CSCI 6350 Spring 2006
Project 3 Classification (Due: Thursday, March 30th)

In this project, we experiment with two classification schemes: decision tree classification (part 1) and Naive Bayes classification. Different classification methods are typically evaluated in terms of how accurate a classifier built based on the training data predict the class labels for the testing data, which is disjoint from the training data. 10-fold cross validation is used for this purpose.

Classification with Naïve Bayes

Implementation

In this part of the project, you are going to implement a Naïve Bayes Classification system. You can use your choice of programming language. The format of data to be used by the classification system should be the same as what is used for C4.5, i.e., it reads class and attribute name information from the filestem.names file, and reads data from filestem.data file.

Your classification system should have two parts:

- learning the classifier :

  $P(C = C_j) = \frac{N(C = C_j)}{\sum_{j=1}^{J} N(C = C_j)}$, where J is the number of categories in data, and

  $P(X_i = V_{ik} \mid C_j) = \frac{N(X_i = V_{ik}, C = C_j) + 1}{N(C = C_j) + \mid X_i \mid}$, where X_i is attribute *i*, V_{ik} is value *k* for attribute *i*, and $|X_i|$ is the number of possible values for attribute *i*.

- classifying new data : to prevent underflow program, *logP* should be used in all computations.

  $c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} \log P(c_j) + \sum_{i \in positions} \log P(x_i \mid c_j)$

Your classification system should also have a chosen strategy that deals with missing values in data.

A cross validation module needs to be written that can perform n-fold cross validation. This does not have to be a shell program.

In addition, if an attribute is designated by user to be ***ignored***, it should not participate in the computation. (This feature is optional. You may assume that all attributes will be used in computations.)

Experiments

All experiments will be performed using the 2nd data set you have used in the decision tree project. This data set contains discrete valued attribute values only.

1. Learning the classifier

Divide the data set into two parts: training data (all but 10 data objects from the original data) and test data (10 data objects not included in the training data). Apply your program on the training data and show the naïve Bayes classifier learned.

2. Classify new data

For each test data, generate the classification for the 10 test data objects:

- Compare the predicted classifications to their actual classifications,
- Compare the prediction results with those generated from C4.5 (based on the decision tree or on the decision rules)

do they agree on the classification?

3. 10-fold cross validation

Perform 10-fold cross validation on the entire data set (test data + training data). Your program should display the prediction accuracy obtained from every 1 of the 10 test data sets.

What is the average prediction accuracy obtained? How does it compare to that obtained from C4.5?

Things to turn in:

- Email your program to cli@mtsu.edu
- Add to the mini-report from last project:
  - a paragraph describing the naïve Bayes classification method.
  - Your implementation specifics:
    - What strategy your program use for missing data
    - Special implementation issues/details you have encountered and solved.
    - What other features you should/could have added to the program?
  - Experimental results and analysis of the these results, include:
    - the classifier learned
    - Classification results generated with new data
    - Comparison of results from naïve Bayes and decision tree
    - 10 fold cross validation result and comparision
    - Analysis
  - Conclusion for the entire project (decision tree and naïve Bayes)