EE 660 **Homework Week 9: Project Proposal**

Posted: Tues., 10/22/2019
Due: Wed., 10/30/2019, 2:00 PM

Please fill in both the Project Proposal form (pp. 1-2) and the Dataset Information Form (p. 3). This is required of everyone (each team submits one HW9 with all their names on it). All fields except "other comments" are required. In each field, replace instructions (black text) with your descriptions. Preferred format is to enter your answers into the Word version of this form, then convert to pdf before submission. If you prefer to use another app instead of Word, then submit a typed version with each field labeled with its title ("Dataset", etc.), and submit as a pdf file.

Please note that this proposal will not be graded like a regular homework. The primary purpose is to give you some feedback on your project topic and plans; the scoring on this homework will be primarily based on whether you put in a reasonable effort and whether the content makes good technical sense.

Insert Project Title Here

Project team: Your name(s) and email address(es)

Qiong wang
wangqion@usc.edu

Project type (specify which):

I will design my own project by using real-world data.

Clear statement of the problem and/or goals.

In this project the goal is to find out the best model to decide whether a person should default credit card or not, and I will answer these following questions:

1. **The confusion matrix, accuracy, and F1 score of the best model.**
2. **How does the age influence the output?**
3. **How does the Education influence the output?**
4. **How does the age and Education together influence the output?**
5. **Find out the most three irrelevant features in data set.**
6. **To make a prediction about which kind of person shouldn't get credict card.**

A plan of preprocessing and feature extraction (if applicable)

Analysis all features, ignore the irrelevant features, and normalized training dataset. Moreover, some features are highly relevant, such as the Amount of previous payment in each month, in order to reduce the redundancy of features, I might calculate the mean, variance, range, etc. as other features.

A plan of your approach

Firstly, separate the dataset to training set, validation set, and test set randomly, then calculate the statistic characteristic about features, preprocess the training data. After that I might apply 3 different method to solve this classification problem. KNN, Logistic Regression, SVM and I will make a model selection by 10-fold-cv. I will get 3 different result according to these three different algorithms, then, I will calculate the F1 score to find out the best model. In the end, I will calculate $E_{D_{Test}}(h_g)$

1. In KNN, I will change the number of K to find out the best model, and calculate the confusion matrix, and draw ROC.
2. In LR, I will use l1-penalty, l2-penalty, and elastic net to find out the best model.
3. In SVM, by changing the penalty C.

A description of any other work of yours that is related to your class project

None

If yours is a team project, roughly describe how work will be divided

N/A

Other Comments

N/A.

EE 660

Dataset Information Form

Fall 2019

Include one form for each dataset you plan to use. (For each dataset's form, you may continue onto an additional page if necessary.)

Dataset or competition title: default of credit card clients Data Set

Link: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Student name(s): Qiong Wang

Brief description of dataset and problem domain: This research aimed at the case of customersâ€™ default payments in Taiwan and compares the predictive accuracy of probability of default among six data mining methods. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients. Because the real probability of default is unknown, this study presented the novel â€œSorting Smoothing Methodâ€□ to estimate the real probability of default. With the real probability of default as the response variable (Y), and the predictive probability of default as the independent variable (X), the simple linear regression result (Y = A + BX) shows that the forecasting model produced by artificial neural network has the highest coefficient of determination; its regression intercept (A) is close to zero, and regression coefficient (B) to one. Therefore, among the six data mining techniques, artificial neural network is the only one that can accurately estimate the real probability of default.

Number of data points: 30000 data points

Number of features or input variables: 23 features

Feature or input-variable types: 3 categorical features and 19 numerical features.

Label (output) type: binary categorical

If Label Type is Categorical, is the number of samples significantly unbalanced (maximal variation of more than a factor of 2)? No.

Problem type: classification

Has Missing Data? NO

If the problem/dataset is a Kaggle competition (current or past), answer:
(i) Is the competition current (give the end date), or past?

(ii) How much information is available on the Kaggle website (e.g., in "kernels" and links therein)? Briefly describe what type of information and code is available.