Maximal depth = 10; Minimal gain = 0.01

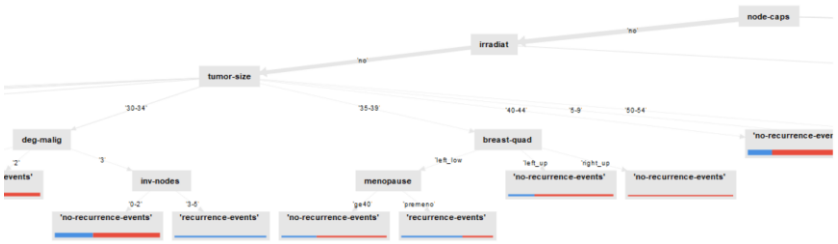

Figura 5 - Decision Tree - Maximal depth = 10; Minimal gain = 0.001

Figura 6 - Decision Tree - Maximal depth = 5; Minimal gain = 0.05

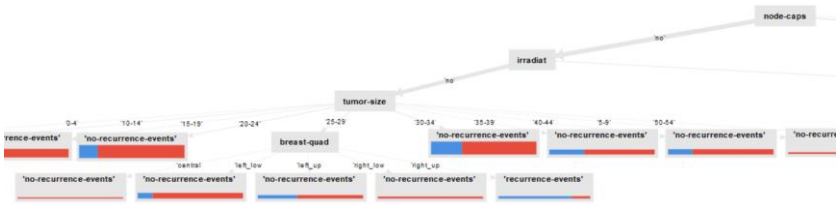

Figura 7 - Decision Tree - Maximal depth = 5; Minimal gain = 0.01

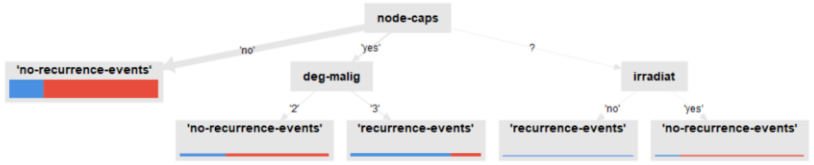

Figura 8 - Decision Tree - Maximal depth = 3; Minimal gain = 0.01

Domanda 3

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.

Da Figura 9, Figura 11, Figura 14 e Figura 15, si può evincere come l'accuratezza aumenti 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, che si può notare in Figura 10, Figura 11 e Figura 12, si verifica quando si aumenta o si mantiene costante il valore di "maximal depth" e si diminuisce progressivamente il valore di "minimal gain".

Infine, in Figura 13, si può notare un buono livello di accuratezza, con un albero di decisione né troppo dettagliato né troppo semplice, che permette di evitare il fenomeno dell'"underfitting" e dell'"overfitting".

accuracy: 70.30% +/- 1.43% (micro average: 70.28%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	0	0	0.00%
pred. 'no-recurrence-events'	85	201	70.28%
class recall	0.00%	100.00%	

Figura 9 - 10-fold Stratified Cross-Validation del Decision Tree - Maximal depth = 10; Minimal gain = 0.1

accuracy: 70.64% +/- 6.20% (micro average: 70.63%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	24	23	51.06%
pred. 'no-recurrence-events'	61	178	74.48%
class recall	28.24%	88.56%	

Figura 10 - 10-fold Stratified Cross-Validation del Decision Tree - Maximal depth = 10; Minimal gain = 0.05

accuracy: 67.48% +/- 6.59% (micro average: 67.48%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	37	45	45.12%
pred. 'no-recurrence-events'	48	156	76.47%
class recall	43.53%	77.61%	

Figura 11 - 10-fold Stratified Cross-Validation del Decision Tree - Maximal depth = 10; Minimal gain = 0.01

accuracy: 66.44% +/- 7.66% (micro average: 66.43%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	37	48	43.53%
pred. 'no-recurrence-events'	48	153	76.12%
class recall	43.53%	76.12%	

Figura 12 - 10-fold Stratified Cross-Validation del Decision Tree - Maximal depth = 10; Minimal gain = 0.001

accuracy: 72.03% +/- 6.22% (micro average: 72.03%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	26	21	55.32%
pred. 'no-recurrence-events'	59	180	75.31%
class recall	30.59%	89.55%	

Figura 13 - 10-fold Stratified Cross-Validation del Decision Tree - Maximal depth = 5; Minimal gain = 0.05

accuracy: 70.28% +/- 7.75% (micro average: 70.28%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	35	35	50.00%
pred. 'no-recurrence-events'	50	166	76.85%
class recall	41.18%	82.59%	

Figura 14 - 10-fold Stratified Cross-Validation del Decision Tree - Maximal depth = 5; Minimal gain = 0.01

accuracy: 74.82% +/- 6.64% (micro average: 74.83%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	24	11	68.57%
pred. 'no-recurrence-events'	61	190	75.70%
class recall	28.24%	94.53%	

