Data Mining and Machine Learning

Mid-Semester Examination, II Semester, 2024–2025

Date : 6 March, 2025
Duration : 3 hours

Marks : 30
Weightage : 20%

1. A number of new vaccines are being deployed to treat a recently discovered disease. Reports are emerging of patients having side effects caused by vaccinations. Some side effects are vaccine-specific, some occur across vaccines.

   For each reported case, there is information available about the nature of the side effect, the vaccine used, demographic details about the patient (age, gender, race, ...) as well as information about prevailing health conditions of the patient (diabetes, hypertension, ...) that may create complications.

   Explain how market-basket analysis can help doctors determine risk factors associated with vaccinations, in general, and specific vaccines, in particular. *(4 marks)*

2. Explain why precision and recall are difficult to achieve simultaneously in a classifier. Describe an example where high precision is preferrable to high recall and another example where the converse is true. *(4 marks)*

3. Suppose we are building a naïve Bayesian classifier and some attribute values are missing in the training data. What problem can this cause with prediction and how can we mitigate the situation? *(3 marks)*

4. The algorithm we described to build a decision tree is deterministic. However, we saw that the decision tree library implemented in Python's `sklearn` library uses a random seed. Why should a random seed be needed? *(3 marks)*

5. (a) Explain whether the following statements are true.

   (i) Polynomial regression can always achieve 100% accuracy with respect to the training data.

   (ii) A decision tree can always achieve 100% accuracy with respect to the training data.

   (b) Explain whether 100% accuracy with respect to the training data is a desirable target to achieve.

   (4 marks)

6. Explain why cross entropy is a more suitable loss function than squared error for logistic regression. (4 marks)

7. 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? (4 marks)

What is a validation set? Explain how a validation set can be used to determine the optimum number of models to use in boosting. (4 marks)