CSCI 183
Fall 2024
Exam 2
12/6/2024
Time Limit: 60 Minutes

Name: Arman Miri
SCU ID: _____

This exam contains 7 pages (including this cover page) and 4 questions.
Total of points is 25.

Grade Table (for teacher use only)

Question	Points	Score
1	9	
2	3	
3	7	
4	6	
Total:	25	

1. (9 points) **Unsupervised Learning**

   (a) (2 points) You are given the following dataset with two features Weight Index and PH:

Medicine	Weight Index	PH
A	1	1
B	2	1
C	4	3
D	5	4

(1,1)
(2,1)

Assume Medicine A and Medicine B are the initial centroids c1 and c2. What are the values of the new c1 and c2 after one iteration of k-means? Show the steps involved.

C → A = $\sqrt{(4-1)^2+(3-1)^2} = \sqrt{3^2+2^2} = \sqrt{13}$

D → A = $\sqrt{(5-1)^2+(4-1)^2} = \sqrt{4^2+3^2} = 5 = \sqrt{25}$

C → B = $\sqrt{(4-2)^2+(3-1)^2} = \sqrt{2^2+2^2} = \sqrt{8}$

D → B = $\sqrt{(5-2)^2+(4-1)^2} = \sqrt{3^2+3^2} = \sqrt{18}$

C → A
D → B
new cluster

new centroids

cluster 1 (A) = $\left(\frac{1+4}{2}, \frac{1+3}{2}\right) = \left(\frac{5}{2}, 2\right)$

cluster 2 (B) = $\left(\frac{2+5}{2}, \frac{1+1}{2}\right) = \left(\frac{7}{2}, \frac{2}{2}\right)$

   (b) (2 points) Monica wants to perform 2-Means algorithm on a customer dataset of her catering business. She decides to choose her initial centroids in the following way:

Choose points with the highest feature values: $(f_1, ?)$ and $(?, f_2)$, where f_1 and f_2 are the two highest values of the features.

Is this a good choice? Demonstrate your answer with an example.

~~Assuming highest feature values means the most similarity, yes.~~ ~~when running k means you want initial centroids with the highest similarity.~~ However, this would not be good. NO, if the clusters don't end up in distinct space. Example if (1,0) and (0,1) are centroids massive overlap slowing algorithm and skewing results

(c) (3 points) Suppose that we have four observations, for which we are given the dissimilarity matrix

$$\begin{bmatrix} & A & B & C & D \\ A & 0 & 0.3 & 0.4 & 0.7 \\ B & 0.3 & 0 & 0.5 & 0.8 \\ C & 0.4 & 0.5 & 0 & 0.45 \\ D & 0.7 & 0.8 & 0.45 & 0 \end{bmatrix}$$

For example, the dissimilarity between the first and the second observations is 0.3. Higher the value, higher the dissimilarity between clusters. Perform agglomerative clustering till you get 1 cluster using single linkage. Justify each merge and show the steps

Merge A,B lowest → next lowest for a,b

(A,B) → A,B,C → Next D → A,B,C,D

.3 .4 .7 → (.3, $\frac{.4+.5}{2}$) → (0, .3, .45, $\frac{.45+.8+.7}{3}$) = (0, .3, .45, .65)

.4 .5 .8

.7 .6 .45

(C,D)

same diff →

A,B → C → A,B,C,D

(.45, .75) → (0, .3, .45, .75)

(d) (2 points) Joey and Chandler are performing agglomerative clustering on a set of data points. They use the same linkage and Euclidean distance metric. They ended up getting different clusters for K = 2. Is this possible? Justify your answer.

No this shouldn't be possible as Agglomerative clustering merges the two closest clusters to each other. If the same data, same distance calc and same linkage they should have the same result. Unless since this is a unsupervised learning Algo they have equidistant clusters to two clusters, in which case it could be possible to have the cluster be assigned to diff ones each time it runs.

2. (3 points) **SQL** You are given the following relations. Write a SQL query for the following:

runs before all

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

Table 1: Students table

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

Table 2: Enrollments table

query "1"
Select *
Slect From Students table
join student_id on
student_id . students table =
student_id enrollments table

if enrollment_id the same and stays the course (in or joined table)

1. Find the total number of courses each student is enrolled in.

Sdect sum(course) From course GroupBy Name;

2. Retrieve the names of students who are not enrolled in any course.

Sdect Name where enroll(ment_id = 'NuLL';

3. Find the average age of students enrolled in courses.

select Physics, From Caurse, AVG (Age);

3. (7 points) **Evaluation**:

(a) (2 points) Given below are images of decision boundaries given by classifiers.

