CSCI 183 Homework 4

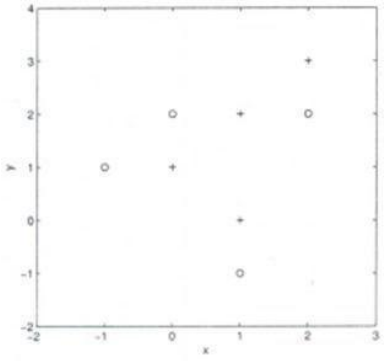
Due Date: March 10th, 2024

Answer the following questions on K-NN:

Q1. Suppose you have given the following data where x and y are the 2 input variables and Class is the dependent variable. (10 points)

x	y	Class
−1	1	−
0	1	+
0	2	−
1	−1	−
1	0	+
1	2	+
2	2	−
2	3	+

Below is a scatter plot which shows the above data in 2D space.



a) **What will be the Euclidean Distance between**

   a. **The two data points A(2,2) and B(2,3)?**

   b. **The two data points C(1,-1) and D(1,0)?**

b) **Suppose you want to predict the class of new data point x=1 and y=1 using Euclidean distance in 3-NN. In which class this data point belongs to and why?**

c) **In the previous question, you are now wanting to use 7-NN instead of 3-NN which of the following x=1 and y=1 will belong to?**

Q2. State True/False for the following statements for k-NN classifiers? Justify your answer. [5 points]

i) The classification accuracy is better with larger values of k
ii) The classification accuracy is best achieved with small values of k
iii) The hypothesis function is the most important aspect of k-NN
iv) k-NN does not require an explicit training step
v) k-NN is a non-parametric method of classification

Answer the following questions on K-Means:

Q1. For which of the following tasks might K-means clustering be a suitable algorithm. Select all that apply and justify your answer! (5 points)

a) Given a set of news articles from many different news websites, find out what are the main topics covered.
b) Given historical weather records, predict if tomorrow's weather will be sunny or rainy.
c) From the user usage patterns on a website, figure out what different groups of users exist.
d) Given a database of information about your users, automatically group them into different market segments.
e) Given sales data from a large number of products in a supermarket, figure out which products tend to form coherent groups (say are frequently purchased together) and thus should be put on the same shelf.
f) Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.

Q2. **Suppose you have an unlabeled dataset $\{x^{(1)}, \ldots, x^{(m)}\}$. You run K-means with 50 different random initializations and obtain 50 different clusters of the data. What is the recommended way for choosing which one of these 50 clusters to use? Explain your answer. (5 points)**

a) Plot the data and the cluster centroids, and pick the clustering that gives the most "coherent" cluster centroids.
b) Manually examine the clusters and pick the best one.
c) The only way to do so is if we also have labels $y^{(i)}$ for our data.
d) For each of the clusters, compute 'Inertia', and pick the one that minimizes the sum of this.

Q3. **Which of the following statements are true? Select all that apply and explain your answer for each choice. (5 points)**

a) On every iteration of K-means, the loss function (inertia) should either stay the same or decrease; in particular, it should not increase.
b) A good way to initialize K-means is to select K (distinct) examples from the training set and set the cluster centroids equal to these selected examples.
c) K-Means will always give the same results regardless of the initialization of the centroids.
d) Once an example has been assigned to a particular centroid, it will never be reassigned to another different centroid
e) For some datasets, the "right" or "correct" value of K (the number of clusters) can be ambiguous, and hard even for a human expert looking carefully at the data to decide.
f) The standard way of initializing K-means is setting $\mu_1 = \ldots = \mu_k$ to be equal to a vector of zeros.

Q4**. Use K-Means Algorithm to create three clusters- [10]. You may choose to code or solve it on paper.**

Point	Coordinates
A1	(2,10)
A2	(2,6)
A3	(11,11)
A4	(6,9)
A5	(6,4)
A6	(1,2)
A7	(5,10)
A8	(4,9)
A9	(10,12)
A10	(7,5)
A11	(9,11)
A12	(4,6)
A13	(3,10)
A14	(3,8)
A15	(6,11)

Assume A2(2, 6), A7(5,10) and A15(6,11) are initialized centers of the clusters. Just show until first iteration.

HW4

KNN

1. (a) $d_{AB} = \sqrt{(2-2)^2 + (2-3)^2} = 1$

$d_{CD} = \sqrt{(1-1)^2 + (1-0)^2} = 1$

x_1	y_1	$\sqrt{(x_1-1)^2+(y_1-1)^2}$
−1	1	2
0	1	1
0	2	1.4142136
1	−1	2
1	0	1
1	2	1
2	2	1.4142136
2	3	2.236068
1	1	0

(b) Using 3-NN, 3 closest points are (1,2), (0,1), (1,0) and all three of these are + so prediction of (1,1) will be **+**

(c) it would be **−** instead, as 7 closest points are all points in dataset except for point B, (4'−' & 3'+').

