Data Mining and Machine Learning

Mid-Semester Examination, II Semester, 2023–2024

Date : 2 March, 2024
Duration : 2 hours

Marks : 30
Weightage : 20%

1. In the market-basket analysis problem, suppose the set of items I has size 10^7, the number of transactions T is 10^{10} and each transaction $t \in T$ contains at most 10 distinct items. Compute upper bounds for F_1 and F_2, the number of frequent itemset of size 1 and 2, respectively, for a support value of 0.1%. *(5 marks)*

2. Recall that a class association rule has the class attribute as its target. To reduce overfitting, a class association rule can be generalized by dropping attributes from its left hand side and checking if the performance improves over random test data.

   Given a decision tree, explain how to interpret paths in the tree as class association rules. How can we apply the generalization strategy for association rules to generalize decision trees? In what way would this be different from generalization through the usual method of pruning? *(5 marks)*

3. Your team has computed the solution to a linear regression problem on n attributes as $\theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n$. Your partner argues that the relative importance of the attributes can be computed from the coefficients. The most significant attribute is the one with the largest coefficient (in magnitude), the second most significant attribute is the one with the second largest coefficient, and so on. Explain whether your partner's claim is justified. *(5 marks)*

4. How can we use a random forest classifier to rank input features in order of importance? Why is this calculation more effective for a random forest than for a single decision tree? (5 marks)

5. We have a dataset $X = \{x_1, x_2, \ldots, x_N\}$ equipped with a symmetric distance function: $d(x_i, x_j) = d(x_j, x_i)$ is the distance between x_i and x_j. We construct an $N \times N$ matrix D such that $D[i,j] = d(x_i, x_j)$. We can cluster the N columns of D using the usual Euclidean distance in N dimensions, since each column is a vector of length N. Explain whether the clusters formed by the columns of D have any meaningful interpretation with respect to the original set X. *(5 marks)*

6. Explain how locally linear embeddings are computed. *(5 marks)*