Benha University
1th Term (Nov. 2021)

Class: The fourth Year

Subject: Big Data
Course Code: SC 446

Final Exam

نموذج إجابة

Faculty of Computers & AI
Date: 09/01/2022
Time: 3 hours
Total Marks: 65 Marks
Examiner(s): Prof. E. Badr

<u>Answer the following questions [4 questions in 2 page]:</u>

Question No. 1 [20 Marks]

a) Write k-means clustering algorithm and explain it by its flowchart?

b) Cluster the following eight points (with (x, y) representing locations) into three clusters:

A1 (2, 10), A2 (2, 5), A3 (8, 4), A4 (5, 8), A5 (7, 5), A6 (6, 4), A7 (1, 2), A8 (4, 9)

Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2).

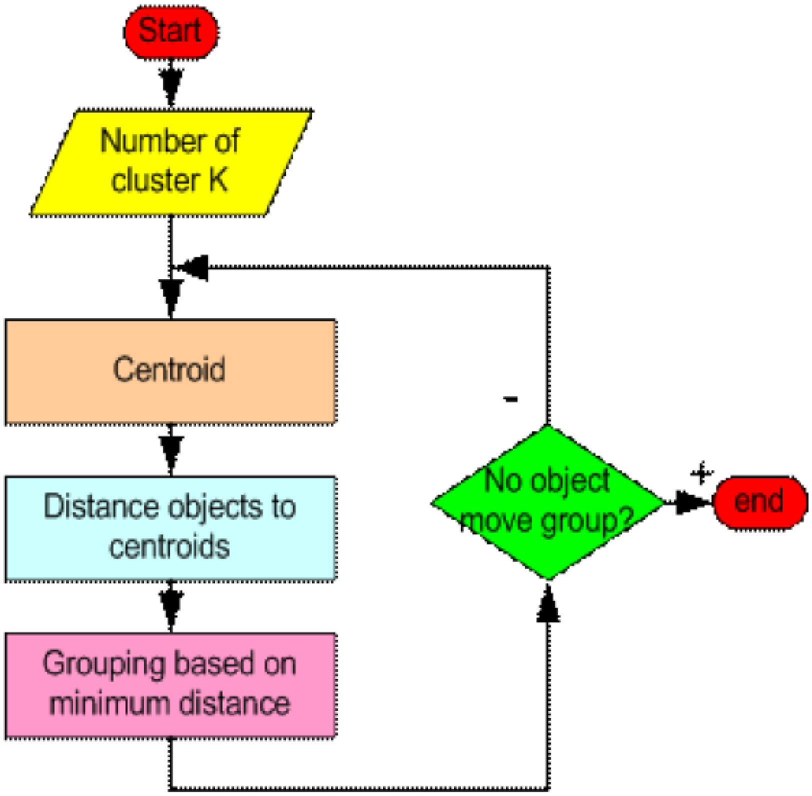
The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as-

$$P(a, b) = |x2 – x1| + |y2 – y1|$$

Use K-Means Algorithm to find the three cluster centers after the second iteration.

Solution:

a)


Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
-
+
end

K-Means Clustering Algorithm-

Step-01:

Choose the number of clusters K.

Step-02:

- Randomly select any K data points as cluster centers.
- Select cluster centers in such a way that they are as farther as possible from each other.

Step-03:

Calculate the distance between each data point and each cluster center.

- The distance may be calculated either by using given distance function or by using euclidean distance formula.

Step-04:

- Assign each data point to some cluster.
- A data point is assigned to that cluster whose center is nearest to that data point.

Step-05:

Re-compute the center of newly formed clusters.

- The center of a cluster is computed by taking mean of all the data points contained in that cluster.

Step-06:

Keep repeating the procedure from Step-03 to Step-05 until any of the following stopping criteria is met-

- Center of newly formed clusters do not change
- Data points remain present in the same cluster

Maximum number of iterations are reached

b)

We follow the above discussed K-Means Clustering Algorithm-

Iteration-01:

- We calculate the distance of each point from each of the center of the three clusters.
- The distance is calculated by using the given distance function.

