DATA 302: Techniques in Machine Learning
Victoria University of Wellington
Trimester 1, 2024

Assignment 2: Machine Learning Pipeline

This assignment has **100** marks in total and is due on **23:59, 26th April, 2024**. Please submit your report as **a single .pdf file** including figures and tables as required, and **your source code as separate files** ("A2_Part1.ipynb" file and "Pipeline.py" file). If you modify "Helper.py" or have other supporting python files, please include them in your submission. Make sure you read the Assessment and Submission sections at the end of this assignment description. This assignment contributes **20%** to your overall course grade.

1 Objectives and Problem Descriptions

The goal of this assignment is to help you understand data manipulation and visualisation tools for machine learning. The purpose is to implement common data handling methods on real-world observations. To validate the effectiveness of the implemented methods, you are also required to perform data analysis tasks to draw useful conclusions. In particular, the following topics should be reviewed:

- CRISP-DM
- Machine Learning Pipeline
- Exploratory Data Analysis (EDA)
- Data Preprocessing
- Feature Selection and Feature Construction

It requires use of *python, numpy, matplotlib, scipy, and scikit-learn*, and serves as an introduction to all those tools. You can run Python based on the template Jupyter notebook and Python code templates provided.

In this assignment, your task is to build a machine learning pipeline to predict customers' credit risks (*good* or *bad*) of a German bank. The dataset *Credit* and the feature descriptions can be downloaded from the Assignment page.

2 Data Understanding [20 marks]

The first part of this assignment is to explore the data and to define the machine learning task. You should:

1. Perform EDA as an initial step to analyse the *Credit* dataset. The analyses should be conducted on the whole dataset, i.e., on the *"Data.csv"* in the data folder. The analyses should explore the data from the four different aspects:

  - Describe the *summary statistics* of the data. This should include the number of instances and number of features. Report the number of categorical and numerical features separately.
  - Identify the *top three numerical features* with the highest correlation with the target variable *Credit Risk* according to the Pearson correlation, and report their correlation values.
  - Plot the distributions of these three numerical features identified in the previous question and the target variable using histograms. One histogram for each feature/variable. Describe how to determine the number of bins to draw the histograms. Based on the histograms, describe the shape of their distributions (i.e., *Positive or Negative or Zero*) with respect to their *skewness* and *kurtosis* (use *Scipy* for obtaining *skewness* and *kurtosis* values).
  - Check for missing values. Write a paragraph to briefly summarise how many features containing missing values and the percentage of missing values for each incomplete feature.

2. Among the three machine learning tasks: *classification*, *regression*, and *clustering*, which one does this problem belong to? Justify your answer.

Provide answers to above questions in your report. Submit your Jupyter Notebook file (.ipynb) or your Python file (.py) that shows how you get the answers.

3 Data Preprocessing [20 marks]

It is crucial to partition the data prior to preprocessing in any supervised learning task to prevent data leakage. Therefore, we must split the whole dataset into a training set and a test data. Any preprocessing model must be trained on the training set only. The trained preprocsing model, then, can be applied to process both the training set and the test set.

You should *determine* the appropriate approaches to perform the following preprocessing steps.

- Encoding categorical data to numerical data.
- Handling missing data.
- Normalising/standardising the data.

Describe your chosen approaches and and your rationale for selecting them in your report. Show the preprocessing steps in the *preprocess()* function which takes the original training and test sets as its input and outputs the **processed** training and test sets. All following questions (Parts 4 and 5) should use the **processed** sets.

4 Feature Ranking [20 marks]

A straightforward feature selection approach is to rank features based on their relevance to the target output. Then, we can select the top-ranked features for use in our machine

learning task. In this part of the assignment, you will use *Mutual Information* to rank features. The higher the mutual information score, the better the feature.
You must use *sklearn.feature_ selection.mutual_ info_ classif* to calculate the mutual information between each feature and the target variable.
You should:

1. Implement the $feature\_ranking()$ method that takes a training set as its input and outputs a feature subset containing top five features.

2. Write a short report that includes:

   - Report the top five features selected by the aforementioned feature selection process.
   - Evaluate and compare the performance on the test set using the subset containing the top five features and the original feature set. Determine which one is better and provide your justification.
   - Use a heatmap to show the Pearson correlation between the top five features. In your report, you should show your heatmap, provide an analysis of the visualisation, and interpret how the features relate to each other.

5 Sequential Forward Feature Selection [40 marks]

Sequential Forward Feature Selection (SFFS) is a well-known feature selection method. The task of this part is to **implement the SFFS algorithm in Python based on the provided code template and then examine the selected features**. The pseudocode of the algorithm can be seen in Algorithm 1.

Algorithm 1 Sequential Forward Feature Selection

```
Input: Training set with $D$ features, number of selected features $d$
Output: A selected feature set $S$
Initialize set of selected features: $S \leftarrow \emptyset$
Initialize set of remaining features: $R \leftarrow \{f_1, f_2, ..., f_D\}$
for $i = 1$ to $d$ do
    Find the feature $f^*$ in $R$ that achieves the best score when combined with $S$
    Remove $f^*$ from $R$
    Add $f^*$ to $S$
end for
return $S$
```

5.1 Implementation

You will need to complete two methods: *sequential_feature_ selection()* and *sequential_ score()*:

- *sequential_feature_ selection()* is a method that takes a training set and the number of features d that you wish to select. The method starts with an empty set, and iteratively adds one feature at a time to the set until d features are selected in the set. The method finally outputs the selected feature set. *sequential_feature_ selection()* will use *sequential_ score()* to determine which feature to add at each step.

- *sequential_ score()* is a method that takes a feature subset S and a training set. It uses *10-fold* cross validation to evaluate the classification performance of the feature subset S. In this part, we will use $KNN(K = 3)$ as the classifier.

5.2 Report

Write a concise report that includes:

1. Is the implemented SFFS algorithm a *filter, embedded,* or *wrapper* feature selection approach? Justify your answer.
2. Is the implemented SFFS algorithm a *feature ranking* or *feature subset* selection approach? Justify your answer.
3. Set the number of selected features d to five and run the sequential selection algorithm.
    - Report the five selected features and the testing performance achieved when using these five features.
    - Use a heatmap to show the Pearson correlation between the five selected features. Provide an analysis of the visualisation, interpreting how the features relate to each other.
    - Compare the testing performance of the implemented SFFS algorithm and the testing performance of the feature ranking algorithm. Discuss which one is better and provide your justification.
4. Execute the SFFS algorithm with the following numbers of selected features {1, 5, 10, 15, 20, 30}. Each number of selected feature yields a corresponding testing performance. Use a scatter plot to visualise the relationship between the number of selected features and the testing performance. Based on the visualisation, provide a detailed discussion on the observed relationship, focusing on how the number of features impacts the testing performance, and identify any trend as the number of features changes.

Assessment

- **Format**: You can use any font to write the report, with a minimum of single spacing and 11-point size (handwriting is not permitted without approval from the course coordinator). Reports are expected to be at most 8 pages that cover all the questions described above.
- **Late Penalties**: Late submissions for assignments will be managed under the "Three Late Day Policy". You will have three automatic extension days, which can be applied to any assignments throughout the course. No formal application is required; instead, any remaining late hours will be automatically deducted when submitting assignments after the due date. You have the flexibility to use only a portion of your late day and retain the remainder for future use. Please note that these three days are for the whole course, not for each assignment. The penalty for assignments that are handed in late without prior arrangement (or use of "late

days") is one grade reduction per day. Assignments that are more than one week late will not be marked. If you require any extension due to exceptional circumstances (like medical), you need to email the course coordinator.