Lab 5: Decision Tree

A. Objectives

- Getting familiar with fundamental concepts and applications of decision tree via **rpart** and **ctree**
- Learn about a case study with Uber data via rpart.

B. Documents

- http://www.statmethods.net/advstats/cart.html
- https://www.tutorialspoint.com/r/r_decision_tree.htm
- https://rstudio-pubs-static.s3.amazonaws.com/123438_3b9052ed40ec4cd2854b72d1aa154df9.html
- Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.

C. Submission

- Files related to one problem is stored in one folder
- Make a zip file and upload the file to blackboard

D. Exercise 1 –Weather dataset

1. Download Weather dataset from Weka repository (http://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/weather.arff) and format in csv. Note that the give file content both the description of data and the data, therefore you need to copy the @data into a text file before re-processing with excel.
2. Import Weather dataset into R and convert to data frame
3. **Use rpart to create a decision tree,** print out the result, plot the tree and answer the following questions:
   a. What is rpart?
      Hint **https://www.rdocumentation.org/packages/rpart/versions/4.1-12/topics/rpart**
   b. When should we use rpart?
   c. In the output of rpart, what are **root node error, n, nsplit, error, xerror** and **xstd**? **Explain the formulas and how to calculate these output values if applicable.**

4. Convert the categorical values in Weather dataset into numeric values (You can define your mapping, e.g., sunny=1, overcast=2 and rainy=3), use ctree to create a decision tree, print out the result, plot the tree and explain the following points:
   a. What is ctree?
   b. When should we use ctree?
   c. In the output of ctree, what are **n, err, number of inner nodes and number of terminal nodes? Explain the formulas and how to calculate these output values if applicable.**
5. Split the Wine dataset into 2 datasets for KNN. 1 dataset has 2/3 of total instances for training set, 1 dataset has 1/3 instances for testing set. For each case k=1, k=5 and k=20, do the following tasks:
   a. Scatterplot the distribution of clusters.
   b. Calculate the proportion of correct KNN classification. What is the best value of k? Explain the differences if applicable.

6. **Compare rpart, ctree and knn.**

E. Exercise 2-Uber dataset

Tutorial link: https://rpubs.com/shifanbamboo/uber

a. Data Exaction and Visualization
-Import Uber data from: https://raw.githubusercontent.com/bjherger/Uber-DS-Challenge/master/data/input/ds_challenge_v2_1_data.csv
-Get information data about First trip with Signup date, Background check date, Vehicle Info.Add data, and First Trip date.

b. Predict whether a user will take a first trip
Use decision tree to build a predictive model via **rpart**. Print out the result, plot the tree and explain.