CIS400/600 Fundamentals of Data and Knowledge Mining

Homework #2

Spring, 2017

Problem 1 (15 pts): Classification error rate is the impurity measure defined as ClassificationError(t) = $1 - \max_i [p(i|t)]$ where *p(i|t)* denotes the fraction of records belonging to class *i* at a given node *t*. Consider the following set of training examples.

X	*Y*	*Z*	No. of Class C1	No. of Class C2
0	0	0	5	40
0	0	1	0	15
0	1	0	10	5
0	1	1	45	0
1	0	0	10	5
1	0	1	25	0
1	1	0	5	20
1	1	1	0	15

(a) Compute a two-level decision tree using the greedy approach. Use the classification error rate as the criterion for splitting. What is the overall error rate of the induced tree?

(b) Repeat part (a) using *X* as the first splitting attribute and then choose the best remaining attribute for splitting at each of the two successor nodes. What is the error rate of the induced tree?

(c) Compare the results of parts (a) and (b). Comment on the suitability of the greedy heuristic used for splitting attribute selection.

Problem 2 (7 pts): While the .632 bootstrap approach is useful for obtaining a reliable estimate of model accuracy, it has a known limitation. Consider a two-class problem, where there are equal number of positive and negative examples in the data. Suppose the class labels for the examples are generated randomly. The classifier used is an unpruned decision tree (i.e., a perfect memorizer).

(a) Show why a bootstrap sample contains approximately 63.2% of the data points in the original dataset of the same size N.

(b) Determine the accuracy of the classifier using the holdout method, where two-thirds of the data are used for training and the remaining one-third are used for testing.

(c) Determine the accuracy of the classifier using ten-fold cross-validation.

(d) Determine the accuracy of the classifier using the .632 bootstrap method.

Problem 3 (8 pts): Consider the data set shown in the table below.

Record	*A*	*B*	*C*	Class
1	0	0	0	+
2	0	0	1	−
3	0	1	1	−
4	0	1	1	−
5	0	0	1	+
6	1	0	1	+
7	1	0	1	−
8	1	0	1	−
9	1	1	1	+
10	1	0	1	+

(a) Estimate the conditional probabilities for $P(A|+)$, $P(B|+)$, $P(C|+)$, $P(A|-)$, $P(B|-)$, and $P(C|-)$.

(b) Use the estimate of conditional probabilities given in the previous question to predict the class label for a test sample ($A = 0, B = 1, C = 0$) using the naïve Bayes approach.

(c) Estimate the conditional probabilities using the m-estimate approach, with $p = 1/2$ and $m = 4$.

(d) Repeat part (b) using the conditional probabilities given in part (c).

Problem 4 (20 pts): Predict survival on Titanic for Kaggle competition.

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

For this problem, you are expected to apply decision tree learning algorithm using either Weka workbench or RWeka package to analyze and predict what kinds of people were more likely to survive.

- Follow the instruction on the Kaggle instruction (https://www.kaggle.com/c/titanic) to download the train and test data sets and read Titanic data description document.
- Use the train set to build decision tree models, experiment with data preprocessing techniques (e.g., deleted, transformed, or added) and the decision tree model parameters setting (e.g. prune or not, minimal leaf size, etc). Use 10-fold cross validation to evaluate the model performance
- Use the models to predict who survived or not in the test set and upload your prediction to the Kaggle site (follow Kaggle submission instruction on https://www.kaggle.com/c/titanic/details/submission-instructions)

Please follow the below instructions for your submission:

(a) Compare the performance (evaluated by both 10 fold cross validation and Kaggle test set accuracy) of at least FOUR decision tree models using different data pre-processing steps and J48 parameters. Organize and present the model comparison using below table template. Discuss what data pre-processing steps and what J48 parameter setting helped improve the test accuracy, and explain why. *(Please include a screen shot of your Kaggle prediction ranking for each model).*

(b) Output the best performing decision tree model (i.e. with the highest Kaggle ranking) and comments on any insights you consider interesting regarding the profile of those passengers who are more or less likely to survive.

Model	Data pre-processing steps	J48 parameters	Evaluation on training data	Evaluation on test data
1	For example: Exclude attributes X and Y. Discretize attribute Z to 5 bins with equal frequency.	Default setting	10-fold CV accuracy 80%	.75998
2	…			