The following illustration shows the calculation of distance between point A1(2, 10) and each of the center of the three clusters-

Calculating Distance Between A1(2, 10) and C1(2, 10)-

P(A1, C1)

= |x2 – x1| + |y2 – y1|

= |2 – 2| + |10 – 10|

= 0

Calculating Distance Between A1(2, 10) and C2(5, 8)-

P(A1, C2)

= |x2 – x1| + |y2 – y1| = |5 – 2| + |8 – 10| = 3 + 2 = 5

Calculating Distance Between A1(2, 10) and C3(1, 2)-

P(A1, C3)

= |x2 – x1| + |y2 – y1| = |1 – 2| + |2 – 10| = 1 + 8 = 9

In the similar manner, we calculate the distance of other points from each of the center of the three clusters.

Next,

- We draw a table showing all the results.
- Using the table, we decide which point belongs to which cluster.
- The given point belongs to that cluster whose center is nearest to it.

Given Points	**Distance from center (2, 10) of Cluster-01**	**Distance from center (5, 8) of Cluster-02**	**Distance from center (1, 2) of Cluster-03**	**Point belongs to Cluster**
A1(2, 10)	0	5	9	C1
A2(2, 5)	5	6	4	C3
A3(8, 4)	12	7	9	C2
A4(5, 8)	5	0	10	C2
A5(7, 5)	10	5	9	C2
A6(6, 4)	10	5	7	C2
A7(1, 2)	9	10	0	C3
A8(4, 9)	3	2	10	C2

From here, New clusters are-

Cluster-01:

First cluster contains points-

- A1(2, 10)

Cluster-02:

Second cluster contains points-

- A3(8, 4)
- A4(5, 8)
- A5(7, 5)
- A6(6, 4)
- A8(4, 9)

Cluster-03:

Third cluster contains points-

- A2(2, 5)
- A7(1, 2)

Now,

- We re-compute the new cluster clusters.
- The new cluster center is computed by taking mean of all the points contained in that cluster.

For Cluster-01:

- We have only one point A1(2, 10) in Cluster-01.
- So, cluster center remains the same.

For Cluster-02:

Center of Cluster-02

= ((8 + 5 + 7 + 6 + 4)/5, (4 + 8 + 5 + 4 + 9)/5)

= (6, 6)

For Cluster-03:

Center of Cluster-03

= ((2 + 1)/2, (5 + 2)/2)

= (1.5, 3.5)

This is completion of Iteration-01.

Iteration-02:

- We calculate the distance of each point from each of the center of the three clusters.
- The distance is calculated by using the given distance function.

The following illustration shows the calculation of distance between point A1(2, 10) and each of the center of the three clusters-

Calculating Distance Between A1(2, 10) and C1(2, 10)-

P(A1, C1)

= |x2 – x1| + |y2 – y1|

= |2 – 2| + |10 – 10|= 0

Calculating Distance Between A1(2, 10) and C2(6, 6)-

P(A1, C2)

= |x2 – x1| + |y2 – y1| = |6 – 2| + |6 – 10| = 4 + 4= 8

Calculating Distance Between A1(2, 10) and C3(1.5, 3.5)-

P(A1, C3)

= |x2 – x1| + |y2 – y1| = |1.5 – 2| + |3.5 – 10| = 0.5 + 6.5 = 7

In the similar manner, we calculate the distance of other points from each of the center of the three clusters.

Next,

- We draw a table showing all the results.
- Using the table, we decide which point belongs to which cluster.
- The given point belongs to that cluster whose center is nearest to it.

