IST565 Data Mining

HW6: naïve Bayes and decision tree for handwriting recognition

Now that we have learned two classification algorithms, decision tree and naïve Bayes, let's think further on the question of choosing algorithms for a specific task. Note that there is no silver bullet in terms of algorithm comparison – no algorithm would outperform all other algorithms on all data sets. Therefore, choosing appropriate algorithms is an important decision, and it requires knowledge of both the data set and the candidate algorithms. In this homework, you will compare naïve Bayes and decision tree for handwriting recognition.

Task description:

The data set comes from the Kaggle Digit Recognizer competition. The goal is to recognize digits 0 to 9 in handwriting images. Because the original data set is too large to be loaded in Weka GUI, I have systematically sampled 10% of the data by selecting the 10^{th}, 20^{th} examples and so on. You are going to use the sampled data to construct prediction models using naïve Bayes and decision tree algorithms. Tune their parameters to get the best model (measured by cross validation) and compare which algorithms provide better model for this task.

Due to the large size of the test data, submission to Kaggle is not required for this task. However, 1 extra point will be given to successful submissions. One solution for the large test set is to separate it to several smaller test set, run prediction on each subset, and merge all prediction results to one file for submission. You can also try use the entire training data set, or re-sample a larger sample.

https://www.kaggle.com/c/digit-recognizer/data

Tip: check out the Kaggle forum to see if there are some patterns other people have found that you can use to build better models.

Report structure:

Section 1: Introduction

Briefly describe the classification problem and general data preprocessing. Note that some data preprocessing steps maybe specific to a particular algorithm. Report those steps under each algorithm section.

Section 2: Decision tree

Build a decision tree model. Tune the parameters, such as the pruning options, and report the 3-fold CV accuracy.

Section 3: Naïve Bayes

Build a naïve Bayes model. Tune the parameters, such as the discretization options, to compare results.

Section 4: Algorithm performance comparison

Compare the results from the two algorithms. Which one reached higher accuracy? Which one runs faster? Can you explain why?

Section 5: Kaggle test result (1 extra point)

Report the test accuracy for the naïve Bayes and decision tree models. Discuss whether overfitting occurs in these models.

Grading rubrics:

1. Are the models constructed correctly?
2. Is the result analysis conclusion convincing?
3. Is sufficient details provided for others to repeat the analysis?
4. Does the analysis include irrelevant content?
5. Successful submission to Kaggle?