

# Esempi di domande di teoria

## Analisi di dati

### Business Intelligence per Big data



Politecnico  
di Torino

AA. 2025-2026  
*Politecnico di Torino*

# Q1 - Clustering



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

# Q2 – Clustering gerarchico MAX linkage

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

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

dove  $\text{dist}(x, y)$  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}

# Q3 – Clustering gerarchico MAX linkage



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

---

- (a) Due cluster  $C_1, C_2$  vengono uniti se esiste una coppia di punti  $p_1 \in C_1, p_2 \in C_2$  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

# Q4 – Metriche classificazione



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

# Q5 - Precisione e richiamo



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

# Q6 – Precisione e richiamo



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

		Predicted	
		T	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

# Q7 - Precisione e richiamo



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”

# Q8 - FP-Growth Header Table



- Dato il seguente dataset transazionale

ABCD

BCE

ABDE

BCDE

BCDE

BCD

E

BD

BD

ABDE

Scrivere la Header Table per FP-Tree con  $\text{MinSup} > 2$ .

# Q9 – FP-Growth Header Table



- Dato il seguente dataset transazionale

ABCD

BCE

ABDE

BCDE

BCDE

BCD

E

BD

BD

ABDE

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

# Q10 - Apriori



Dato il seguente dataset di transazioni →

Applicare l'algoritmo Apriori con un  $\text{minsup} > 2$ . Quali sono gli itemset di lunghezza 1, 2 e 3 che vengono generati da Apriori dopo i passi di join e prune (con principio Apriori), prima del conteggio del supporto nella base dati?

Transactions	
0	B C
1	A B C
2	C D
3	A C
4	C D
5	A B E
6	A B C
7	D E
8	A B C D
9	A D E

# Q11 - Valutazione classificatore



Stiamo lavorando su un dataset composto da 20 campioni contenute 4 differenti classi. Il vettore seguente rappresenta la ground truth per le classi da 0 a 3:

gt = [2, 3, 2, 0, 1, 3, 0, 3, 0, 3, 3, 2, 1, 1, 0, 0, 0, 3, 3, 3]

Un classificatore predice il seguente vettore di classi:

pr = [2, 0, 0, 3, 2, 0, 0, 0, 0, 2, 0, 2, 3, 1, 2, 3, 3, 2, 0, 3]

Quale delle seguenti affermazioni è vera?

- (a) La recall della classe 0 è più bassa della precision della stessa classe
- (b) La precisione media supera lo 0.4
- (c) L'accuratezza supera lo 0.4
- (d) Nessuna risposta è corretta
- (e) La precisione per la classe 0 è la più bassa delle precisioni tra tutte le classi

# Q12 - Clustering



Abbiamo a disposizione 5 punti con le seguenti coordinate (x, y):

A (0, 0)

B (0, 4)

C (6, 2)

D (4, 2)

E (2, 4)

Vogliamo applicare il clustering di tipo K-Means per 2 iterazioni con i seguenti centroidi iniziali: (0, 0) e (5, 5). La metrica di distanza è

$d(a, b) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$  dove a e b sono due punti.

Quali cluster si ottengono?

(a) {A, E} {B, C, D}

(b) {A, B} {C, D, E}

(c) Nessuna risposta è corretta

(d) {A, B, E} {C, D}

(e) {A, E} {B, C, D}

# Q13 - DBSCAN



For two n-dimensional points  $P_1 = (P_{11}, P_{12}, \dots, P_{1n})$  and  $P_2 = (P_{21}, P_{22}, \dots, P_{2n})$  the Manhattan distance is defined as follows:

