Introduction to Data Science

Project - Phase 2

Instructors: **Dr. Bahrak, Dr. Yaghoobzadeh**

TA(s): Hamid Salemi, Mohammad Araghi

Deadline: Friday, Khordad 25th, 11:59 PM

Introduction

In this section, we will build upon the previous two phases to complete the project by training machine learning models to predict the target variable.

Important Note

The accuracy and output results will make up only a small part of your overall grade. Most of your grade is based on the **quality of your analysis** and the **quality of your project report**.

Part 1: Preprocessing

In the previous section (EDA), you were asked to perform various data preprocessing operations as needed in the data analysis stages. In this section, perform these operations with a machine learning algorithms approach **(Note: if you have already completed the following steps in the previous phase, there is no need to repeat them here).**

Explain your reasons for choosing the steps you took in preprocessing.

Part 2: Feature Engineering and Selection

Based on the nature of your data, apply specific feature engineering techniques to enhance the quality of your features. You may change these techniques after you check the performance of your models, to improve the metrics. Also, you may decide to select only specific features among all of your features for the next steps.

Explain your criterion and the reasons for the techniques you used in feature engineering and selection.

Part 3: Dimensionality Reduction

Using the PCA method, reduce the dimensions of numerical features to two dimensions. How much of the initial data variance is transferred to the new space?

If we aim to retain 95% of the original variance, what is the minimum number of dimensions required in the new space? **Save both the original data and the dimension-reduced one for the next parts.**

Part 4: Evaluation Metric

Choose appropriate evaluation metrics based on the nature of the data and the project goal, and **explain your reasons for choosing them**.

Part 5: Model Training

In this section, you need to implement three methods to predict your target variable.
First, split the initial data (including all features) into training and test sets.

Method One: Neural Network

Design and train a neural network for your goal. Then report the following:

- Error plot during training
- Model performance on the test data based on the evaluation metric (if the problem is classification, also report precision, recall, and f1-score, and explain which metric is more important in your problem)
- The network architecture, loss function, and optimization algorithm you chose, and explain your reasons for choosing them.

Methods Two & Three

Choose two methods from the following based on the problem goals and train the models:

- □ Linear/Logistic Regression
- □ SVM
- □ Decision Tree
- □ KNN

Then report the following:

- Error plot during training
- Model performance on the test data based on the evaluation metric (if the problem is classification, also report precision, recall, and f1-score, and explain which metric is more important in your problem)
- Model hyperparameters

Finally, compare the three implemented methods. **Which method performed better? Provide your analysis of this comparison.** Note that having 3 models is mandatory for this comparison, and adding more models, based on how this extra information improves the quality of your comparison, has a bonus score.

Part 6: Feature Analysis

Train the best-performing method from the previous section using the dimension-reduced data. **How did the model performance change? Provide your analysis.**

Part 7: Overall Report and Discussions

This is the last step of your project! Provide a brief report about your main steps from phase 0 till the end of this phase. We don't want detailed information in this report; only mentioning key decisions and ideas is enough. This will show the roadmap of your project. Also mention the problems and challenges you faced and your solutions for them, along with some alternatives.

Notes

- Upload your work in this format on the website: DS_Project_P2_[Std number].zip. If the project is done in a group, include all of the group members' student numbers in the name.
- If the project is done in a group, only one member must upload the work.
- This phase will not be accepted after its deadline i.e. there will be no late and grace policy!

Good luck!