

Cognome e Nome:

Matricola:

*(Scrivere solo nello spazio bianco. Se necessario, usare il retro del foglio.**Non sono ammessi elaborati su fogli diversi.)***Esercizio 1 (punti 7 su 30)**

Dato il seguente dataset:

Item	X <sub>1</sub>	X <sub>2</sub>	X <sub>3</sub>	X <sub>4</sub>	Y
1	2	3	-1	-2	0
2	1	6	4	5	1
3	-5	2	-1	-4	0
4	-7	3	-5	4	1
5	-10	8	-9	6	1
6	-18	17	-12	2	0
7	1	5	-10	7	1
8	6	-5	-8	3	0

Disegnare una rete neurale (eventualmente anche con un solo neurone) che, presi in input i valori degli attributi predittivi  $X_1$ ,  $X_2$ ,  $X_3$ ,  $X_4$ , di una delle righe, più un eventuale termine di bias, produca il valore  $Y$  della stessa riga. Si richiede quindi di specificare i pesi, l'eventuale termine di bias e le (o la) funzioni (o funzione) di attivazione.

**Esercizio 2 (punti 5 su 30)**

Usando il coefficiente di Pearson, determinare l'attributo predittivo che sia maggiormente correlato all'attributo dipendente Y. Per chi non la ricordasse, la formula per il calcolo del coefficiente di Pearson è la seguente:

$$PC = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

- ▶ Dove  $n$  nel nostro caso è 8, cioè il numero di istanze;
- ▶  $\bar{x}$  è la media di tutti i valori di una variabile  $X_i$ ;
- ▶  $s_x$  è la sua deviazione standard:  $s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$

**Esercizio 3 (punti 6 su 30)**

Facendo sempre riferimento al dataset dell'esercizio 1, per ognuna delle 4 dipendenze candidate del tipo  $X_i \rightarrow Y$ , dire se vale come FD o, in caso contrario, dire se vale come RFD, nel qual caso specificare le soglie numeriche di similarità e/o le soglie per il g3 error, a seconda che si tratti di RFD che rilassa sul confronto, sull'extent, o su entrambi.

**Esercizio 4 (punti 7 su 30)**

Data la seguente signature matrix:

Shingle	S <sub>1</sub>	S <sub>2</sub>	S <sub>3</sub>	S <sub>4</sub>
0	1	0	0	0
1	0	0	1	1
2	0	1	0	0
3	0	1	1	0
4	1	0	1	1
5	1	0	0	1

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:

$$h_1(x) = (4x^2 + 1) \bmod 6; \quad h_2(x) = (x^3 + 2) \bmod 6; \quad h_3(x) = (x^4 + 3) \bmod 6; \quad h_4(x) = (2x + 4) \bmod 6.$$

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), per ogni coppia di colonne calcolare la probabilità che la coppia venga selezionata per il confronto, ipotizzando che la matrice delle signature sia partizionata in 2 bande di 3 righe ciascuna.

**Esercizio 5 (punti 5 su 30)**

Dato il seguente dataset di transazioni:

Transazione	Data	Prodotto	Categoria	Quantità
111	01/05/2022	Pane	Salumeria	1
111	01/05/2022	Insalata	Ortofrutta	3
111	01/05/2022	Banane	Ortofrutta	6
112	03/06/2022	Mozzarella	Salumeria	1
112	03/06/2022	Pane	Salumeria	1
112	03/06/2022	Insalata	Ortofrutta	1
113	10/05/2022	Mozzarella	Salumeria	1
113	10/05/2022	Pane	Salumeria	1
114	01/06/2022	Mozzarella	Salumeria	2
114	01/06/2022	Pane	Salumeria	2
114	01/06/2022	Banane	Ortofrutta	4

Usando l'algoritmo Apriori, rispondere alle seguenti domande:

- Estrarre regole di associazione con supporto e confidenza di almeno il 50% che permettano di capire quali prodotti incentivano la vendita di pane?;
- Usando le categorie, trovare le regole di associazione che hanno supporto e confidenza di almeno il 75%.