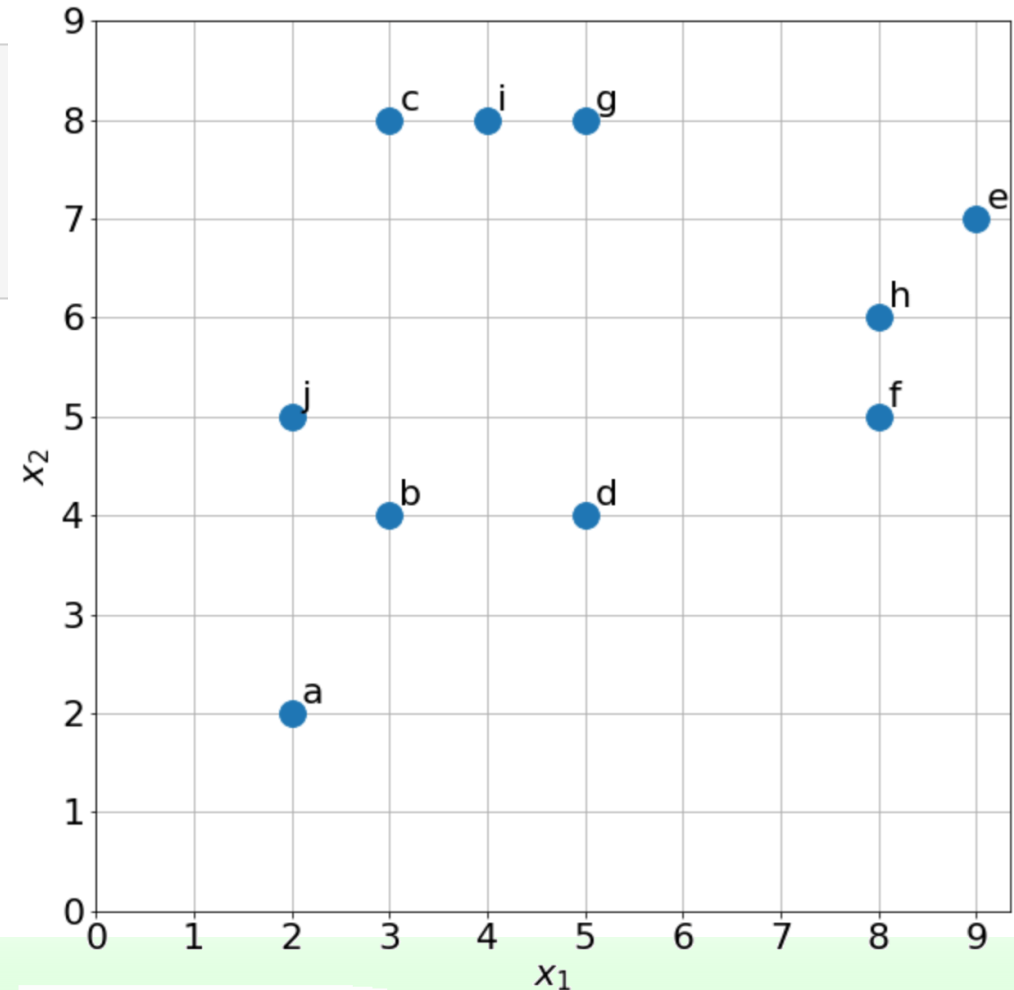
$$\text{dist}(P_1, P_2) = \sum_i |P_{1i} - P_{2i}|$$

Using the Manhattan distance, apply the DBSCAN algorithm to the following points in the bidimensional space.

Use the following hyperparameters:  $\epsilon = 2.5$ , minpoints=2 (at least 2 points as neighbors)

For each point write:

- The assigned label (N=noise, B=border, C=core)
- The assigned cluster id (order of cluster ids is not important, use -1 for noise points)



# Q14 - Precision/recall



A random forest classifier has been trained on a 2-dimensional dataset (features  $X_1$ ,  $X_2$ ). Each point in the dataset is labelled as either A, B or C (star, cross, triangle respectively).

The following figure represents a test set that is used to validate the classifier.