2.

i) The classification accuracy is better with larger values of k

False, if we have k = # of data points then the majority group always win, which can actually have lower accuracy.

ii) The classification accuracy is best achieved with small values of k

False lower k, (e.g. k=1) can be easily affected by naises which can also lead to low accuracy.

iii) The hypothesis function is the most important aspect of k-NN

False KNN algorith has no hypothesis function, it's non-parimetric.

iv) k-NN does not require an explicit training step

True KNN has no hypothesis function and no optimization is needed, so no training. it basically just use dataset to predict.

v) k-NN is a non-parametric method of classification

False, KNN is non-parametric, but can also be regression when the target is numerical, and it will just take the average of the k point's target value.

K-Means:

1.

a) **Suitable**,
K-means can be used to find clusters of news articles that are similar to each other, which could correspond to the main topics.

b) **Not suitable**,
We don't have any target in our data set, so we can't predict the new data point's target value

c) **Suitable**,
K-means can cluster user patterns to identify different behaviors or groups, which can be used for understanding different user segments. However, we still can't know what kind of group each cluster is.

d) **Suitable**,
similar from c), different users with different behaviors can be clustered into different market segments

e) **Suitable**,
if a group of products are bought in similar patterns, and similar sale on each, then potentially, customers see them as bundled goods (such as garbage can and garbage bag)

f) **Not suitable**,
again, K-mean isn't able to predict future value as it has no target information on its data set

2.
I would choose to use option d)

Option a can be hard if we have higher dimensions for features, such as 4+ features, and I think it would be hard to plot.

Option b can be quite subjective, and can be easily influenced by the people picking the clusters.

Option c is impossible as we are assuming no target labels.

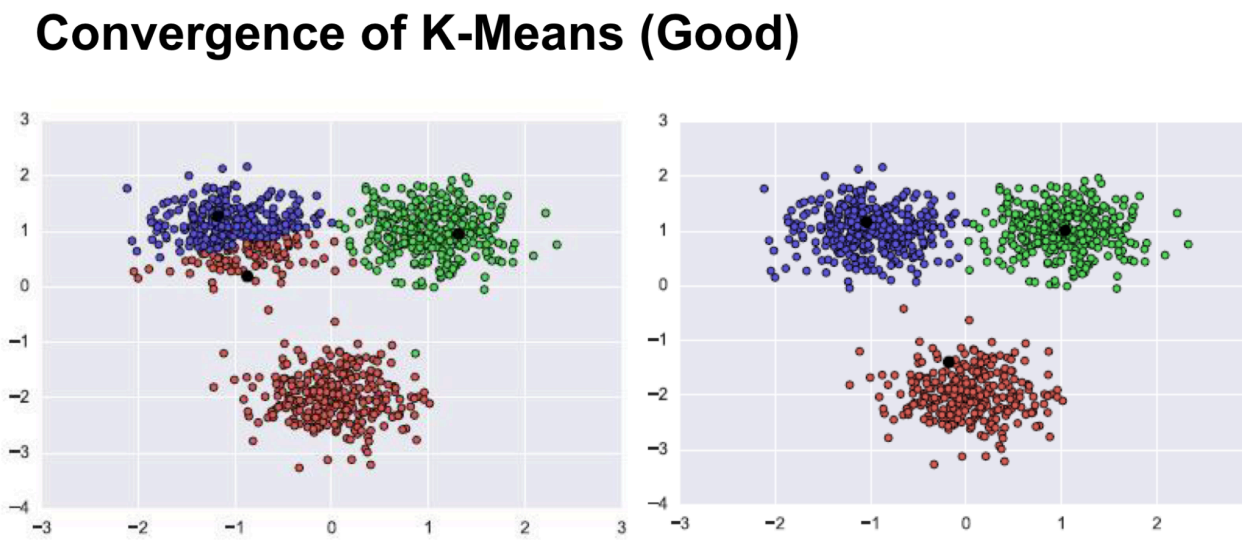
Option d sounds more feasible, as the points are considered more related to each other in each cluster when the inertia is low, and could be a good determinator on how good the centroids are.