Given Points	**Distance from center (2, 10) of Cluster-01**	**Distance from center (6, 6) of Cluster-02**	**Distance from center (1.5, 3.5) of Cluster-03**	**Point belongs to Cluster**
A1(2, 10)	0	8	7	C1
A2(2, 5)	5	5	2	C3
A3(8, 4)	12	4	7	C2
A4(5, 8)	5	3	8	C2
A5(7, 5)	10	2	7	C2
A6(6, 4)	10	2	5	C2
A7(1, 2)	9	9	2	C3
A8(4, 9)	3	5	8	C1

From here, New clusters are-

Cluster-01:

First cluster contains points-

- A1(2, 10)
- A8(4, 9)

Cluster-02:

Second cluster contains points-

- A3(8, 4)
- A4(5, 8)
- A5(7, 5)

- A6(6, 4)

Cluster-03:
Third cluster contains points-

- A2(2, 5)
- A7(1, 2)

Now,

- We re-compute the new cluster clusters.
- The new cluster center is computed by taking mean of all the points contained in that cluster.

For Cluster-01:
Center of Cluster-01
= ((2 + 4)/2, (10 + 9)/2)
= (3, 9.5)

For Cluster-02:
Center of Cluster-02
= ((8 + 5 + 7 + 6)/4, (4 + 8 + 5 + 4)/4)
= (6.5, 5.25)

For Cluster-03:
Center of Cluster-03
= ((2 + 1)/2, (5 + 2)/2)
= (1.5, 3.5)

This is completion of Iteration-02.

After second iteration, the center of the three clusters are-

- C1(3, 9.5)
- C2(6.5, 5.25)
- C3(1.5, 3.5)

Question 2 [20 Marks]

a) Derivatve two equations (using least square method) that determine the constant A and B for the best fit curve Y = AX + B ?

b) Fit the least square line to the following data and find Y(10)

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

Solution:

a)

Since $d_i^2 = [y_i - (y_i)_{curve}]^2$

Since $(y_i)_{curve} = Ax_i + B$

Then $d_i^2 = [y_i - (Ax_i + B)]^2$

Taking the summation from 1 to n

$$\sum_{i=1}^{n} d_i^{\,2} = \sum_{i=1}^{n} [y_i - (Ax_i + B)]^2$$

This summation is a function of A and B only why?

$$E(A,B) = \sum_{i=1}^{n} [y_i - (Ax_i + B)]^2$$

$$\frac{\partial E(A,B)}{\partial A} = \sum_{i=1}^{n} 2[y_i - (Ax_i + B)]^1(-x_i) = 0$$

$$A\sum_{i=1}^{n} x_i^{\,2} + B\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \text{ ----------------------------(1)}$$

$$\frac{\partial E(A,B)}{\partial B} = \sum_{i=1}^{n} 2[y_i - (Ax_i + B)]^1(-1) = 0$$

$$A\sum_{i=1}^{n} x_i + nB\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\quad (2)$$

b)

x	**y**	**x^2**	**xy**
1	**1**	**1**	**1**
3	**2**	**9**	**6**
4	**4**	**16**	**16**
6	**4**	**36**	**24**
8	**5**	**64**	**40**
9	**7**	**81**	**63**
11	**8**	**121**	**88**
14	**9**	**196**	**126**
$\sum = 56$	$\sum = 40$	$\sum = 524$	$\sum = 364$

From the equation (2) 56A+8B=40--------→1
From the equation (1) 524A + 56B =364---→2
Then B = 6/11 and A =7/11 so y= 7/11 (x) + 6/11 therefore y(10)= 76/11

Question 3 [20 Marks]

Choose the correct answer:

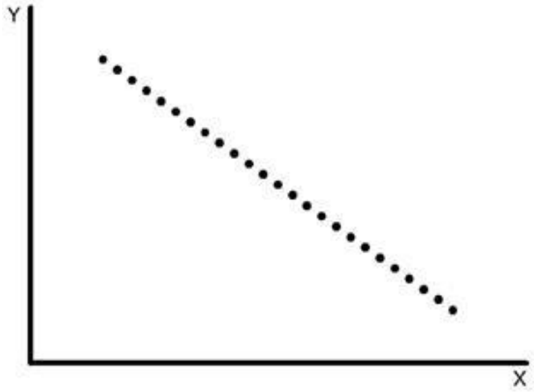
1) True-False: Linear Regression is a supervised machine learning algorithm.
A) TRUE B) FALSE