The decision boundaries of the model are shown in the figure:

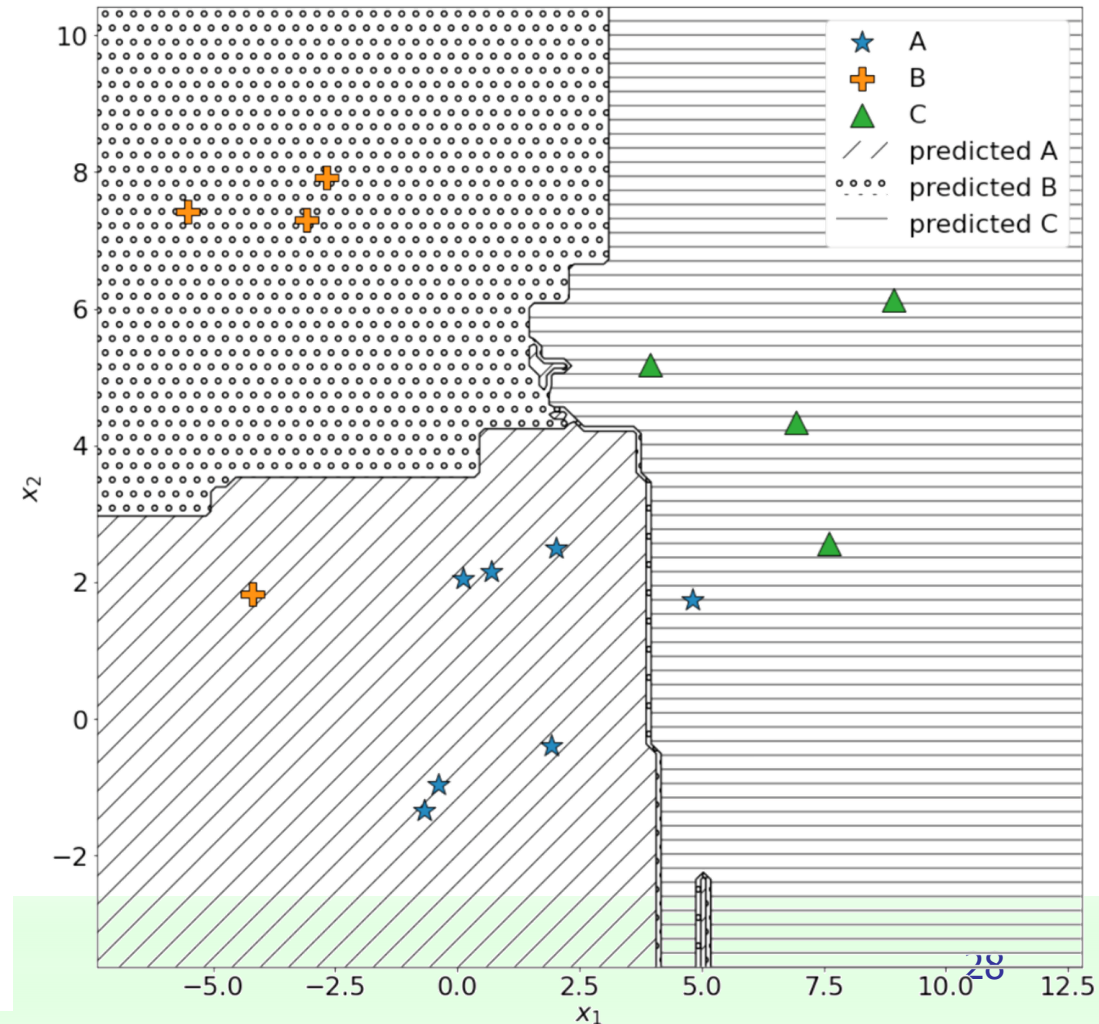
- Diagonal lines represent areas of the input space where the model predicts class A
- Small circles represent areas of the input space where the model predicts class B
- Horizontal lines represent areas of the input space where the model predicts class C

Write in the box below:

