Introduction to Machine Learning Experiment 1
Decision Tree Algorithm

March 12, 2015

This assignment asks you to implement the Decision Tree classifier and test your implementation on the given data set.

1 Data

The data set is from *UCI Machine Learning Repository*[1]. The *Adult Data Set*[2] aims at determining whether a person's income exceed a certain amount (50k), given the census data.

Each record (a line in the data file) corresponds to a person, and has *15 attributes* including the class label (the last attributes, $<= 50k$ or $> 50k$). There are both numerical and categorical attributes, and some of the values are missing (denoted by "?"). So you may need to handle them differently. We use 32561 of the records as training set (*adult.data*) and 15281 as test set (*adult.test*). The data set is moderately unbalanced with 75% of the records belonging to the majority class ($<= 50k$). A detailed description of the data set is in *adult.names*, which is in a C4.5 machine-readable format.

2000 records are set aside for a black-box test. 14 feature attributes are given in *test.features*. You should submit the predicted class labels in the same format as *test.result.example* (see Task 3 for details).

2 Tasks

Task 1

Implement the Decision Tree algorithm (no pruning for Task 1), train the model *only* on the training set and evaluate its performance on the test set. There are many ways to measure classification performance, you need to at least report the *accuracy*, which is

[1] `http://archive.ics.uci.edu/ml/index.html`
[2] `http://archive.ics.uci.edu/ml/datasets/Adult`

the proportion of the records that are correctly classified[3]:

$$Accuracy = \frac{\text{number of correctly classified records}}{\text{number of test records}}$$

You should also measure the impact of the training set size on performance. You must *randomly sample 5%, 50% and 100%* records from the whole training set to train your model, and test it on the whole test set. The random selection may affect the performance as well. So you should repeat the experiment, including the random sampling step, *at least 5 times* and report *min, max, and average* accuracies.

Task 2

Select and implement one of the *post-pruning* strategies. Use the same training set sizes as in Task 1 but *randomly select 40% of your selected training sets* as validation sets. Report the tree sizes and the accuracies on test set, and compare them with those in Task 1. Again you need to repeat *at least 5 times* and report *min, max, and average* accuracies.

Task 3

Use 100% of the training data, and select *the first 40%* of them as validation set, other 60% as training set, to train a post-pruned decision tree (as in Task 2). Then use this decision tree to predict the class labels for the 2000 records in *test.features*. You need to submit the predicted class labels in a text file.

3 Submission

Source code

With necessary comments. No restriction on programming languages, but make sure that TA can run your code easily.

README

A text file that briefly describes how to run your code and (re-)produce the reported results. Please also have your name, your student number and your contact information included in it.

Report

A pdf file that includes the following information:

- Your design of the experiments: the splitting criteria, the pruning strategy, and how you evaluate the performance. Don't just copy&paste the source code. Help us understand your design.

[3] `http://en.wikipedia.org/wiki/Accuracy#In_binary_classification`

- The experiment results: the results of Task 1 and Task 2. You are welcome to use more evaluation metrics or design further experiment to access the performance.
- Your analysis and discussion: For example, whether the model has an over-fitting problem and why. Tell us what you find in the experiment.

Result file for Task 3

A text file that contains 2000 predicted class labels. The file should have 2000 lines. Each line should contain a class label ($<= 50k$ or $> 50k$) and ends with a '.'. You should also name this file as *yourStudentId.test.result*, and make sure the TA can easily find it in your submission.

4 Deadline & Other Information

DEADLINE: Mar. 29, 2015

Upload the packed file with your name and student number in the filename to learn.tsinghua.edu.cn. Late submissions WILL NOT BE ACCEPTED.

Feel free to contact the TA for further information.

maojiaxin@gmail.com 15210588612