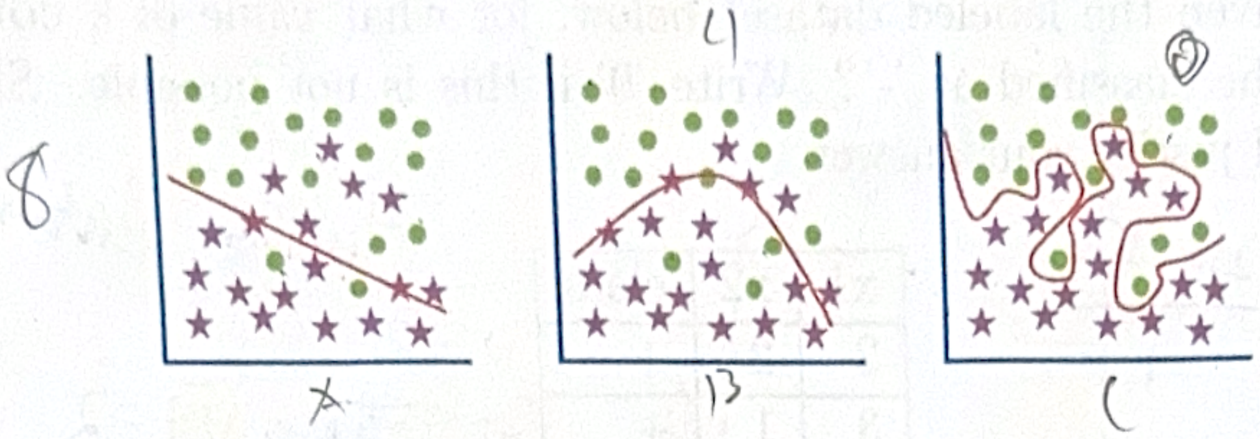


Figure 1: Model A, Model B and Model C from left to right

i. What would be the descending order of training accuracy of the models?

ii. Which model is the best model and why?

i. Highest C, B, A lowest

ii. Model B is the best as it is not overfitting like C and has greater accuracy compared to A, 8 wrg vs 4 wrong

(b) (2 points) Joey has a non-linearly separable, imbalanced dataset with two features. He applied 5-NN and got very high testing error. He used SMOTE hoping to improve the performance but with no luck. Why do you think SMOTE is not working? [Give a reason other than dependency of K]

the model may be thrown off by outliers or distance between points, there is a clear underfitting here. in which case a new model would be best. Distance weighted knn will help. this could be due to the non linearly separable data

(c) (3 points) You are given the following list of test points for a univariate linear regression model: $\hat{y} = 9.2 + 0.8x$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

IN → test attribute value = [43, 44, 45, 46, 47]

true → true value (target) = [41, 45, 49, 47, 44]

out → output = 43.6, 44.4, 45.2, 46, 46.8

Calculate the error of the model. Show the steps.

$\frac{1}{5}\left(|(43.6-41)| + |(44.4-45)| + |(45.2-49)| + |(46-47)| + |(46.8-44)|\right)$

$\frac{1}{5}(2.6 + .6 + 3.8 + 1 + 2)$

$\frac{11}{5} = 2.2$

$\frac{1}{3}(11)$

$\frac{1}{5}(10) = 2$

4. (6 points) **K-NN**

(a) (1 point) Given the labeled dataset below, for what value of k **could** the query point (1,1) be classified as '-'? Write '0' if this is not possible. Show the steps involved and justify your answer.

