Data Mining Assignment I
Classification using K Nearest Neighbours (KNN)

Submit your solutions to:
`pergu@it.uu.se`

Due date: 14/11/2005, 17:00

1 Background

This assignment is taken from the field of forensic science. The goal is to classify a number of glass samples in order to determine what kind of glass each of them is (building window glass, vehicle window glass, vehicle headlight glass, *etc*).

The classification is to be based on the K nearest neighbours method which you will need to implement in MATLAB.

2 Assignment

The data set can be found in the file `glassdata.mat` which is accessible through the course's homepage. When you load this file in MATLAB, you will find three matrices:

1. *GlassData* a 163*9 matrix where each row contains one observation, and each column corresponds to a certain meassurement (i.e., refractive index and the amount of magnesium),
2. *GlassClasses* a 163*1 matrix that contains the classes (1 to 6) corresponding to each observation in *GlassData*, and
3. *TestData* a 30*9 matrix of observations where the class of the objects is unknown.

Your assignment is the following:

- Implement the K nearest neighbour algorithm in MATLAB. Use the *Manhattan distance measure*
- Preprocess the data to fit the needs of the algorithm. There are no missing values and you may assume that there is no noise.
- Divide the data into suitable test and training sets and find the best value of K. *Note that no rules of thumb are accepted as the choice of K.*
- Use your implementation to classify the unclassified data that can be found in the *TestData* matrix.

3 Some Information on MATLAB

On most (if not in all) SUN machines in the department labs, MATLAB is started by typing `matlab` in the shell. MATLAB has an extensive help function, but if you need additional help there are links to several different tutorials at the course's homepage.

To save the variables that you have in your system in the file `result.mat`:

1. List all current variables by writing *whos* at the MATLAB prompt or by looking in the *workspace* window.
2. Remove the variables that you do not want to save by writing *clear VarName*
3. Save the remaining variables into the file `result.mat` by writing *save result*.

4 Report

As in all assignments of this course, you are allowed and encouraged to work in pairs. Your report should contain information about:

- Your names and e-mails.
- Pointer to MATLAB files and running directions. Please make sure that your code runs on MATLAB 6.0. (The version of MATLAB that is started with the `matlab` command in the shell.) Needless to mention, you are not allowed to update these files after the deadline.
- Pointer to a `.mat` file that contains (only) the variables:

  K containing the best value of neighbours.

  TestClasses containing the result of the classification of *TestData* in the same form as the *GlassClasses* matrix.
- A brief description of design decisions in your implementation. For example, how the data was preprocessed and divided into training and test sets.
- The reasoning behind your choice of K.
- A discussion of how the *apriori* knowledge of the importance of different characteristics of the classes can be used in the preprocessing.

Good luck!