Business Intelligence per Big Data

Esempi di domande di teoria
Analisi di dati

AA 2022-2023 - *Politecnico di Torino*

Domanda 1

Quale delle seguenti affermazioni relative al clustering è corretta?

- (a) utilizzando l'algoritmo k-means, il centroide di un cluster corrisponde sempre ad un punto appartenente a quel cluster
- (b) nell'algoritmo DBSCAN, l'appartenenza di un punto a un determinato cluster è caratterizzata da un peso tra 0 e 1
- (c) il risultato dell'algoritmo DBSCAN può essere visualizzato tramite un dendrogramma
- (d) Nell'algoritmo k-means, inizializzazioni diverse dei centroidi possono produrre come risultato cluster diversi
- (e) quando outliers e punti di rumore sono presenti nei dati, è più opportuno utilizzare l'algoritmo k-means rispetto a DBSCAN
- (f) per ridurre il valore di SSE (sum of squared error) bisogna ridurre il valore di K dell'algoritmo k-means

Domanda 2

La policy di MAX (complete) linkage prevede che la distanza fra due cluster X e Y sia calcolata come:

$$dist(X, Y) = max_{x \in X, y \in Y} dist(x, y)$$

dove $dist(x, y)$ e' una distanza che puo' essere definita fra coppie di punti.

Per un dataset di 5 punti viene calcolata la seguente matrice di distanze.

	a	b	c	d	e
a	0	20	11	16	8
b	20	0	13	24	2
c	11	13	0	19	22
d	16	24	19	0	3
e	8	2	22	3	0

Si applica il clustering gerarchico agglomerativo per estrarre 3 cluster. Viene utilizzata la policy di “MAX linkage” (complete linkage).

Quali sono i 3 cluster ottenuti?

- ○ (a) {a, d}, {b, e}, {c}
- ○ (b) {d}, {b, e}, {a, c}
- ○ (c) {a}, {b}, {c, d, e}
- ○ (d) {a}, {c}, {b, d, e}
- ○ (e) Non e’ possibile rispondere alla domanda con le informazioni a disposizione
- ○ (f) {b}, {d}, {a, c, e}
- ○ (g) Nessuna delle altre risposte e’ corretta
- ○ (h) {a, b}, {c, d}, {e}

Domanda 3

La metrica **MAX** (o complete linkage) nel clustering gerarchico agglomerativo prevede che:

- ○ (a) Due cluster $\mathbf{C_1}$, $\mathbf{C_2}$ vengono uniti se esiste una coppia di punti $\mathbf{p_1} \in \mathbf{C_1}$, $\mathbf{p_2} \in \mathbf{C_2}$ la cui distanza è la maggiore nella matrice delle distanze
- ○ (b) Un cluster **C** verrà unito ad un singolo punto **p** se la distanza tra **p** e **C** è quella massima nella matrice delle distanze
- ○ (c) I cluster ottenuti siano molto sensibili al rumore
- ○ (d) Nessuna delle risposte è corretta
- ○ (e) Un cluster **C** verrà unito ad un singolo punto **p** se la massima distanza tra **p** ed i punti di **C** è la maggiore nella matrice delle distanze
- ○ (f) I primi due punti ad unirsi nel dendrogramma sono quelli più distanti tra loro

Domanda 4

Quale delle seguenti metriche **non** è indicata per valutare le performance di un algoritmo di classificazione?

- ○ (a) Silhouette Index
- ○ (b) Accuracy
- ○ (c) Recall
- ○ (d) Matrice di confusione
- ○ (e) F-measure
- ○ (f) Precision

Domanda 5

- Precisione(C) è la frazione di predizione corrette rispetto a tutte le predizioni fatte per la classe C
- Recall(C) è la frazione di predizioni corrette rispetto a tutti i punti che appartengono alla classe C

Vengono dati due vettori, y_pred, y_true, che contengono le predizioni effettuate da un classificatore e la ground truth, rispettivamente.

```
y_true: [B A A C A B B B C B]
y_pred: [A A B A B C A B C C]
```

Quali sono precisione e richiamo per la classe A?

