Iran University of Science and Technology

School of Computer Engineering

Assignment #3

ADVANCED DATA MINING, SPRING 2025

DR. MINAEI

Hossein Fanaei, Nafiseh Ahmadi

Problem 1)

1. Explain how LLMs are utilized for 6 different classification tasks (only name the task, then elaborate on its demanding aspects). Recommend a solution for each limitation you mention.
2. Suppose that we would like to select between two prediction models, M1 and M2. We have performed 10 rounds of 10-fold cross validation on each model, where the same data partitioning in round i is used for both M1 and M2. The error rates obtained for M1 are 30.5, 32.2, 20.7, 20.6, 31.0, 41.0, 27.7, 26.0, 21.5, 26.0. The error rates for M2 are 22.4, 14.5, 22.4, 19.6, 20.7, 20.4, 22.1, 19.4, 16.2, 35.0. Comment on whether one model is significantly better than the other considering a significance level of 1%.
3. Many partitional clustering algorithms that automatically determine the number of clusters claim that this is an advantage. List two situations in which this is not the case.

Problem 2)

Use the similarity matrix in the table below to perform single and complete link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged.

	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

Problem 3)

Using the given dissimilarity matrix, compute the silhouette coefficient for each point, each of the two clusters, and the overall clustering. Cluster 1 contains {P1, P2}, Cluster 2 contains {P3, P4}.

	P1	P2	P3	P4
P1	0	0.10	0.65	0.55
P2	0.10	0	0.70	0.60
P3	0.65	0.70	0	0.30
P4	0.55	0.60	0.30	0

Problem 4)

In this question, you will use several classification algorithms on the real Titanic dataset, which contains information about the passengers aboard the Titanic. Your goal is to train classifiers that can predict whether a passenger survived based on their features. To do this, you must complete the Classification_Titanic.ipynb notebook by implementing the sections marked with TODO comments. Finally, compare and analyze the performance of the different classifiers.

Problem 5)

Complete the clustering.ipynb notebook and answering questions provided throughout the exercise.