CIS 635 Data Mining

Homework 8

Description

In this homework you will be writing R scripts to compare data sets and algorithms using the performance statistics covered in lecture. You will also be modifying data sets to measure the effects of overfitting and the curse of dimensionality.

Instructions

Part 1 – which algorithm is best

Write an R script that takes a data table as input and returns the performance statistics (precision, recall, and accuracy) for the five difference algorithms, decision trees (rpart), naive Bayes (naiveBayes), K nearest neighbor (knn) , support vector machines (svm) and artificial neural networks (nnet). The return value of the script will be a 5 by 3 matrix of the statistics for each algorithm. For knn use k=3, for svm, using the linear kernel and for nnet use 4 hidden nodes. To calculate the statistics you will be using 10 fold cross validation.

Run your script on the 4 data sets (example1.txt, example2.txt, example4.txt and example5.txt) and report the results. You may get strange results for NB and SVM on example1; if so, just ignore them. For each set, report which algorithm performed the best (disregard knn for this part). Then also give a short, 1 sentence explanation for why it did well. After that, give a short, 1 sentence explanation for why knn did so well on all the sets. You may find it helpful to scatter plot each data set.

Part 2 – the curse

For this part you will be starting with data set example4.txt with only 2 variables and a class and slowly expanding the size of it to see the effects of the curse of dimensionality. The 2 variables contain the helpful information for classifying the data. The data that you will add to the table will be random uniform (runif) data that should not be helpful. You are to write 1 or 2 R scripts for this experiment. The script will iterate 20 times, each time adding a new column of uniform data from 1 to 100 (be sure that the data is integers). After adding the column calculate the accuracy using decision tree, naive Bayes and knn. Either show the results in a table or plot them on 3 line plots (line plots are better). In 2 sentences (or less) , explain the results.