Business Intelligence per i Big Data

Esercitazione 6 – Seconda Parte

BOZZA DI SOLUZIONE

Domanda 1

(a) Come mostrato in Figura 1, l'attributo più selettivo risulta essere "Age", perché rappresenta il nodo radice dell'albero di decisione.

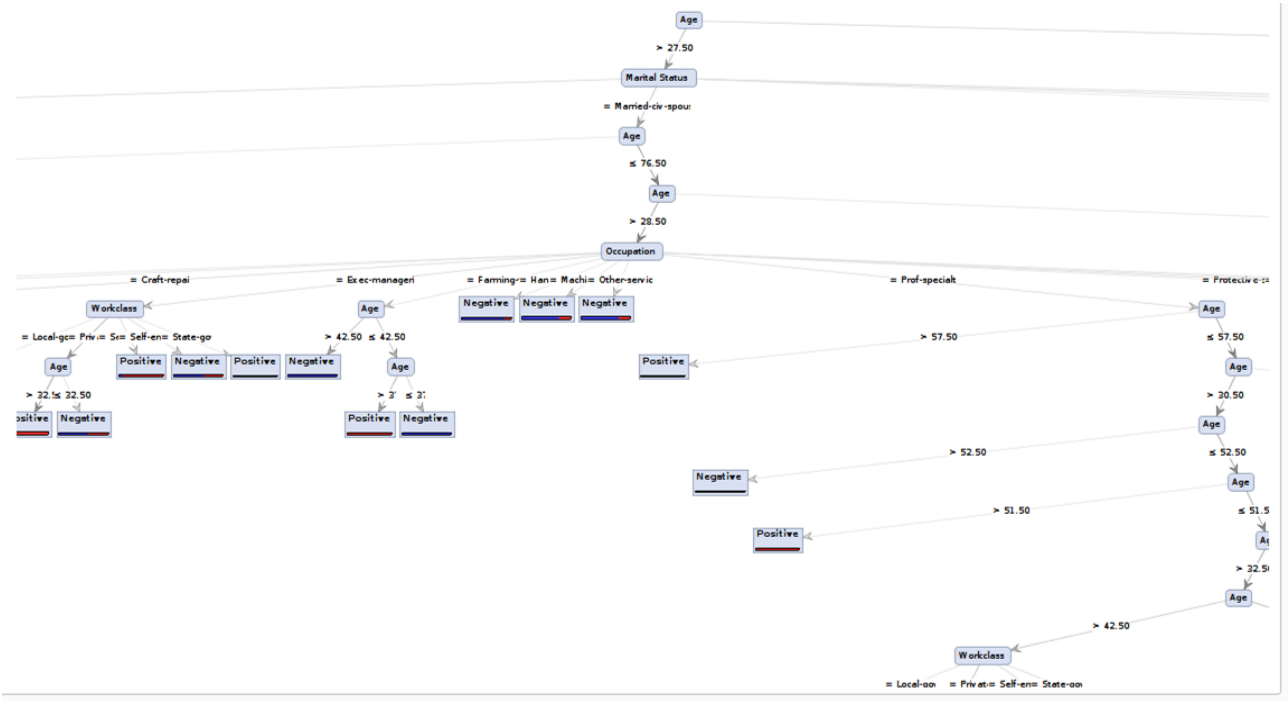

Figura 1

(b) L'altezza dell'albero, ovvero la lunghezza massima di un percorso che collega la radice ad una foglia dell'albero è 15.

(c) Un partizionamento puro è un split sui valori di un attributo tale per cui i record corrispondenti appartengono tutti alla medesima classe. Per esempio, consideriamo la porzione sinistra dell'albero di decisione rappresentato in Figura 2. I valori dell'attributo "Age" sono splittati in due gruppi: <= 57 and > 57. Mentre la prima partizione è "impura", perché copre record etichettati sia con la classe "Positive" sia con la classe "Negative", la seconda è pura perché tutte le relative istanze appartengono alla classe "Positive".

