Data Mining and Machine Learning

Mid-Semester Examination, II Semester, 2021–2022

Date	: 16 March, 2022	Marks	: 30
Duration	: 2 hours + 0.5 hours upload time	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. *(5 marks)*

2. Consider the following situation when building a decision tree for binary classification. We have a node with 60 samples, with 40 belonging to the majority class and 20 to the minority class. We have only one attribute A available to query, which splits the node into two subsets S_1 and S_2. S_1 has 35 samples with 21 in the majority class and S_2 has 25 samples with 19 in the majority class.

   Compute the impurity gain using misclassification rate as a measure of impurity and contrast it with the impurity gain due to a nonlinear impurity measure such as Gini index or entropy. What can you conclude from this? *(5 marks)*

3. Explain why squared error is a natural loss function for normal regression while cross entropy is more suitable for logistic regression. *(5 marks)*

4. We want to build a naïve Bayes classifier for junk email. Each message is modelled as a bag of words. However, an email message has some structure that can be exploited: we can separate out the sender's address, the subject line and the body of the message. We assume all three parts are constucted from a common vocabulary, but with different probability distributions. The corresponding generative model first generates the sender's address with some probability distribution, then the subject line, with a different distribution, and finally the body, with yet another distribution. When classifying an email as junk, we would like to give weightage w_1 to the sender's address, w_2 to the subject line and w_3 to the body, $w_1 + w_2 + w_3 = 1$. Explain how to modify the standard naïve Bayes classifier to achieve this. *(5 marks)*

5. How can we use a decision tree to rank input features in order of importance? Compare the effectiveness of this calculation if we use a random forest rather than a single decision tree. *(5 marks)*

6. Suppose we apply gradient boosting to solve a regression problem, using a sequence of regression trees. Describe a strategy to estimate the optimum number of regression trees to use. *(5 marks)*