x1	x2	class
2	2	+
3	1	+
6	2	-
4	1	-
4	2	-
5	3	-
3	3	+

eucidan distance

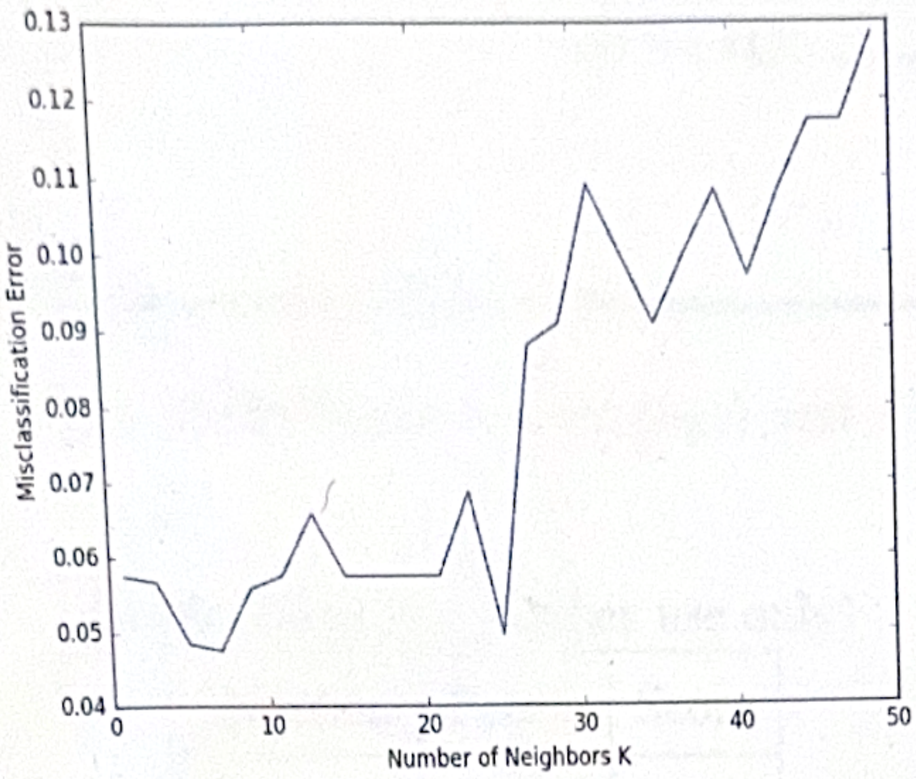
k	class	# + +
1	+	1 +
2	+	2 +
3	+	3 +
4	+	3 + 1 -
5	+	3 + 2 -
6	+	3 + 3 -
7	-	3 + 4 -

→ $\sqrt{2}$
→ $\sqrt{2^2+0^2} \to \sqrt{4} = 2$
→ $\sqrt{5^2+1^2} \to \sqrt{26}$
→ $\sqrt{3^2+0^2} \to \sqrt{9} = 3$
→ $\sqrt{3^2+1^2} \to \sqrt{10}$
→ $\sqrt{4^2+2^2} \to \sqrt{20}$
→ $\sqrt{2^2+2^2} \to \sqrt{8}$

K=7 1,1 will be -

(b) (3 points) **Modified K-NN:** You are given a "black box" where you input a set of instances $P_1, P_2, \ldots P_n$ and a new test example Q, and the black box outputs the nearest neighbor of Q, say P_i and its corresponding class label $C_i, i \in [0, 1]$. Is it possible to construct a 5-NN classification algorithm based on this black box alone? If so, how and if not, why not? Describe in detail.

Assumin Q can give only the 1NN: this won't be possible q can only give the NN to a point to run an algo like 5NN you need the 5 nearest neighbors. this even if ran multiple times will give the same point. If the black box worked and stored the NN and could calcualte a new one on each iteration then you can call it 5 times to get the 5NN needed Black box is doing the distance and then NN for us.



(c) (2 points) Joey records the following plot for K-NN on a dataset.

Joey made the following observations. State whether the following observations are correct/incorrect. Justify your answer:

1. The best range of K should be 30 to 50.
2. The dataset could potentially be imbalanced.
3. The dataset could potentially have a lot of outliers
4. The dataset has two classes.

1. False the best value between 5 to 10 you want the lowest misclassification rate, the higher K higher rate 5 to 10 is lowest

2) True, As the miscalculation rate rises with runs it's possible that it is learning one class due to an imbalance thus higher misclassifications

3) True, the swings in data could be a sign of this as KNN is sensitive to outliers

4) true, otherwise the misclassification rate should be 0 as all points the same. Unless they very enough however the increasing misclassification should show a class is misclassified