CMPSC497 Fall 2020 Programming Assignment 2.

Assigned: Friday October 16, 2020

Due: Monday, October 30, 2020 (by midnight, submit a package of codes in .ipynb and a report in .pdf via Canvas)

Maximum: 100 point

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

Based on the result of data preprocessing (in Programming Assignment 1), you are asked to learn classification models that predict the *default status of clients in next month* [1], and perform a number of tasks in order to evaluate the performance of models under different settings. You will learn a number of classification models using the preprocessed data, along with the features you identified previously in Assignment 1 and/or additional new features you would like to propose in this assignment. To find the best classifier, you are asked to exercise what you learn in the course to address various issues arising in the model training process and to fine-tune parameters in different models.

Dataset and new features

- *Credit Card Clients Dataset [1]*. Please add the features you propose (those in Assignment 1 and the new ones) to the dataset in accordance with the following format: Column 1 is the client ID and Column 2-24 are the features in the original dataset. Insert your proposed features in the subsequent columns, i.e., Column 25 and so on, and move the output class to the last column, e.g., Column 26 if you add only one new feature.

Classification Experiment

- Task 1. In this task, you will train a *logistic regression classifier* on the Credit Card Clients Dataset to predict whether a client will default in the next month.
    - There are 14 numerical features and 9 categorical features. Please train **Logistic Regression Model 1** based on normalized numerical features and one-hot encoded categorical features, and train **Logistic Regression Model 2** based on unnormalized numerical feature and one-hot encoded categorical feature. Feel free to train additional models if you would like to compare the effect of encoding on categorial features (but please state your goal in the report).

- Please use 5-Fold cross-validation for experiments. (See textbook and https://en.wikipedia.org/wiki/Cross-validation_(statistics))

Please summarize the definitions and mathematical formulae of **confusion matrix**, **precision metric**, **recall metric**, **f-measure metric,** and **accuracy metric.** Please compare the performance of **Logistic Regression Model 1** and **Logistic Regression Model 2** in terms of these four metrics.

Imbalanced Issue

- Task 2. Note that data imbalance exists in this dataset. Please explain why we want to avoid imbalance issue in training classifiers. Briefly summarize at least 3 methods to deal with data imbalance issue. Create a new dataset by either downsampling or oversampling and repeat the steps in Task 1 to compare the performance of generated dataset with the original dataset. Note that you only need to perform sampling on the training set but not on the testing set. Please explain why.

Feature Selection

- Task 3. Please list the reasons why we perform feature selection. Please perform feature selection based on the correlation results in PA 1 (using chi-square for categorical data and mutual information for numerical data). Generate partial datasets by only using top k (k =1, 3, 5) most correlated categorical features and numerical features for model training (i.e., k categorical features + k numerical features). Follow the setup in Task 1 to compare the performance of partial datasets with the original dataset.

Model Comparison

- Task 4. In addition to logistic regression, *decision tree*, *support vector machines*, and *multi-layer perceptron neural network* are also widely used for classification. Please follow the setup in Task 1 to compare the performance of these three models with the logistic regression model on both balanced and imbalanced datasets. In this task, please use the default settings of the hyper-parameters for the corresponding models.

Parameter Fine-tuning

- Task 5. Classification performance not only affected by the models used but also by their parameter settings. Please choose at least **two** of the four models you have applied so far, list the tunable hyper-parameters along with their value ranges. Then, perform hyper-parameter tuning on each of them, using five categorical features and five numerical features that you choose in Task 3 on the imbalanced (original) dataset. Please select/submit a trained classifier, along with your fine-tuned parameters, for testing. In other words, you want to choose the one that has the best classification accuracy (obtained in your validation). You are

required to perform parameter tuning on two models but encouraged to perform it on all four models in order to find the best model. For this task, your grade will be based on your model's performance on our testing dataset.

- Suggestion: grid_search() in sklearn is a very useful facility to perform parameter fine-tuning. It may help you reduce effort for tedious coding. Please see https://scikit-learn.org/stable/modules/grid_search.html

Deliverables

- Please submit the report (.pdf) and the complete Jupyter Notebook (.ipynb) on Canvas.
- The report should contain the experimental results for various tasks in the assignment.

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 in Python.
- pandas (https://pandas.pydata.org/). A framework for easy-to-use data analysis and manipulation in Python

Reference

- [1] I-Cheng Yeh, Che-hui Lien, The comparison of data mining techniques for the predictive accuracy of probability of default card clients, Expert Systems with Applications, Volume 36 Issue 2 Part 1, 2009, Pages 2473-2480
- https://www.sciencedirect.com/science/article/pii/S0957417407006719 (accessible on PSU VPN or PSU access)