CSCI 183 Homework 4

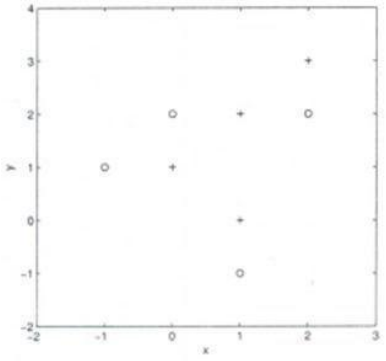
Due Date: March 10th, 2024

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.