Homework 2

Classification and Clustering

In this homework, you need to learn how to use the machine learning tool for classification and clustering

Follow the steps below:

1. Choose a machine learning tool
   (For example: python scikit-learn, R CRAN, weka…)
2. Download the dataset
   https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
3. Complete the tasks

Submission:

Hand in your code and the report (“word” or “pdf” or “ipython notebook” or “html”), and upload it to e3

Deadline:

- 2016/10/31 23:59

Presentation:

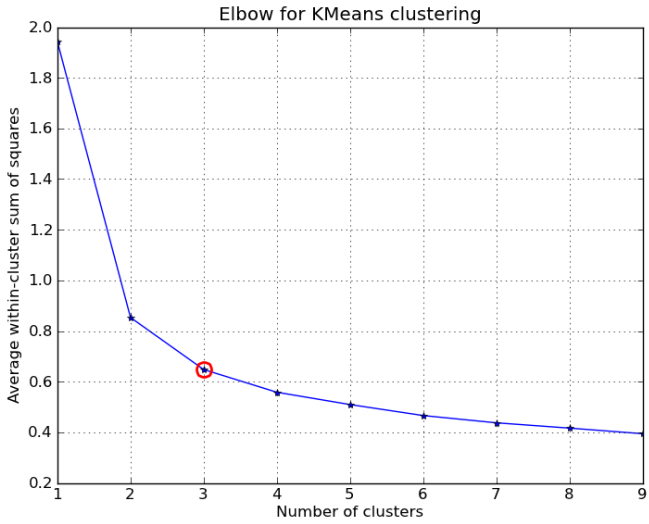
- We will choose some people to present in 2016/11/7 DM course

Classification

1. Split the data randomly to training data and test data (70% / 30%)
   將資料切成 70%的訓練資料，30%的測試資料
2. What is the **accuracy** of (1) Logistic Regression (2) k-Nearest Neighbors (3) Naive Bayes (4) Random Forest (5) SVM model in test data
   用五種模型訓練，在測試資料的準確度分別是多少
3. Draw the **ROC curve** in Logistic Regression
   畫出 Logistic Regression 的 ROC curve
4. Calculate the **precision and recall** in k-Nearest Neighbors
   計算 k-Nearest Neighbors 的 precision 和 recall
5. Draw the **Confusion Matrix** of Naive Bayes
   畫出 Naive Bayes 的 Confusion Matrix
6. What is the performance with different parameters in SVM
   不同參數差異在 SVM 模型的表現

Clustering

7. According to Gender, Education, and Marital status, how many kinds of customer should be divided into? (Hint: use the elbow method with K-means) 你認為根據 Gender , Education, and Marital status 三個欄位，可以將所有客戶分成幾類? (提示:利用 K-Means，elbow method)



8. Feel free and try more ^_^ 盡情的玩你所選的 tool 和這份資料