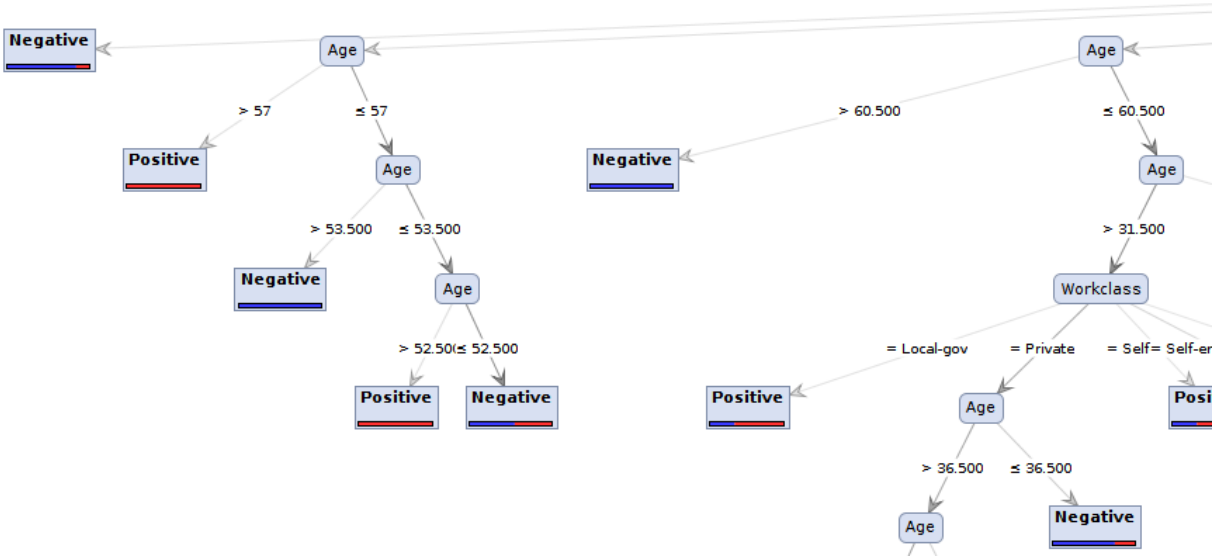

Figura 2

Domanda 2

Il parametro “maximal depth” permette di specificare l’altezza massima dell’albero di decisione generato. Usando la configurazione di default (con maximal depth = 20) l’altezza dell’albero risulta essere 15 e quindi il processo ricorsivo di learning dell’albero viene completato. Al contrario, settando un valore di maximal depth inferior a 15 (ad es. 5) la ricorsione viene interrotta e quindi la qualità del modello generato potenzialmente decresce.

Il parametro “minimal gain” permette di scegliere se splittare ulteriormente un nodo dell’albero oppure no. In particolare, un nodo viene splittato se il suo gain è superiore alla soglia minima (minimal gain). Valori elevati di minimal gain producono un numero limitato di partizionamenti e, di conseguenza, alberi di decisione più piccoli. Valori troppo elevati di minimal gain (ad es., 0.9) impediscono completamente lo split dei valori degli attributi e quindi l’albero risultante conterrà un singolo nodo. Mentre valori di minimal gain bassi (ad es., 0.01), producono generalmente uno splitting degli attributi abbastanza fitto.

Domanda 3

In generale, riducendo il valore del minimal gain e aumentando la maximal depth si genera un modello di classificazione più dettagliato e quindi più accurato. Tuttavia, sulla base dei risultati riportati nelle Figure -, impostando valori di maximal depth superiori a 20 e minimal gain inferiori a 0.005 si produce l’effetto denominato “overfitting”, ovvero il modello risulta troppo “focalizzato” sui dati di train per classificare in modo accurato nuovi dati di test. Mentre modelli con valori di maximal depth inferiori a 10 e minimal gain inferiori a 0.05 producono un modello troppo semplice per classificar e nuovi record.

accuracy: 79.60% +/- 2.62% (mikro: 79.60%)

	true Negative	true Positive	class precision
pred. Negative	698	134	83.89%
pred. Positive	70	98	58.33%
class recall	90.89%	42.24%	

Figura 3 – Decision Tree. Minimum gain = 0.05. Maximum depth = 10

accuracy: 76.60% +/- 0.49% (mikro: 76.60%)

	true Negative	true Positive	class precision
pred. Negative	766	232	76.75%
pred. Positive	2	0	0.00%
class recall	99.74%	0.00%	

Figura 4 – Decision Tree. Minimum gain = 0.05. Maximum depth = 5

accuracy: 79.60% +/- 2.62% (mikro: 79.60%)

	true Negative	true Positive	class precision
pred. Negative	698	134	83.89%
pred. Positive	70	98	58.33%
class recall	90.89%	42.24%	

Figura 4 – Decision Tree. Minimum gain = 0.005. Maximum depth = 20

Domanda 4

Incrementando il valore di K, il classificatore considera un numero maggiore di dati di train "vicini" al dato di test e quindi l'accuratezza media cresce: 73.20% con K=1, 77.00% con K=3, 77.40% con K=5, 78.30% con K=10 (Figure 5-10). Considerando un numero molto elevato di record di train "vicini" (ad es., K>10) la presenza di dati rumorosi comincia ad inficiare le performance di classificazione e dunque l'accuratezza media di classificazione diminuisce leggermente (Figura 11-12).

