Homework Assignment 4

Due date: Tue, November 2 (in class)

1 Nearest Neighbor Algorithm

In this assignment you will implement and experiment using the k-nearest neighbors algorithm.

Code

Write your own implementation of the k-nearest neighbor algorithm and experiment with the datasets glass.arff, heart-statlog.arff, sonar.arff that are available in the $WEKADATA directory. Note: these files include numerical attributes and a discrete class; your code should work with any dataset of this type. Your code

- should be usable for reading files in the weka format,
- should include facilities to normalize features (so they have zero mean and standard deviation of 1),
- should have a facility to estimate accuracy using stratified cross validation (write this as a function so it can be used multiple times)
- should include a facility to pick the value of $k$ by cross validation (write this as a function that picks $k$ for a given dataset or an appropriate subset of the dataset)

Experiments

1. Use this code to estimate the performance of the algorithm with and without normalization and across a range of values for $k$ (say 1 to 25). To visualize the results for each dataset plot accuracy, measured using stratified 10 fold cross validation, as a function of $k$ with and without normalization of features.

2. Use double cross validation to evaluate performance under automatic selection of $k$. To perform this, in each fold of the standard cross validation, you run another cross validation using the training set only and use it to choose $k$ for that fold. You then estimate the error on the test set for that fold using the selected $k$. Finally you calculate the average error rate across the folds.

   How does the accuracy compare to the results in the previous part? How stable is the choice of $k$ across different folds?

3. Implement a simple scheme for feature selection by evaluating the information gain for each feature, ranking features by information gain, and selecting the top $N$ attributes. (Notice that in this scheme we are not accounting for correlation among attributes when selecting.)

   Evaluate this method on the dataset sonar-withmanyfeatures.arff which is an extension of sonar.arff where we have added irrelevant features. For the evaluation, fix $k$ to some value giving reasonable performance on the previous part and then plot the performance of cross validation as a function of $N$ where $N$ varies from 1 to the full set of attributes.

   How does the accuracy compare to results with the original dataset? how does performance vary with the number of features selected?

4. [Extra Credit (up to 10 points)] Implement the wrapper method for forward search feature selection and evaluate it in a similar way.
\section*{2 What You Need to Submit}

Please submit hard-copies only. Your submission should include:

1. Source code of all the programs used to run your experiments. If it’s not obvious please give instructions on how to run your code. Please also make sure that your code is well documented and written with good style.

2. A short report on the experiments you ran and their results. Please include all numerical results and plots as requested above and please discuss any trends or lack thereof in the results.

\section*{3 Grading}

You assignment is graded based on the quality of the code, quality of documentation, and the quality of the report and observations.