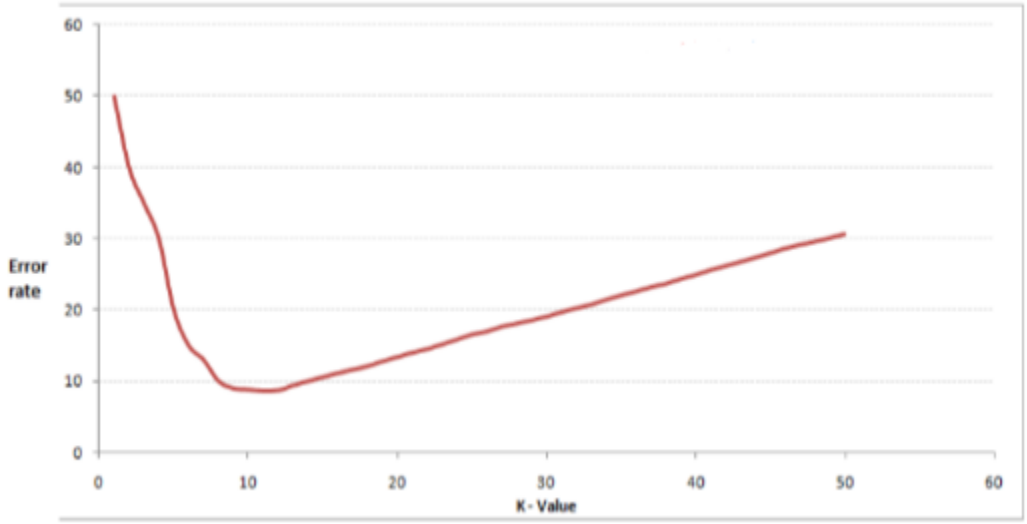
Aditya Ranade and Austin Wang
Ghosh
CSCI183
24 May 2022

CSCI183 - Homework 4

Q1.In the image below, which would be the best value for k assuming that the algorithm you are using is k-Nearest Neighbor. Explain your answer (5 points)

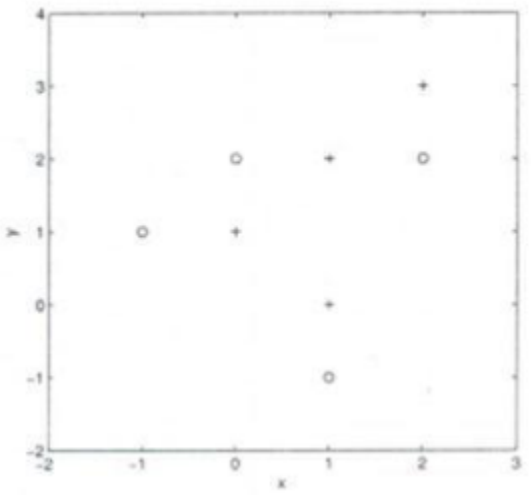


ANSWER: K=10 because the error rate is the lowest there, so using this value of k would make the most sense.

Q2. 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 (i)Euclidean Distance (ii) Manhattan Distance between the two data point A(2,2) and B(2,3)?

Q2a. (i) sqrt((2-2)^2 + (2-3)^2) = 1
Q2b. (i) sqrt((1-1)^2 + (-1-0)^2) = 1

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?

b. If we use 3-NN, x=1 and y=1 will belong to the "+" class, represented by a "+" in the 2D graph of the plotted points. The reason for this is when this test point is plotted, we see that in the 3-NN circle, there are 3 "+" points, and using the majority rule, it is classified as the "+" point.

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?

c. If we use 7-NN instead of 3-NN, x=1 and y=1 will belong to the "-" class, represented by a circle in the 2D graph of the plotted points. The reason for this is when this test point is plotted, we see that in the 7-NN circle, there are 4 "-" points and 3 "+" points, and using the majority rule, it is classified as the "-" point.

Q3. State True/False for the following statements for k-NN classifiers? Justify your answer.

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

False, you need to find a value of k that is neither too small nor too large.

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

False, you need to find a value of k that is neither too small nor too large, and the decision boundary may not represent the best classification

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

False, there is no explicit hypothesis function for k-NN

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

True, it simply uses the training data at the test time to make predictions.

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

True, this is a fact (there are no values of theta that are used in this form of classification).

Q4. Suppose you have trained a k-NN model and now you want to get the prediction on test data. Before getting the prediction suppose you want to calculate the time taken by k-NN for predicting the class for test data.
Note: Calculating the distance between 2 observation will take D time.

a) What would be the time taken by 1-NN if there are N(Very large) observations in test data?

ANSWER: N * D, as the value of N is very large.

b) What would be the relation between the time taken by 1-NN, 2-NN, 3-NN.

N*D for all because the memory is stored at the beginning, meaning that for all of them, the time will be about the same, because any value of k in the k-NN algorithm results in the same amount of time taken.

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!

a) Given a set of news articles from many different news websites, find out what are the main topics covered.

TRUE - This is clustering because you are sorting articles into different groups by topic. Therefore, each of the clusters represents a different segment of topics.

b) **Given historical weather records, predict if tomorrow's weather will be sunny or rainy.**

FALSE - This is not clustering because you are also trying to predict the weather, and there are no labels on the inputs, therefore it becomes impossible to classify.

c) From the user usage patterns on a website, figure out what different groups of users Exist.

TRUE - This is clustering because you are trying to sort users into groups. Therefore, each of the clusters represents a different group/type of users.

d) Given a database of information about your users, automatically group them into different market segments.

TRUE - This is clustering because we are again trying to sort users into groups, and each cluster will represent a different market segment

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.**

TRUE - This is clustering because you are sorting products into groups based on how commonly they are purchased together, and this becomes a common example of segmenting the products into different groups.

f) Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.

FALSE - This is not clustering because you are trying to predict future sales. This would be considered more of a regression problem, and K-means does not label the data in the input, therefore it becomes impossible to perform regression.

Q2.Suppose you have an unlabeled dataset. 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.

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 for our data.
d) For each of the clusters, compute 'Inertia', and pick the one that minimizes the sum of

This.

ANSWER: D - We should compute the inertia and pick the clusters that minimize the sum of it because lower Inertia means the cluster has data points that are closer to the central point. Therefore, the lower the Inertia the better the clusters are.

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

a) On every iteration of K-means, the loss function (inertia) should either stay the same or decrease; in particular, it should not increase.

TRUE - the loss function should always be decreasing as we try to minimize the loss and improve the clusterings. If the inertia is decreasing, that means we are getting better clusterings.

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.

TRUE - this is the recommended method that should be used for the initialization of the K-means algorithm.

c) K-Means will always give the same results regardless of the initialization of the centroids.

FALSE - the way optimal centroids are calculated is affected by the starting point because that will determine around where clusters are made. K-means should be done multiple times from different starting points as the best location of the centroids may change as this is done.

d) Once an example has been assigned to a particular centroid, it will never be reassigned to another different centroid

FALSE - as the cluster center is re-calculated and is changed for the next iteration of the K-means algorithm, the points/examples that are next to the new centroid may be re-classified t another cluster. The K-means algorithm is executed until the centroids no longer change, and the cluster assignments no longer change. Until this is achieved, an example may continue to change which centroid it is being assigned to.

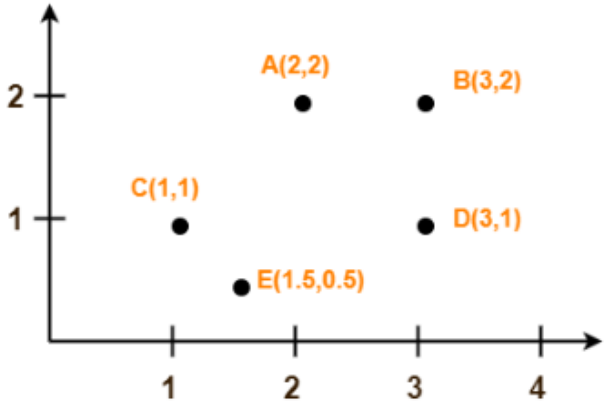
e) For some datasets, the "right" or "correct" value of K (the number of clusters) can be ambiguous, and hard even for a human expert looking carefully at the data to decide.

TRUE - finding the right K value is hard because there isn't an exact loss or error value that is optimal. Instead, there can be a range of K values that are good but not always perfect. If we use the elbow method for example, what value of k to choose can sometimes be hard as there may not be an exact point on the curve that highlights which value of k will give us the best number of clusters for the given dataset.

f) The standard way of initializing K-means is setting to be equal to a vector of zeros.

FALSE - we want to set it up so that each point is near at least one center of a cluster. This will help us better determine the groupings and find points that belong in each cluster.

Q4. Use K-Means Algorithm to create two clusters- [10]
Assume A(2, 2) and C(1, 1) are initialized centers of the clusters. Just show until Iteration-1.



A-B: d = sqrt((2-3)^2 + (2-2)^2) = 1
C-B: d = sqrt((1-3)^2 + (1-2)^2) = 2.236

A-D: d = sqrt((2-3)^2 + (2-1)^2) = 1.414
C-D: d = sqrt((1-3)^2 + (1-1)^2) = 2

A-E: d = sqrt((2-1.5)^2 + (2-0.5)^2) = 1.581
C-E: d = sqrt((1-1.5)^2 + (1-0.5)^2) = 0.707

A-C: d = sqrt((2-1)^2 + (2-1)^2) = 1.414

Clusters based on the distances calculated: {A, B, D} , {C, E}

<u>New centers</u>
A cluster center ((2 + 3 + 3)/3, (2 + 2+ 1)/3) = (2.667, 1.667)
C cluster center ((1 + 1.5)/2, (1 + 0.5)/2) = (1.25, .75)