CMSC 828G Homework 4

Data Mining,CMSC 828G, Spring 2002

Due Date: Tuesday, May 9, at the start of class

1. A new CS graduate student needs to decide about which classes to take in order to satisfy the Ph.D. class requirements. She asked several students in the department and collected the following data. For example, student 1 took cmsc828g, cmsc656, and cmsc710. Assume that the student thinks it is a good idea to take course combinations that other students have taken. You are required to help her to make a decision as follows.

   (a) Apply the Apriori association rule mining algorithm on the collected data to learn association rules. Show each step. Assume minimum support is 30%.

   (b) Apply the FP association rule mining algorithm on the collected data to learn association rules. Show each step. Assume minimum support is 30%.

   (c) Assume, on the other hand, that the student is trying to maximize the likelihood of getting into the classes, so taking class combinations that are less popular is better. Design an alternative algorithm that can be used. Write down the steps of your algorithm and use the collected data to show how the algorithm works. Can you use an Apriori style approach? Does the monotonicity property hold for infrequent items?

Student ID	cmsc828g	cmsc656	cmsc710	cmsc726	cmsc754
1	1	1	1	0	0
2	1	1	1	1	1
3	1	0	1	1	0
4	1	0	1	1	1
5	1	1	1	1	0
6	1	0	1	0	1
7	1	0	1	1	1
8	1	0	1	1	0
9	1	1	1	1	0
10	1	0	0	1	0

2. The department of Computer Science is thinking of automating the procedure of assigning new teaching assistants to classes. The following data was collected from previous semesters. The class ID is one of the data attributes. The attribute Native Speaker refers to whether or not the TA is native English speaker. The goal is to predict the performance of the TA as being good or bad.

   (a) Using the information gain measurement, discussed in class, construct a decision tree classifier. Show each step.

   (b) Using the gain ratio measurement, discussed in class, construct a decision tree classifier. Show each step.

class ID	Instructor	Semester	Native Speaker	Class Size	Performance
cs134	John	Fall	Yes	Large	Bad
cs145	Smith	Fall	Yes	Small	Bad
cs166	John	Spring	No	Medium	Good
cs270	Rachel	Spring	Yes	Large	Good
cs211	Smith	Summer	No	Medium	Bad
cs240	John	Summer	No	Large	Good
cs256	Rachel	Fall	Yes	Small	Good
cs270	Smith	Spring	Yes	Medium	Bad
cs256	Rachel	Fall	No	Small	Bad
cs270	Rachel	Spring	Yes	Medium	Good

3. Suppose that, as usual, our data set consists of a set of n vectors $X = \{x(1), \ldots, x(n)\}$, each with p attributes or features, $x_1(i), \ldots, x_p(i)$. Suppose there is also a distinguised attribute, $x_c(i)$, the class of $x(i)$, that we are interested in predicting.

   It is often the case that many of these attributes have relatively little value for discriminating between different classes. For example, we might have 100 features, but we require only 5 or 6 to do a good job at classifying x.

   (a) Suppose we build a classifier using all 100 features. Including all those extra features has a number of disadvantages. Discuss them.

   (b) Suppose we are considering eliminating one of the features. Write a formula to express the information lost by eliminating a feature.

   (c) Suppose we are learning a naive bayes classifier. Describe a procedure for selecting the best subset of features of size k. Measure best in terms of the classification error of your classifier. (be precise, i.e., there should be an formula.)

   (d) Is your approach above efficient? If so, give the running time and explain why this is reasonable. If not, describe a possible heuristic algorithm for feature selection.

4. In this problem, you are asked to explore several classification algorithms using the WEKA machine learning toolbox. Please save enough time to do this part of the homework, some of the algorithms require some time to run. The steps involved are as follows.

   (a) Download the WEKA machine learning software. The software is available at http://www.cs.waikato.ac.nz/ml/weka/

   (b) Begin by exploring the use of the tool on the mini_census data set. It is available on the class web page.

   (c) Build a table to compare the following algorithms with respect to their accuracy (correctly classified instances) using 5-fold, 10-fold and 20-fold cross-validation. You do not need to perform the split manually; the software includes an option for doing cross-validation.

   - K-nearest neighbors algorithm (IBK with no cross validation, window size = 0, no distance weighting) with K set to 1, 20 and 50.
   - Naive Bayesian algorithm (NaiveBayes with no kernel estimator).
   - Support Vector Machine algorithm (SMO) with the default settings. See if you can get a better model by varying the values of the upper bound on the coefficients (c) and the degree of the polynomial (exponent) (higher-order polynomials can be very slow; you are only required to use linear kernels for the given census test data; for your own data set, if it is small, you may want to try higher-order polynomials). Specify the settings of the best model you were able to get. For SVMs, you need only run 5-fold cross validation.
   - C4.5 decision tree algorithm (J48.J48). See if you can get a better tree (readable and with high accuracy) by varying the settings of the algorithm. For example, increase and decrease confidence factor, do or do not use reduced error pruning, do or do not use binary splits, vary the minimum number of instances in a leaf, and so on. Specify the settings of the best tree you were able to get.

   (d) Comment on the training time and the classification time of the classifiers.

   (e) Repeat (c) and (d) for another data set of your choice. Please do not use a data set that is included in the tool. Also, the data set should be big enough, i.e. not less than 2000 instances.