- ○ (a) Precisione: 0.3333, Richiamo: 0.5
- ○ (b) Nessuna delle altre risposte è corretta
- ○ (c) Precisione: 0.3333, Richiamo: 0.25
- ○ (d) Precisione: 0.25, Richiamo: 0.3333
- ○ (e) Precisione: 0.3, Richiamo: 0.5
- ○ (f) Precisione: 0.5, Richiamo: 0.3333
- ○ (g) Precisione: 0.3333, Richiamo: 0.2
- ○ (h) Precisione: 0.3, Richiamo: 0.25
- ○ (i) Precisione: 0.2, Richiamo: 0.3333

Domanda 6

Data la matrice di confusione in figura, quale delle seguenti affermazioni **non è** corretta?

		Predicted T	Predicted F
Actual	T	90	0
	F	10	0

- (a) Il richiamo della classe F è del 10%
- (b) Il richiamo della classe T è del 100%
- (c) Tutti i 100 elementi vengono etichettati come classe T
- (d) L'accuratezza del modello è del 90%
- (e) La precisione della classe T è del 90%
- (f) La precisione della classe F è 0

Domanda 7

Un classificatore binario viene addestrato per distinguere immagini di gatti da immagini di cani. Il test set usato per valutare le performance del modello è bilanciato, con 10,000 immagini di cani e 10,000 immagini di gatti.

Il classificatore predice solamente 50 immagini come appartenenti alla classe “gatto”. Tutte queste predizioni sono corrette.

Quale delle seguenti affermazioni riguardanti il classificatore è vera?

- ○ (a) Ha recall alta per la classe “cane”
- ○ (b) Ha recall alta per la classe “gatto”
- ○ (c) Ha F1 score alto per la classe “cane”
- ○ (d) Ha precisione bassa per la classe “gatto”
- ○ (e) Ha precisione alta per la classe “cane”
- ○ (f) Ha accuratezza alta
- ○ (g) Nessuna delle altre affermazioni è corretta
- ○ (h) Ha F1 score alto per la classe “gatto”

Domanda 8

Dato il seguente dataset transazionale

ABCD
BCE
ABDE
BCDE
BCDE
BCD
E
BD
BD
ABDE

Scrivere la Header Table per FP-Tree con MinSup>2.

Domanda 9

Dato il seguente dataset transazionale

ABCD
BCE
ABDE
BCDE
BCDE
BCD
E
BD
BD
ABDE

Scrivere FP-Tree con MinSup>2. In particolare, riportare l'elenco dei percorsi che caratterizzano FP-Tree. Per ogni percorso specificare la sequenza di nodi nella forma (elemento, supporto locale).

Domanda 10

Dato il seguente dataset transazionale

ABCD
BCE
ABDE
BCDE
BCDE
BCD
E
BD
BD
ABDE

Scrivere la Node link chain per l'item E

Domanda 11

The following transactional database is given:

	Transactions
0	BC
1	ADE
2	ADE
3	ABC
4	CE
5	BC
6	BD
7	ADE
8	ABCE
9	AE

Apply the Apriori algorithm to extract frequent itemsets. Use minsup = 2 (an itemset is frequent if it appears in at least 2 transactions).

What are the length-3 itemsets generated by Apriori **after the join step and prune step** (applying the Apriori principle), **before counting the support** in the database?

- ○ (a) ABC, ABD, ACE, ADE
- ○ (b) It is not possible to answer the question with the available information.
- ○ (c) ABC, ADE
- ○ (d) ABD, ABE, ACD, ECD
- ○ (e) ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE
- ○ (f) ABC, ABD, ABE, ACD, ACE, ADE, ECD
- ○ (g) ABC, ABE
- ○ (h) None of the other answers is correct.
- ○ (i) ABC, ACE, ADE