2) Which of the following methods do we use to find the best fit line for data in Linear Regression?
A) Least Square Error B) Maximum Likelihood C) Logarithmic Loss D) Both A and B

3) Which of the following evaluation metrics can be used to evaluate a model while modeling a continuous output variable?
A) AUC-ROC B) Accuracy C) Logloss D) Mean-Squared-Error

4) Which of the following statement is true about outliers in Linear regression?
A) Linear regression is sensitive to outliers B) Linear regression is not sensitive to outliers
C) Can't say D) None of these

Question Context 5:
Consider the following data where one input(X) and one output(Y) is given.



5) What would be the root mean square training error for this data if you run a Linear Regression model of the form (Y = A0+A1X)?
A) Less than 0 B) Greater than zero C) Equal to 0 D) None of these

Question Context 6:
Suppose, you got a situation where you find that your linear regression model is under fitting the data.
6) In such situation which of the following options would you consider?
1.Add more variables 2.Start introducing polynomial degree variables 3.Remove some variables
A) 1 and 2 B) 2 and 3 C) 1 and 3 D) 1, 2 and 3

7) In practice, Line of best fit or regression line is found when ______________
a) Sum of residuals ($\sum(Y - h(X))$) is minimum
b) Sum of the absolute value of residuals ($\sum|Y-h(X)|$) is maximum
c) Sum of the square of residuals ($\sum (Y-h(X))^2$) is minimum

d) Sum of the square of residuals (∑ (Y-h(X))^2) is maximum

8) If Linear regression model perfectly first i.e., train error is zero, then

a) Test error is also always zero b) Test error is non zero

c) Couldn't comment on Test error d) Test error is equal to Train error

9) How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?

a) 1 b) 2 c) 3 d) 4

10) In a simple linear regression model (One independent variable), If we change the input variable by 1 unit. How much output variable will change?

a) by 1 b) no change c) by intercept d) by its slope

11) In the mathematical Equation of Linear Regression Y = β1 + β2X + ϵ, (β1, β2) refers to …………

a) (X-intercept, Slope) b) (Slope, X-Intercept) c) (Y-Intercept, Slope) d) (slope, Y-Intercept)

Question No. 4 **[5 Marks]**

Formulate the travelling salesman problem as a mathematical model and apply this mathematical model on the complete graph K$_4$?

Solution:

Label the cities as 1, 2, . . ., *n* , in which *n* is the total number of cities and arbitrarily assume 1 as the origin.

Define the decisions variables:

$$x_{ij} = \begin{cases} 1 & \text{if the route includes a direct link between cities } i \text{ and } j, \\ 0 & \text{otherwise.} \end{cases}$$

In addition, for each city i = 1, 2, . . ., n , let $u_i \in \Re^+$

be an auxiliary variable and let c_{ij} be the distance between cities *i* and *j* and . Then, the MTZ formulation to the TSP is the following:

$$\begin{aligned} \min \quad & \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} c_{ij} x_{ij}, \\ \text{subject to} \quad & \\ & \sum_{i=1, i\neq j}^{n} x_{ij} = 1, \quad j = 1, 2, \ldots, n, \\ & \sum_{j=1, j\neq i}^{n} x_{ij} = 1, \quad i = 1, 2, \ldots, n, \\ & u_i - u_j + n x_{ij} \leq n - 1, \quad 2 \leq i \neq j \leq n, \\ & x_{ij} \in \{0, 1\} \quad i, j = 1, 2, \ldots, n, \quad i \neq j, \\ & u_i \in \mathbb{R}^+ \quad i = 1, 2, \ldots, n. \end{aligned}$$

GOOD LUCK
Prof. Dr. E. Badr