Figura 15 - 10-fold Stratified Cross-Validation del Decision Tree - Maximal depth = 3; Minimal gain = 0.01

Domanda 4

Come si può notare dalle figure sottostanti, 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. Si raggiunge un massimo nell'accuratezza per k=10 per poi iniziare una progressiva attenuazione per valori di k maggiori di 10 (15, 20, …)

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.

accuracy: 66.44% +/- 7.28% (micro average: 66.43%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	30	41	42.25%
pred. 'no-recurrence-events'	55	160	74.42%
class recall	35.29%	79.60%	

Figura 16 - 10-fold Stratified Cross-Validation del K-NN con k = 1

accuracy: 70.26% +/- 7.23% (micro average: 70.28%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	27	27	50.00%
pred. 'no-recurrence-events'	58	174	75.00%
class recall	31.76%	86.57%	

Figura 17 - 10-fold Stratified Cross-Validation del K-NN con k = 3

accuracy: 73.77% +/- 5.98% (micro average: 73.78%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	26	16	61.90%
pred. 'no-recurrence-events'	59	185	75.82%
class recall	30.59%	92.04%	

Figura 18 - 10-fold Stratified Cross-Validation del K-NN con k = 5

accuracy: 74.84% +/- 6.23% (micro average: 74.83%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	25	12	67.57%
pred. 'no-recurrence-events'	60	189	75.90%
class recall	29.41%	94.03%	

Figura 19 - 10-fold Stratified Cross-Validation del K-NN con k = 7

accuracy: 75.20% +/- 5.43% (micro average: 75.17%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	25	11	69.44%
pred. 'no-recurrence-events'	60	190	76.00%
class recall	29.41%	94.53%	

Figura 20 - 10-fold Stratified Cross-Validation del K-NN con k = 10

accuracy: 74.13% +/- 5.67% (micro average: 74.13%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	18	7	72.00%
pred. 'no-recurrence-events'	67	194	74.33%
class recall	21.18%	96.52%	

Figura 21 - 10-fold Stratified Cross-Validation del K-NN con k = 15

accuracy: 73.79% +/- 5.61% (micro average: 73.78%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	17	7	70.83%
pred. 'no-recurrence-events'	68	194	74.05%
class recall	20.00%	96.52%	

Figura 22 - 10-fold Stratified Cross-Validation del K-NN con k = 20

Utilizzando il classificatore Naïve Bayes si ottiene (Figura 23) un'accuratezza media inferiore a quella del K-NN. Questo risultato può essere dovuto al fatto che gli attributi del dataset preso in esame (breast.xls) non sono indipendenti tra loro, che è l'ipotesi necessaria del classificatore Naïve Bayes affinché il risultato sia di buona qualità.

accuracy: 72.45% +/- 7.70% (micro average: 72.38%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	41	35	53.95%
pred. 'no-recurrence-events'	44	166	79.05%
class recall	48.24%	82.59%	

Figura 23 - 10-fold Stratified Cross-Validation del Naïve Bayes

Domanda 5

La Figura 24 mostra la matrice di correlazione ottenuta dal dataset analizzato (breast.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 per lo più 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. Ad esempio, per l'attributo "age" potrebbe aver senso assegnare un valore numerico a ciascun elemento nominale, ma utilizzando "unique integers" come "coding type" si otterrebbe un mapping senza che i valori dell'attributo "age" siano prima ordinati. Dunque, potrebbe accadere che il range "30-39" risulti maggiore rispetto a quello "40-49", il che porterebbe a risultati inconsistenti.

Quindi, con la sola matrice di correlazione, non è possibile stabile l'indipendenza degli attributi. Osservando i risultati precedentemente ottenuti (il classificatore Naïve Bayes ha un'accuratezza inferiore al K-NN) si può ipotizzare che gli attributi non siano indipendenti tra loro. Una probabile dipendenza potrebbe essere quella tra gli attributi "age" e "menopause", i quali per definizione sono strettamente correlati tra loro.

Attribut...	age	menopa...	tumor-s...	inv-nodes	node-ca...	deg-mal...	breast	breast-...	irradiat
age	1	?	?	?	?	?	?	?	?
menopa...	?	1	?	?	?	?	?	?	?
tumor-size	?	?	1	?	?	?	?	?	?
inv-nodes	?	?	?	1	?	?	?	?	?
node-caps	?	?	?	?	1	?	?	?	?
deg-malig	?	?	?	?	?	1	?	?	?
breast	?	?	?	?	?	?	1	?	-0.019
breast-q...	?	?	?	?	?	?	?	1	?
irradiat	?	?	?	?	?	?	-0.019	?	1

Figura 24 - Matrice di correlazione