accuracy: 73.20% +/- 2.48% (mikro: 73.20%)

	true Negative	true Positive	class precision
pred. Negative	635	135	82.47%
pred. Positive	133	97	42.17%
class recall	82.68%	41.81%	

Figura 5 – K-NN. Matrice di confusione. K=1

accuracy: 77.00% +/- 4.49% (mikro: 77.00%)

	true Negative	true Positive	class precision
pred. Negative	671	133	83.46%
pred. Positive	97	99	50.51%
class recall	87.37%	42.67%	

Figura 6 – K-NN. Matrice di confusione. K=3

accuracy: 77.40% +/- 3.23% (mikro: 77.40%)

	true Negative	true Positive	class precision
pred. Negative	676	134	83.46%
pred. Positive	92	98	51.58%
class recall	88.02%	42.24%	

Figura 7 – K-NN. Matrice di confusione. K=5

accuracy: 76.60% +/- 2.65% (mikro: 76.60%)

	true Negative	true Positive	class precision
pred. Negative	669	135	83.21%
pred. Positive	99	97	49.49%
class recall	87.11%	41.81%	

Figura 8 – K-NN. Matrice di confusione. K=7

accuracy: 78.30% +/- 2.28% (mikro: 78.30%)

	true Negative	true Positive	class precision
pred. Negative	704	153	82.15%
pred. Positive	64	79	55.24%
class recall	91.67%	34.05%	

Figura 9 – K-NN. Matrice di confusione. K=8

accuracy: 78.20% +/- 3.16% (mikro: 78.20%)

	true Negative	true Positive	class precision
pred. Negative	704	154	82.05%
pred. Positive	64	78	54.93%
class recall	91.67%	33.62%	

Figura 10 – K-NN. Matrice di confusione. K=10

accuracy: 77.30% +/- 1.95% (mikro: 77.30%)

	true Negative	true Positive	class precision
pred. Negative	685	144	82.63%
pred. Positive	83	88	51.46%
class recall	89.19%	37.93%	

Figura 11 – K-NN. Matrice di confusione. K=15

accuracy: 76.60% +/- 3.23% (mikro: 76.60%)

	true Negative	true Positive	class precision
pred. Negative	675	141	82.72%
pred. Positive	93	91	49.46%
class recall	87.89%	39.22%	

Figura 12 – K-NN. Matrice di confusione. K=17

Come mostrato in Figura 13, Naïve Bayes ottiene un'accuratezza media più elevata di K-NN (80.50% contro78.30%) sul dataset analizzato.

accuracy: 80.50% +/- 3.88% (mikro: 80.50%)

	true Negative	true Positive	class precision
pred. Negative	635	62	91.10%
pred. Positive	133	170	56.11%
class recall	82.68%	73.28%	

Figura 13 – Naïve Bayes. Matrice di confusione.

Domanda 5

Figura 22 mostra la matrice di correlazione ottenuta dal dataset analizzato. Essa riporta la correlazione mutua (e simmetrica) tra coppie di attributi. Per esempio, l'attributo "Age" risulta essere molto correlato con l'attributo "Marital status" Dato che sussistono correlazioni significative tra attributi, ad es., tra "Age" e "Marital Status" (correlazione = 0.401), tra "Occupation" e "Workclass" (correlazione = 0.239), l'ipotesi Naïve risulta essere irrealistica per il dataset analizzato. Tuttavia, le performance di Naïve Bayes risultano essere mediamente buone (vedi riposta alla domanda precedente).

Attributes	Age	Workclass	Education	Marital Status	Occupation	Relationship	Race	Sex	Native Country
Age	1	0.083	0.023	0.401	-0.008	-0.202	-0.003	-0.090	-0.037
Workclass	0.083	1	0.101	0.048	0.239	0.031	0.062	0.034	-0.025
Education	0.023	0.101	1	0.047	0.135	0.048	0.037	0.025	0.073
Marital Status	0.401	0.048	0.047	1	-0.010	0.036	0.065	0.174	0.019
Occupation	-0.008	0.239	0.135	-0.010	1	0.007	0.014	-0.115	-0.019
Relationship	-0.202	0.031	0.048	0.036	0.007	1	0.121	0.210	0.061
Race	-0.003	0.062	0.037	0.065	0.014	0.121	1	0.129	0.153
Sex	-0.090	0.034	0.025	0.174	-0.115	0.210	0.129	1	-0.028
Native Country	-0.037	-0.025	0.073	0.019	-0.019	0.061	0.153	-0.028	1

Figura 14 – Matrice di correlazione