CSCI 183
Fall 2024
Quiz 3
11/22/2024
Time Limit: 40 Minutes

Name: Arman Miri
SCU ID: _____

This exam contains 4 pages (including this cover page) and 5 questions.
Total of points is 15.

Grade Table (for teacher use only)

| Question | Points | Score |
|---|---|---|
| 1 | 6 | |
| 2 | 2 | |
| 3 | 2 | |
| 4 | 3 | |
| 5 | 2 | |
| Total: | 15 | |

1. (6 points) Short Answer Questions

(a) What happens if you run K-Means multiple times with different initializations? [Check all that apply]

- ☐ It always produces the same result. won't produce same result
- ☒ It may produce different cluster assignments.
- ☒ It fails to converge in some cases. → if one inital case did then it can reverse!
- ☒ It assigns overlapping clusters. → if clusters drift

(b) What is the minimum no. of features required to perform K-Means clustering? [Write any assumptions]

- ☒ 1
- ☐ 3
- ☐ 0
- ☒ 2

need to have ∝ 1 features minimum
Cannot find pattern on 0, 1 and 3 would work and need data career Junior &?

(c) What is the computational complexity of K-Means per iteration for a dataset with $n$ points, $k$ clusters, and $d$ dimensions?

- ☐ $O(n \cdot d)$
- ☒ $O(k \cdot n \cdot d)$
- ☐ $O(k^2 \cdot n \cdot d)$
- ☒ $O(n^2 \cdot d)$

Ond is k-m
k clusters so add in, it's greedy algo

inertia is squared euclidean distance maybe → ?

(d) (TRUE/FALSE?)In K-Means, the final cluster representative (cluster center) is always an observation from the dataset. Justify your answer.

false, the cluster center would be the centroid which is initialy randomized in good practice, The more iterations however the more it moves so you could consider it an observation but it is derived not seen

Depends on is denoting observation

(e) (TRUE/FALSE?) If a cluster is split by picking one of the points of the cluster as a new centroid and then reassigning the points in the cluster either to the original centroid or the new centroid, the total inertia would increase. Justify your answer.

true, this would happen if the inteia of the points to the new centroid is greater than the previous, or rather the new centroid is overall further from all points

(f) (TRUE/FALSE?) The computational time of KNN and distance weighted KNN are always same for same K. Justify your answer.

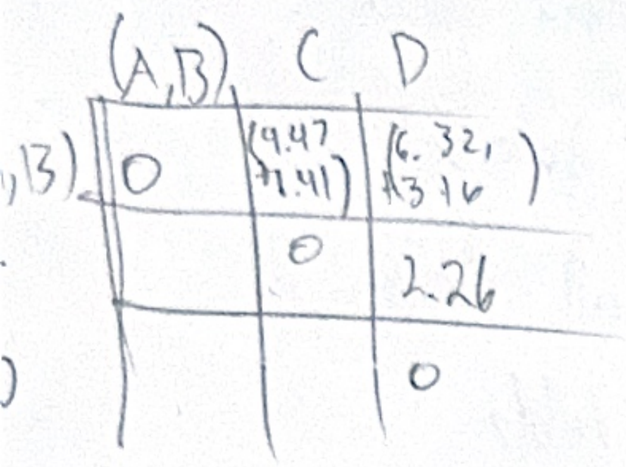
True, both compute the k nearest neighbors, but distanced weighted does it to a query point which will adjust the weight of distance otherwise should be the same False KNN has less comp since weight that distance weighted extra computation

2. (2 points) You are given the proximity matrix of 4 data points (A,B,C,D) in Table 1. Perform agglomerative clustering using the complete-linkage method. What would be the data points in each cluster when the the dataset is divided into two clusters? Show the steps.

complete linkage furthest points

Table 1: Proximity Matrix

| | A | B | C | D |
|---|---|---|---|---|
| A | 0 | 2.83 | 4.47 | 6.32 |
| B | | 0 | 1.41 | 3.16 |
| C | | | 0 | 2.26 |
| D | | | | 0 |




Process →

← Step 2 groups?

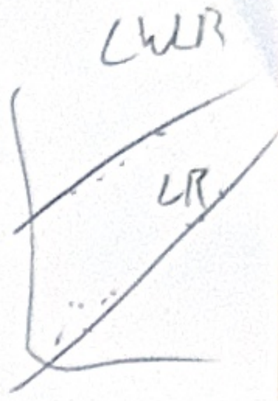
compute dist matrix between pt
let data be own cluster
Repeat
Merge two closest clusters
Update distance Matrix
Until only a sigle cluster remains

3. (2 points) Joey and Chandler are performing 3-Means clustering on the same dataset (same #instances and same #features). They chose the same distance measure and same initial cluster centers. They ended up getting 3 clusters but Joey's clusters had different set of data points as compared to Chandler's. Is this possible? [0.5 point] Justify your answer. [1.5 points]

yes kmeans is unsupervised thus

Ids say they run the algorithm and have some data point equidistant to two clusters. since this is unsupervised the computer decides, thus it's possible it assigns them to diff groups on diff runs. It's very difficult to get identical results on any ML Algo rerunning it. It will be close not-exact. However they should converge on the same point in the end given the given parameters, tested on HW go2 slightly diff results such Pima 98.1 vs 98 vs 9.8.2, etc.

4. (3 points) Joey and Chandler are performing Regression on the same dataset (same #features and same #instances). Joey performs 5-NN and Chandler performs LWLR with the same K value. Joey gets a higher testing error (on the same dataset) than Chandler. Is this possible? Explain your answer.




Yes this is possible, Joey used LWLR which performs based off the nearest points not the whole data set as shown in the chart. Thus when testing his algorithm on the total dataset it resulted in a higher error as it didn't factor all data. While SNN would factor all data even if using 5 diff folds thus allowing it to better predict on the dataset

5. (2 points) You are given the following relations. Write a SQL query to find the names, age and student IDs of students who are enrolled in some course, sorted by age in descending order.

query = '' in
Select *
From Student table left join Enrollments table ← combine
on enrollments table. Studen_id = Students table. Student_id

| student_id | name | age |
|---|---|---|
| 1 | Alice | 20 |
| 2 | Bob | 21 |
| 3 | Charlie | 19 |

Table 2: Students table

| enrollment_id | student_id | course |
|---|---|---|
| 101 | 1 | Mathematics |
| 102 | 1 | Physics |
| 103 | 2 | Chemistry |

Table 3: Enrollments table

Select From Name Where in ''

~~Select * From course where enrollment_id = True Orders by Age;~~ x query course sort by age

Select Name, Age, StudentId Where Course = 'Physics' Orders By Age;