# Tutorial 7

1. (Decision Trees)

   Customer churn is the loss of clients or customers. Banks, telephone service, companies, Internet service providers, pay TV companies and insurance firms often use customer churn analysis and customer churn rates as one of their key business metrics.

   This is because the cost of retaining an existing customer is far less than acquiring a new one. Companies from these sectors often have customer service branches which attempt to win back defecting clients, because recovered long-term customers can be worth much more to a company than newly recruited clients.

   In this problem , a wireless telecommunications company wants to predict whether a customer will churn (switch to a different company) in the next six months. With a reasonably accurate prediction of a person's churning, the sales and marketing groups can attempt to retain the customer by offering various incentives. Variables of our concern are listed below.

   (i) Age (years)

   (ii) Married (true/false)

   (iii) Duration as a customer (years)

   (iv) Churned contacts -Number of the customer's contacts that have churned (count)

   (v) Churned (true/false)—Whether the customer churned

   (a) Build a decision tree for predicting customer churn, using the feature variables Age, Married, Cust_years and Churned_contacts.

   (b) Consider the decision tree in exercise 1 to predict binary variable Churned. Use the tree to predict customer churn for the following observations.

| | Age | Married | Cust_years | Churned_contacts |
|---|---|---|---|---|
| 2821 | 26 | 1 | 2 | 2 |
| 96 | 23 | 1 | 3 | 3 |
| 5085 | 56 | 1 | 5 | 2 |
| 758 | 36 | 1 | 5 | 2 |
| 487 | 45 | 0 | 2 | 1 |
| 987 | 28 | 0 | 2 | 2 |
| 6061 | 22 | 1 | 3 | 0 |
| 3745 | 22 | 0 | 3 | 2 |
| 4709 | 60 | 1 | 2 | 1 |
| 2769 | 32 | 0 | 3 | 1 |

2. (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).

   In Tutorial 6, 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?


Source: `http://suruchifialoke.com`

3. 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, a similar complexity parameter exists, which is denoted as $C_p$. Heuristically, smaller values of $C_p$ 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 feature variables. We want to find the best $C_p$ value in terms of mis-classification error rate.

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

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

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

      i. Set the $j$th group to be the test set.

      ii. Fit a decision tree on the other 9 sets with the value of $C_p$.

      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 $C_p$ value in terms of mis-classification error rate.