3.

a) **True**,
Since for each iteration, we are picking points closer and closer to the centroid, and repicking the centroid from the center of the cluster, we should be able to see a decrease in each iteration, and remaining the same is also possible, but probably more on the ones where it almost converges.

b) **True**,
Picking the data points can be a good way for initialization, and it's also possible that picking random points (not data points) can result in a good clustering result.

c) **False**,
note different initialization can lead to different results. For example:



d) **False**,
Normally at the first few iterations, it's common for points to switch back and forth between clusters.

e) **True**,
if the dataset don't have clear boundaries, and different data points are nested together, it would be really hard to separate them in a definitely correct way,

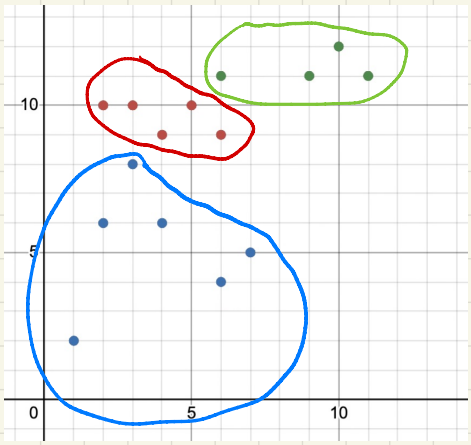
f) **False**,
Firstly, pickling all initialization points to be the same would make it impossible to cluster points, as each point will have the same distance to all centroids.

4.

initialization : $C_1 = A_2 = (2,6)$

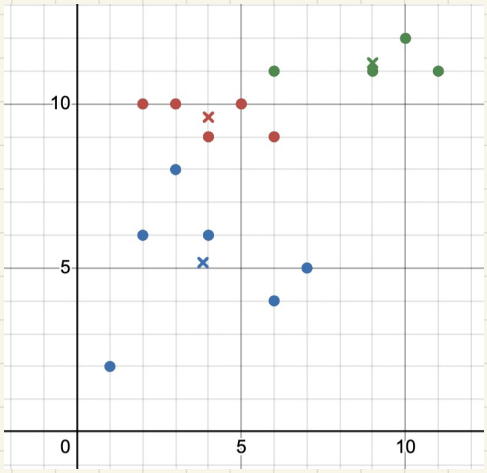
$C_2 = A_7 = (5,10)$

$C_3 = A_{15} = (6,11)$



first iteration :

x_1	y_1	$\sqrt{(x_1-2)^2+(y_1-6)^2}$	$\sqrt{(x_1-5)^2+(y_1-10)^2}$	$\sqrt{(x_1-6)^2+(y_1-11)^2}$
2	10	4	3	4.1231056
2	6	0	5	6.4031242
11	11	10.29563	6.0827625	5
6	9	5	1.4142136	2
6	4	4.472136	6.0827625	7
1	2	4.1231056	8.9442719	10.29563
5	10	5	0	1.4142136
4	9	3.6055513	1.4142136	2.8284271
10	12	10	5.3851648	4.1231056
7	5	5.0990195	5.3851648	6.0827625
9	11	8.6023253	4.1231056	3
4	6	2	4.1231056	5.3851648
3	10	4.1231056	2	3.1622777
3	8	2.236068	2.8284271	4.2426407
6	11	6.4031242	1.4142136	0



new centroid : $C_1^{(1)} = \left(\frac{2+6+1+7+4+3}{6}, \frac{6+4+2+5+6+8}{6} \right)$

$= \left(\frac{23}{6}, \frac{31}{6} \right)$

$C_2^{(1)} = \left(\frac{2+6+5+4+3}{5}, \frac{10+9+10+9+10}{5} \right)$

$= (4, 9.6)$

$C_3^{(1)} = \left(\frac{11+10+9+6}{4}, \frac{11+12+11+11}{4} \right)$

$= (9, 11.25)$

x_1	y_1	$\sqrt{\left(x_1-\frac{23}{6}\right)^2+\left(y_1-\frac{31}{6}\right)^2}$	$\sqrt{\left(x_1-4\right)^2+\left(y_1-9.6\right)^2}$	$\sqrt{\left(x_1-9\right)^2+\left(y_1-11.25\right)^2}$
2	10	5.1693541	2.0396078	7.1107313
2	6	2.013841	4.1182521	8.75
11	11	9.2406109	7.1386273	2.0155644
6	9	4.4032816	2.0880613	3.75
6	4	2.4608038	5.9464275	7.8461774
1	2	4.2491829	8.1706793	12.229575
5	10	4.9721446	1.077033	4.1907637
4	9	3.8369548	0.6	5.482928
10	12	9.2044675	6.4621978	1.25
7	5	3.1710496	5.4918121	6.5622024
9	11	7.7924465	5.192302	0.25
4	6	0.84983659	3.6	7.25
3	10	4.9046463	1.077033	6.1288253
3	8	2.9533409	1.8867962	6.823672
6	11	6.2227182	2.4413111	3.0103986

new cluster: