Jack Yeung
3/31/24
B455

Assignment05

1 Question 1

Suppose we have the following data points in a 2-dimensional space:
A = (1, 1)
B = (2, 2)
C = (2, 4)
D = (1, 2)
Perform single-link clustering using Euclidean distance and answer the following:
a) Describe step by step process of generating clusters using single-link clustering for the given data points. (Show your step-by-step calculations)
calculate the distance for all variations

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

$$d_{ac} = \sqrt{(2-1)^2 + (4-1)^2} = \sqrt{10}$$

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

$$d_{bc} = \sqrt{(2-1)^2 + (4-2)^2} = 2$$

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

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

merge ad and bd, which only leaves C

b) Draw the dendogram for final clusters.

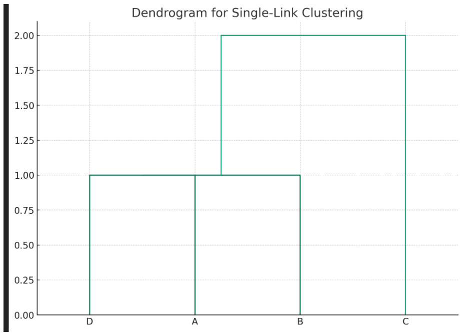

Figure 1: dendrogram

2 Question 2

Based on the K-Means clustering, answer the following question.
a) Why is it crucial to carefully select the number of clusters in a clustering task?

When selecting the right amount of clusters for the k-means algorithm it is important to consider what the goal of the clustering is. The goal of k-means and other clustering algorithms is to find groups within a dataset. If the amount of clusters is too high you may be splitting a group into a smaller unimportant groups. Conversely if there are not enough clusters you may lose information on sub populations. By getting the correct number of clusters, you aim to find unique groups which provide some sort of value or information to the target column.

b) Discuss the implications of choosing an inappropriate value for k.

As covered in the previous question, if you have an inappropriate amount of k you will struggle to make sense of the data. The groups will not provide any insight that can be used to train a model.

3 Question 3

Consider a dataset consisting of the following observations: Data Points: [2,3,5,7,9] Using the k-nearest neighbor (k-nn) density estimator, calculate the estimated density for the following data

k-nn formula:

$$\hat{p} = k/2Nd_k$$

a) x = 4 with k = 2

find distances from x to each point: [2,1,1,3,5]
sorted: [1,1,2,3,5]
distance to kth point = 1
insert into formula: 2/(2 * 5 * 1) = .2

b) x = 5 with k = 3

find distances from x to each point: [3,2,0,2,4]
sorted: [0,2,2,3,4]
distance to kth point = 2
insert into formula: 3/(2 * 5 * 2) = .15

Show your step-by-step calculations, including the determination of the distances to the nearest neighbors and the final density estimate.

4 Question 4

Answer the following questions
a) What are outliers ?

an outlier is an instance of a dataset which is drastically different from other points.

b) Explain the significance of outlier detection.

Because outliers will skew our models, it is important to locate and deal with them. The goal is to statistically find which data points do not belong into the data set. In the case of a Gaussian distribution we eliminate data points which statistically should not show up. A common way to do this would be to find how many standard deviations the point is away from the mean. Or in the case of a multivariate distribution, we would find outlier points with large mahlanobis distances.

c) Explain the concept of local outlier factor (LOF) and its role in outlier detection.

local outlier factor is a useful tool to locate outliers in non parametric distributions. the score of LOF is calculated by comparing the density of neighborhoods to the average density. This is done by finding the distance from a point to it's kth nearest neighbor proportionate to the amount of points in the neighborhood. Larger values indicate a higher chance of being an outlier.

5 Question 5

a) Describe the process of nonparametric density estimation using histograms. What are the advantages and disadvantages of this method?

Using non parametric densities, you can estimate the probability of a data point without assuming an underlying structure to the distribution of data points. This allows this model to be much more flexible to datasets that do not follow a distribution model like Gaussian. Although it is simplistic, the model is sensitive to different bin widths as well as discontinuity.

b) Describe the kernel estimator approach for nonparametric classification. How does it differ from the k-nearest neighbor approach?

A kernel estimator approach allows for an estimation for a smooth, continuous probability density function. This approach differs from k nearest neighbors as kernel smoothing is used for overall densities whereas knn models focus on local densities of neighborhoods not over the whole space.

6 Question 6

a) Consider the following dataset representing the values of a variable: [1.2,2.4,2.5,3.1,3.5,4.2,4.8,5.3,5.5,6.1]. Using the histogram method, estimate the probability density function p(x) for a given dataset. Assume a bin size h of 0.5 and calculate the density estimate for the query point x=3.

Using the histogram method

$$\hat{p}(x) = \frac{\#\{x^t \text{ in the same bin as } x\}}{Nh}$$

assuming bins that start at 0 and range every .5, the point would fall into the bin of [2.5,3) which contains 1 other point. Inserting into the formula we get

$$\hat{p} = 1/(10 * .5)$$

$$\hat{p} = .2$$

b) Explain the differences between using a Gaussian kernel and an ellipsoidal kernel in multi-variate density estimation.

Gaussian vs ellipsoidal kernels in multi-variate describe the shape of the kernel smoothing process. Gaussian kernels will perform equally in each direction. A ellipsoidal kernel will have smooth each axis differently, depending on the distribution of each axis.

c) Explain Condensed Nearest Neighbor algorithm

condensed nearest neighbor is an algorithm that aims to reduce the computational resources needed for the knn algorithm. The goal of this algorithm is to remove data points that do not border other any other class. I.e datapoints that are surronded by the same class, and do not contribute to the discriminant function.