CMPSC497 Fall 2018 Programming Assignment 1.

Assigned: Wednesday, September 12, 2018

Due: Wednesday, September 26, 2018 (by midnight, submit a package of codes and report via Canvas)

Maximum: 100 point

Note: This assignment is to be done by an individual student, no team work allowed.

Data analysis and preprocessing is an important step before applying data mining and machine learning models for targeted tasks, such as classification and clustering. Data analysis aims to study the datasets in order to get insights of the data and understand difficulty of the target tasks, *e.g.* observe the value distribution of features and classes, and examine the correlation between classes and feature values. Data preprocessing aims to handle the noise of the data to make it suitable for further steps.

In the programming assignment 1, you are asked to perform data analysis and data preprocessing using the following dataset. You can use built-in function in *sklearn* and *matplotlib* for these tasks.

Dataset

- Bank Marketing Dataset [1] (https://archive.ics.uci.edu/ml/datasets/Bank+Marketing). The dataset is related with direct marketing campaigns of a Portuguese banking institution. The classification goal is to predict whether a client will subscribe a term deposit.
- Each data record, describing a client, contains the basic information of the client and whether the client subscribed the term project. Please treat column 1 to 20 as *features* and column 21 as the *class*.

Data Analysis

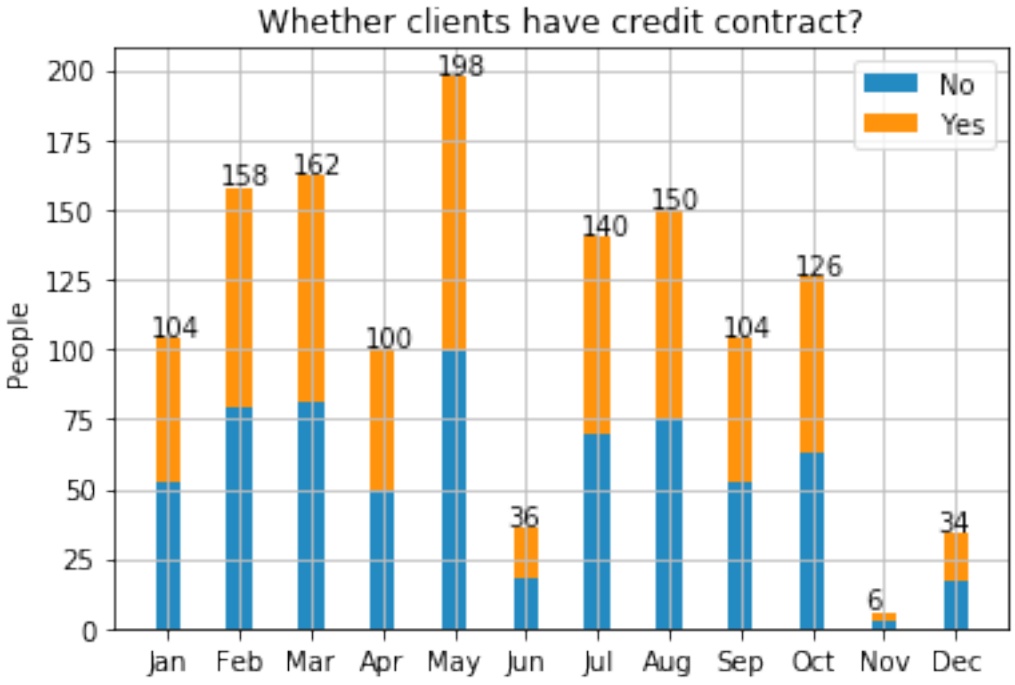
- Task 1. Plot the distribution of values in the *class* attribute of the dataset using a bar chart. *Please describe what you observe*, *e.g.* whether the data distribution is imbalanced.

- Task 2. Read the reference and answer the following questions.

