Tutorial 5

1. (KNN and N-fold Cross Validation)

   Loan managers often need to take into account an applicant's demographic and socio-economic profiles in deciding whether to approve a loan to the applicant, to minimize losses due to defaults. In this exercise we will build and evaluate a classifier based on the German Credit Data to predict whether an applicant is considered as having good or bad credit risk. The features or predictors include (1) loan duration (in months), (2) credit amount, (3) Installment rate in percentage of disposable income and (4) age in years.

   (a) Read and explore the data from the file `German_credit.csv`.

   (b) Standardize the input features.

   (c) Randomly select 800 customer records to form the training data, and the remaining 200 records will be the test data.

   (d) Use 1-nearest neighbor classifier for the training data to predict if a loan applicant is credible for the 200 test points. Compute the accuracy of the classifier.

   (e) Use N-folds cross validation with $N = 5$ to find the average accuracy for the 1-nearest neighbor classifier.

   (f) Repeat question 1e for K-nearest neighbor classifiers where $K = 1, 2, ...100$.

   (g) Compare the 100 classifiers above, which few values of K give the best average accuracy?

2. (Decision Trees)

   Consider the famous Iris Flower Data set which was first introduced in 1936 by the famous statistician Ronald Fisher. This data set consists of 50 observations from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor).

   Four features were measured from each observation: the length and the width of the sepals and petals (in cm).

Source: `http://suruchifialoke.com`

   (a) Use decision tree method to predict Iris species based on all four features.

   (b) Visualize the decision tree above, using the `rpart.plot` function.

   (c) What are the more important features in the fitted tree above?