Esercizi

Business Intelligence per Big Data

Parte II

Sol

Valutazione classificatore

Stiamo lavorando su un dataset composto da 20 campioni contente 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

Sol

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 }

Apriori

Dato il seguente dataset di transazioni ➔

	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

Applicare l'algoritmo Apriori con un 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?

MongoDB

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?

MongoDB 2

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"
  }
]
```

(a)

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

(b)

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

(c)

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

(d)

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

(e)

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

(f)

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

(g)

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

(h)

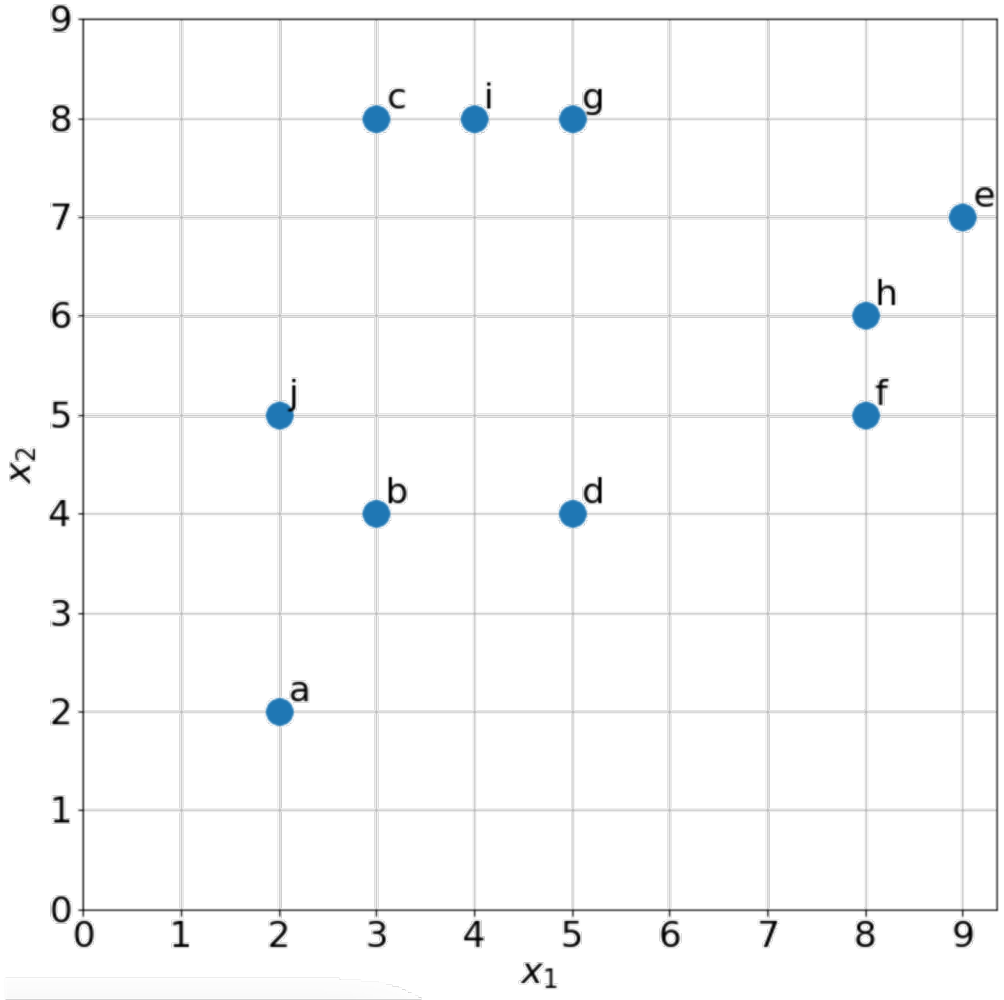
```
db.employees.aggregate([
  {
    $group: {
      _id: "$department",
      avg: {
        $avg: "$compensation"
      }
    }
  },
  {
    $match: {
      "age": {
        $lt: 25
      },
    }
  }
])
```

DBSCAN

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

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

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



Use the following hyperparameters: ϵ = 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)

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:

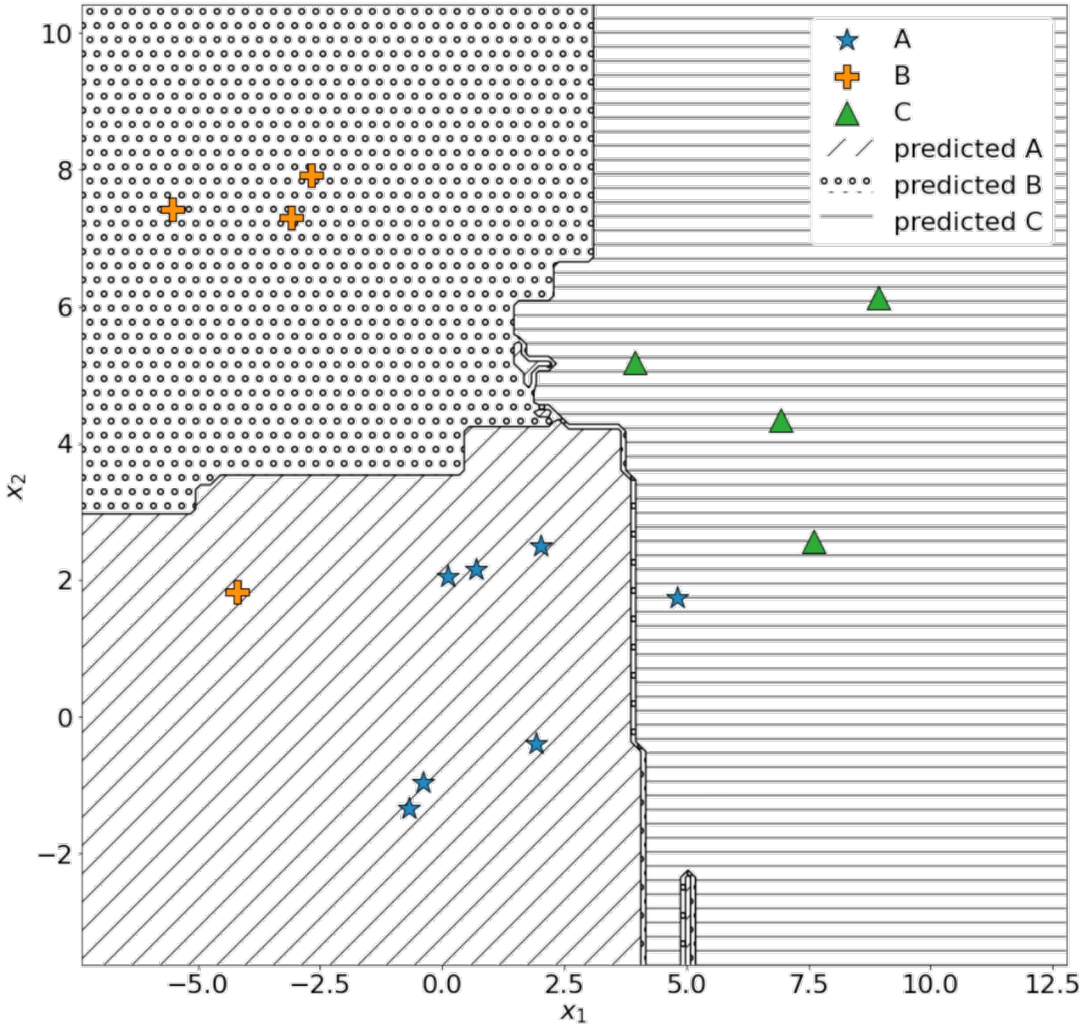
- Diagonal lines represent areas of the input space where the model predicts class A
- Small circles represent areas of the input space where the model predicts class B
- Horizontal lines represent areas of the input space where the model predicts class C

Write in the box below:

precision(B)

precision(C)

recall(B)



K-means

You are given a dataset containing 9 points in 2 dimensions (x1, x2). 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.