Fondamenti di Data Science e Machine Learning - Prof. G. Polese - Anno Accademico 2019/20

Prova Scritta (ore 9:00) 15/06/2020

Cognome e Nome: **Matricola:**

Esercizio 1 (punti 7 su 30)

Data la seguente tabella:

Tuple number	Height	Weight	Shoe size
1	175	70	40
2	175	75	39
3	175	69	40
4	176	71	40
5	178	81	41
6	169	73	37
7	170	62	39

Relativamente agli attribute *Height*, *Weight* e *Shoe size*:

a) Elencare eventuali FD tra tali attributi, con lato destro a singolo attributo

b) Elencare eventuali RFD con RHS singolo che rilassano solo sul confronto, solo sull'extent, o su entrambi, mostrando le relative soglie di similarità o della misura di coverage g3 error

Esercizio 2 (punti 7 su 30)

Data la seguente signature matrix:

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

a) Calcolare la similarità di Jaccard tra ogni coppia di colonne;

b) Calcolare la signature di minhash per ogni colonna usando le seguenti 3 funzioni hash:

$h_1(x) = (2x + 1) \bmod 6$; $h_2(x) = (3x + 2) \bmod 6$; $h_3(x) = (5x + 2) \bmod 6$.

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.

Fondamenti di Data Science e Machine Learning - Prof. G. Polese - Anno Accademico 2019/20

Prova Scritta (ore 9:00) 15/06/2020

Esercizio 3 (punti 6 su 30)

Date la seguente tabella di transazioni su un sito di commercio elettronico:

Trans.ID	Data	Prodotto	Quantità
11	11/3/2020	Computer	1
11	11/3/2020	Borsa computer	1
11	11/3/2020	Mouse	1
12	12/3/2020	Televisore	1
12	12/3/2020	Staffa TV	1
12	12/3/2020	Antenna	1
13	15/3/2020	Computer	1
13	15/3/2020	Borsa computer	1
14	18/3/2020	Televisore	1
14	18/3/2020	Computer	1
14	18/3/2020	Mouse	1
14	18/3/2020	Borsa computer	1

a) Elencare i frequent item sets con l'algoritmo Apriori supporto di almeno il 50%
b) Trovare tutte le regole di associazione con supporto e confidenza di almeno il 50%

Esercizio 4 (punti 6 su 30)
Dati i seguenti punti in uno spazio bidimensionale:

(1,2)(2, 4)(2,5)(3,4)(3,5)(5,2)(6,3)(6,8)(9,3)(11,2)(10,4)(12,3)

a) Mostrare i passi di un algoritmo di clustering gerarchico (mostrando ad ogni passo cluster e centroidi) per raggruppare i suddetti punti in 2 cluster, usando la funzione di distanza euclidea.
b) Come a) ma usando la distanza del coseno.

Esercizio 5 (punti 4 su 30)
Supponendo che i primi 6 punti dell'esercizio 4 siano etichettati come *True* ed i restanti come *False*, classificare i seguenti punti utilizzando l'algoritmo *KNN* prima con K=3 e poi con K=5:

(6,10)(4,8)(7,9)(9,10)