**precision(B)**

**precision(C)**

**recall(B)**



# Q15 - K-means

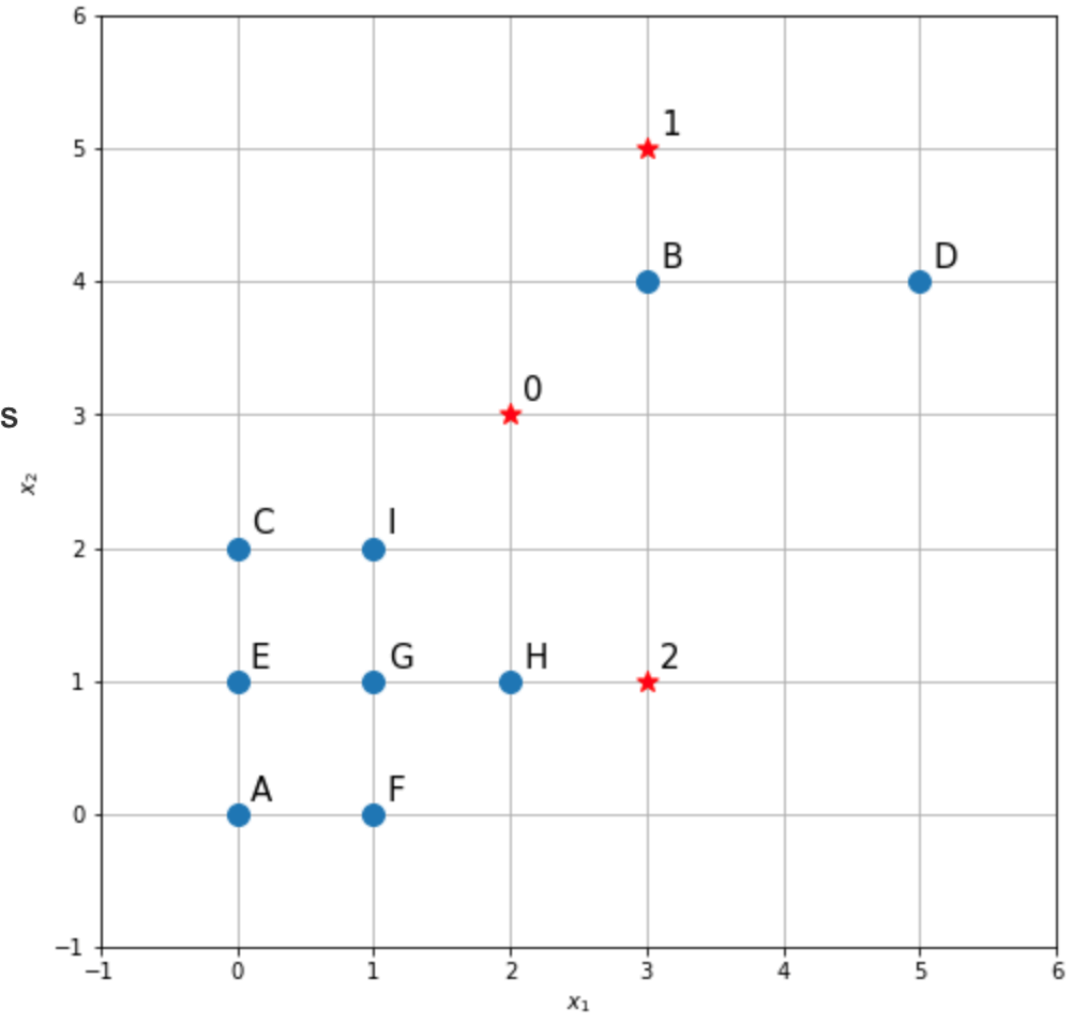


You are given a dataset containing 9 points in 2 dimensions ( $x_1, x_2$ ). Each point is labelled A through I.

You apply K-means clustering with  $K = 3$ . The figure below represents the 9 points (blue dots) and the initializations for the three centroids (red stars). Each centroid is labelled with a number (0, 1, 2).

What are the new centroids computed after 1 iteration of the K-means algorithm?

Use the Euclidean distance when computing any distance.



# ESERCIZI MONGODB

# MongoDB 1



Dalla documentazione di MongoDB sull'operatore \$in:

## \$in

Sintassi: { field: { \$in: [ valore1, valore2, ... valoreN ] } }

L'operatore \$in seleziona i documenti in cui il valore *field* e' uguale a uno dei valori nell'array specificato

Viene data la seguente collection MongoDB.

