Data Mining and Machine Learning

Sample Midsemester Exam Questions, II Semester, 2020–2021

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%.

2. Association rules can be used for classification by using the class attribute as the target for each rule. To reduce overfitting, a rule can be generalized by dropping attributes from the left hand side of rules and checking if the performance improves over random test data. Explain how to use this idea for generalizing decision trees. In what way would it be different from generalization through the usual method of pruning?

3. (a) A new test for tuberculosis is administered to a sample of 1000 patients, 100 of whom actually have tuberculosis. The test is found to be 80% accurate—if a person has the disease, the test is positive 80% of the time and if a person does not have the disease, the test is negative 80% of the time.

   Suppose we regard the test as a classifier for tuberculosis. What is its precision and recall?

   (b) An airport security system consists of a full body scanner followed by manual frisking. If the full body scanner beeps, the passenger is checked manually and then allowed to proceed if there is nothing amiss. If the full body scanner does not beep, no frisking is done. In terms of the entries in the confusion matrix, what ratio should the full body scanner maximize to ensure that no suspicious person is let through unchecked? Explain your answer.

4. Suppose we build a decision tree for binary classification on a given set of training data, without any pruning, and we discover a leaf node that is not pure—it has representatives of both classes.

   (a) What can we infer us about the attributes being used for classification?

   (b) Describe a real-life situation where this could happen.

5. 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?

6. When we use bagging, we need not keep aside a separate test set to validate our model. Explain.