Fondamenti di Data Science e Machine Learning - Prof. G. Polese - Anno Accademico 2020/21

Prova Scritta (ore 10:00) 09/06/2021

Cognome e Nome: **Matricola:**

Esercizio 1 (punti 6 su 30)

Data la seguente signature matrix:

Shingle	**S_1**	**S_2**	**S_3**	**S_4**
0	1	1	0	1
1	0	0	1	1
2	1	1	0	0
3	0	0	1	0
4	0	1	1	0
5	1	0	0	0

a) (*1 punto*) Calcolare la similarità di Jaccard tra ogni coppia di colonne;

b) (*3 punti*) Calcolare la signature di ogni colonna usando le seguenti 4 funzioni hash:

$h1(x) = (x^4 + 1) \bmod 6$; $h2(x) = (x^2 + 3) \bmod 6$; $h3(x) = (x^3 + 7) \bmod 6$; $h4(x) = (3x^2 + 5) \bmod 6$.

Mostrare l'evoluzione della matrice delle signature di minhash simulando l'esecuzione dell'algoritmo per il loro calcolo. Inoltre, calcolare le similarità di Jaccard tra tutte le coppie di signature di minhash.

c) (2 *punti*) Sulla base delle similarità di Jaccard calcolate al punto b), calcolare per ogni coppia di colonne la probabilità che la coppia venga selezionata per il confronto, ipotizzando che la matrice delle signature sia partizionata in 2 bande di 2 righe ciascuna.

Esercizio 2 (punti 8 su 30)

Dato il seguente dataset di training, dove l'attributo *Gioca* è quello dipendente:

Tempo	**Temperatura**	**Umidità**	**Vento**	**Gioca**
Soleggiato	Calda	Alta	NO	*NO*
Soleggiato	Calda	Alta	SI	*NO*
Nuvoloso	Calda	Alta	NO	*SI*
Piovoso	Mite	Alta	NO	*SI*
Piovoso	Fredda	Normale	NO	*SI*
Piovoso	Fredda	Normale	SI	*NO*
Nuvoloso	Fredda	Normale	SI	*SI*
Soleggiato	Mite	Alta	NO	*NO*
Soleggiato	Fredda	Normale	NO	*SI*
Piovoso	Mite	Normale	NO	*SI*
Soleggiato	Mite	Normale	SI	*SI*
Nuvoloso	Mite	Alta	SI	*SI*
Nuvoloso	Calda	Normale	NO	*SI*
Piovoso	Mite	Alta	SI	*NO*

Costruire un albero di decisione ternario su 3 livelli (radice ed altri 2 livelli), scegliendo per ogni livello l'attributo che più di tutti migliora l'entropia, ipotizzando di utilizzare per ogni attributo un criterio di splitting binario o ternario che verifichi semplicemente l'occorrenza dei singoli valori possibili (2 o 3 per attributo). Inoltre, per ogni nodo dell'albero indicare la distribuzione dei campioni ed il valore dell'entropia.

Esercizio 3 (punti 6 su 30)
Facendo riferimento al dataset dell'esercizio 2

a) (*punti 2)* Trovare eventuali FD con RHS singolo;

b) (*punti 4)* Trovare eventuali RFD con RHS singolo che rilassano sull'extent e con un g3-error inferiore al 15%.

Esercizio 4 (punti 3 su 30)
Utilizzando la tecnica OneHotEncoding trasformare gli attributi categorici a 3 valori del dataset dell'esercizio 2 in attributi numerici.

Esercizio 5 (punti 7 su 30)
Dati i seguenti punti in uno spazio bidimensionale:

(3,5)(3, 3)(2,4)(6,3)(5,3)(4,2)(1,2)(1,1)(2,1)

Usare l'algoritmo *K-means* per suddividere tali punti in 3 cluster, utilizzando la distanza euclidea e svolgendo 2 iterazioni dell'algoritmo. Per la scelta dei 3 centroidi alla prima iterazione utilizzare il metodo che prevede la selezione del solo primo centroide in modo casuale (scegliere il punto più vicino a quello le cui coordinate si ottengono sommando tutte le prime componenti e poi le seconde dei 9 punti forniti), mentre i restanti 2 il più lontano possibile dai centroidi già scelti.