

CSI660 – Data Mining Assignment #1): Mining with Trees and Naïve Bayes

Due: Friday 02/27/04

Worth: 20% of Final Grade

Late Policy: You lose one full grade for each week (including partial weeks) you are late.

Read the instructions carefully.

The purpose of this assignment is to understand the computational properties of using decision trees and Naïve Bayes classifiers for mining. Answer each question carefully, if you are not sure of a question's intent, then ask. For this assignment, you may implement your own decision tree and naïve Bayes algorithm or use standard existing code. Please nominate which option you are taking.

All data sets are on the web site.

For the Vote and Pima data sets

1. Report the predictive accuracy for models built from 10%, 20% ... 90% of the training data (the remainder is used for testing). Plot the predictive accuracy against training set size. What do you notice about the predictive accuracy? How can you explain this phenomenon?
2. Perform four fold cross validation for both naïve Bayes and decision tree classifiers. Report the mean and variance of the predictive accuracy. By examining the predictive accuracy and the models built for each situation how do decision trees and naïve Bayes what fundamental difference do you notice between the two types of classifier.
3. The formal definition of overfitting is:

Hypothesis $h \in H$ **overfits** training data if there is an alternative hypothesis $h' \in H$ such that

$$error_{train}(h) < error_{train}(h')$$

and

$$error_{\mathcal{D}}(h) > error_{\mathcal{D}}(h')$$

3a) How can you **empirically** verify that a decision tree is not overfitting the training data. Produce a decision tree that you believe has not overfitted the training data using your approach.

3b) Can a Naïve Bayes learner ever overfit the training data? Why?

4. Obtaining lift curves is important to analyze the performance of the classifier. Plot a lift curve for a decision tree learner and naïve Bayes classifier.
5. The decision tree classifier rank orders confidences with respect to the conditional probability $P(y=+|x)$. Can you suggest an improvement to obtain a better ranking? Explain your answer.

For bonus marks (optional)

Over and under-sampling are common approaches to allows a classifier to make predictions for rare events. Discuss some of the conditions the two approaches will work/fail under. Verify your assertions formally or with empirical evidence.