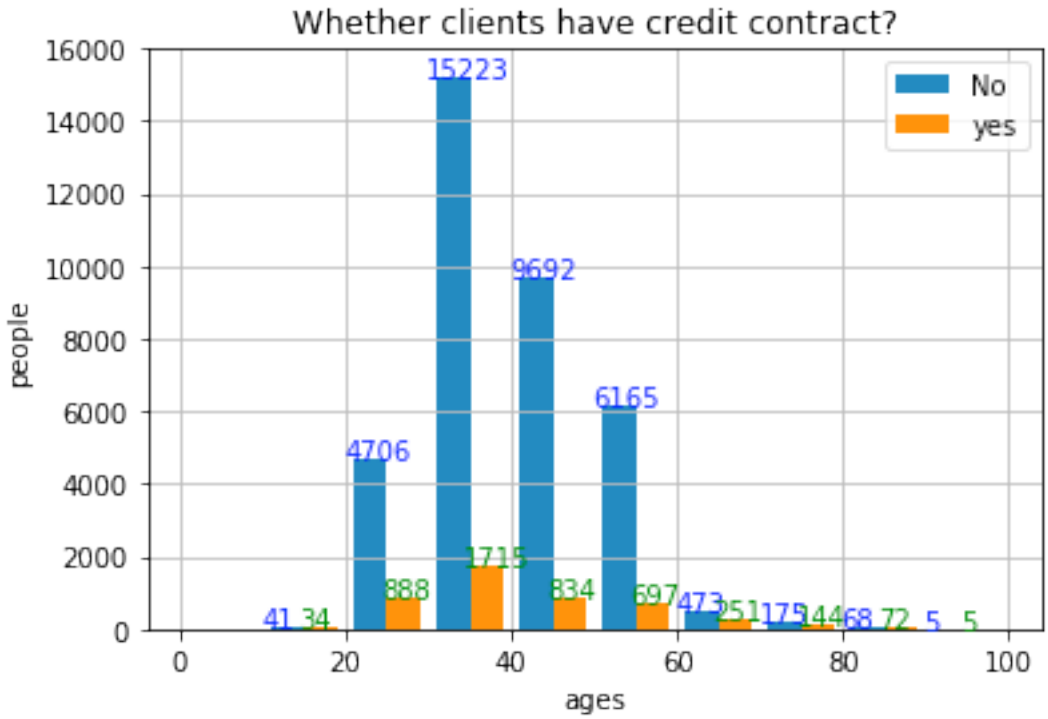
a) Please summarize the characteristics and differences of chi-square function (https://en.wikipedia.org/wiki/Chi-squared_test) and mutual information functions (https://en.wikipedia.org/wiki/Mutual_information).
b) Can we simply apply chi-square function and mutual information function on Bank Marketing Dataset for feature selection? Please explain. (hint: the difference between categorical and numerical data)
c) Employ chi-square or mutual information as appropriate to obtain a measure between values of each feature and the class. Rank features by their measures of chi-square and mutual information.
Note: Please make two lists: one for chi-square and the other for mutual information. An attribute only belongs to one list.

- Task 3. Based on the two ranked lists obtained in Task 2, plot the value distribution of (i) the highest ranked three categorical features, (ii) the lowest ranked three categorical features, (iii) the highest ranked three numerical features, and (iv) the lowest ranked three numerical features. Describe what you observe from these value distributions.
Note: Please plot a Bar chart and a Histogram for a categorical feature and a numerical feature, correspondingly. See below for examples. For Histogram, please evenly divide the overall value range into 10 intervals. For each bar and interval, please color the portion of records/instances corresponding to different classes and show the overall count.

Bar Chart



Histogram



Data preprocessing

- Task 3. Normalize the range of values of numerical features. If values are all positive or all negative, normalize them into [0, 1] or [-1, 0], respectively. Otherwise, normalize them into [-1, 1]. For each normalized numerical feature, submit the ranges of its original and normalized values.
- Task 4. Encode categorical features using *one-hot representation* scheme. For example, assuming that there is a 'state' feature with three categorical values, 'PA', 'NY' and 'NJ'. Create three new binary features, namely 'state_is_PA', 'state_is_NY' and 'state_is_NJ' to replace 'state', where the feature values are either 0 or 1. For each new binary feature, count and report the number of value 1, e.g., "state_is_PA": 15000, "state_is_NY": 20000 and "state_is_NJ": 10000.

Packages

- sklearn (http://scikit-learn.org/). A machine learning framework in Python
- matplotlib (https://matplotlib.org/). Website provides tutorials on how to plot bar chart and histogram in python.
- NumPy (http://scikit-learn.org/). A fundamental package for scientific computing with Python.

Reference

[1] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014.

[2] Please refer to http://scikit-learn.org/stable/modules/preprocessing.html#.