Assignment 05

Instructions

1. Each assignment can contain both theoretical and practical questions.
2. Use LaTeX (preferred) or Word for theoretical question responses.
3. Practical questions are in the provided Jupyter notebook. Use Google Colab (Preferred) or Jupyter Notebook to complete questions directly in the Jupyter Notebook. Include code changes and reasoning in the Jupyter Notebook. Convert the Jupyter Notebook into an HTML page for submission.
4. Submit a PDF or Word file with reponses to theoretical questions, a Jupyter Notebook, and an HTML page (both files) with completed practical questions.
5. A 25% penalty applies to submissions on the first day after the due date, and a 50% penalty for submissions 24 to 48 hours late. No submissions will be accepted beyond 48 hours past the due date.

Theoretical Questions

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)

b) Draw the dendogram for final clusters.

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?

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

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

a) x = 4 with k = 2

b) x = 5 with k = 3

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

Question 4

Answer the following questions

a) What are outliers ?

b) Explain the significance of outlier detection.

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

Question 5

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

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

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.

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

c) Explain Condensed Nearest Neighbor algorithm.

Practical Questions

Please refer to and answer Questions 7, 8, and 9 in the provided Jupyter Notebook