Eseguendo la seguente query, quale risultato viene ritornato?

```
db.collection.count({
  $or: [
    {
      firstName: "John"
    },
    {
      occupation: {
        $in: [
          "consultant",
          "HR"
        ]
      }
    }
  ],
  yearOfBirth: {
    $gte: 1990,
    $lte: 1995
  }
})
```

```
[
  {
    "firstName": "John",
    "lastName": "Smith",
    "yearOfBirth": 1990,
    "occupation": "accountant"
  },
  {
    "firstName": "Mike",
    "lastName": "Brown",
    "yearOfBirth": 1991,
    "occupation": "HR"
  },
  {
    "firstName": "Mike",
    "lastName": "Williams",
    "yearOfBirth": 1992,
    "occupation": "HR"
  },
  {
    "firstName": "Mary",
    "lastName": "Smith",
    "yearOfBirth": 1993,
    "occupation": "accountant"
  },
  {
    "firstName": "Robert",
    "lastName": "Williams",
    "yearOfBirth": 1994,
    "occupation": "software engineer"
  },
  {
    "firstName": "Jennifer",
    "lastName": "Davis",
    "yearOfBirth": 1987,
    "occupation": "db administrator"
  },
  {
    "firstName": "Sarah",
    "lastName": "Davis",
    "yearOfBirth": 1988,
    "occupation": "consultant"
  },
  {
    "firstName": "Lisa",
    "lastName": "Brown",
    "yearOfBirth": 1989,
    "occupation": "consultant"
  }
]
```

# MongoDB 2



Viene data una collection "employees", contenente le informazioni dei dipendenti di un'azienda. Per ciascun dipendente, informazioni su età, dipartimento e salario sono note. Il seguente è un esempio di documento estratto dalla collection:

```
{
  "_id" : ObjectId("62b97ce2850f3cffcbaab699"),
  "first" : "SUSAN",
  "last" : "DAVIS",
  "age" : 26,
  "compensation" : 65000,
  "department" : "HR"
}
```

Si vuole estrarre da questa collection il compenso medio dei dipendenti con età minore di 25 anni, separatamente per ogni dipartimento. Quale delle seguenti query soddisfa la richiesta?

(a)

```
db.employees.aggregate([
  {
    $match: {
      "age": {
        $lt: 25
      },
    }
  },
  {
    $group: {
      _id: null,
      avg: {
        $avg: "$compensation"
      }
    }
  }
])
```

(c)

```
db.employees.aggregate([
  {
    $group: {
      _id: "$department",
      avg: {
        $avg: "$compensation"
      }
    }
  },
  {
    $match: {
      "age": {
        $gt: 25
      },
    }
  }
])
```

(e)

```
db.employees.aggregate([
  {
    $match: {
      "age": {
        $gt: 25
      },
    }
  },
  {
    $group: {
      _id: "$department",
      avg: {
        $avg: "$compensation"
      }
    }
  }
])
```

(g)

```
db.employees.aggregate([
  {
    $group: {
      _id: null,
      avg: {
        $avg: "$compensation"
      }
    }
  },
  {
    $match: {
      "age": {
        $lt: 25
      },
    }
  }
])
```

(b)

```
db.employees.aggregate([
  {
    $match: {
      "age": {
        $gt: 25
      },
    }
  },
  {
    $group: {
      _id: null,
      avg: {
        $avg: "$compensation"
      }
    }
  }
])
```

(d)

```
db.employees.aggregate([
  {
    $match: {
      "age": {
        $lt: 25
      },
    }
  },
  {
    $group: {
      _id: "$department",
      avg: {
        $avg: "$compensation"
      }
    }
  }
])
```

(f)

```
db.employees.aggregate([
  {
    $group: {
      _id: null,
      avg: {
        $avg: "$compensation"
      }
    }
  },
  {
    $match: {
      "age": {
        $gt: 25
      },
    }
  }
])
```

(h)

```
db.employees.aggregate([
  {
    $group: {
      _id: "$department",
      avg: {
        $avg: "$compensation"
      }
    }
  },
  {
    $match: {
      "age": {
        $lt: 25
      },
    }
  }
])
```

