Tutorial 6

1. (DT and N-fold Cross Validation) 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`

   In Tutorial 5, we used decision tree method to predict Iris species based on all four features. We now would want to use N-fold CV to check on how good the method is, based on the accuracy.

   We'll use 5-fold CV where we would want to **keep the ratio of the three species the same (1:1:1) in both training set and test set**.

   What's the average accuracy of the decision tree method?

2. Recall that we studied N-fold cross-validation for the K-nearest neighbor classifier, in which the value of k is varied to control the complexity of the decision surface for the classifier.

   For decision tree classification, when fitting a tree using function `rpart()`, we use the argument

   `control = rpart.control( minsplit =1)`

   where `minsplit = 1` is to specify the minimum number of observations that must exist in a node in order for a split to be attempted. By default, `minsplit = 20`. This `minsplit` argument helps to draft the complexity of a tree, complex with many layers and branches or simple with few layers and less branches.

   For this `control = rpart.control()`, there is a similar complexity parameter exists, which is denoted as `cp` where by default `cp = 0.01`:

   `control = rpart.control(cp = 0.01)`.

   Heuristically, smaller values of `cp` correspond to decision trees of larger sizes, and hence more complex decision surfaces.

   For this problem, we will investigate n-fold cross validation for a decision tree classifier.

   Consider the data set 'bank-sample.csv' we discussed in the lectures. For this exercise, we will fit a decision tree with subscribed as outcome; and job, marital, education, default, housing, loan, contact and poutcome as 8 feature variables. **We want to find the best `cp` value in terms of misclassification error rate.**

   (a) Randomly split the entire data set into 10 mutually exclusive data sets.

   (b) Let `cp` take on the values 10^k for $k = -5, -4, ..., 0, ..., 3, 4, 5$.

(c) At each `cp` value, run the following loop for $j = 1, 2, ..., 10$:

i. Set the jth group to be the test set.

ii. Fit a decision tree on the other 9 sets with the value of `cp`.

iii. Predict the class assignment of `subscribed` for each observation of the test set.

iv. Calculate the number of mis-classification(s) by comparing predicted versus actual class labels in the test set.

(d) Determine the best `cp` value in terms of mis-classification error rate.
Note: mis-classification error rate is the complement of the accuracy.