Machine Learning Pipeline for Credit Risk Analysis with the German Credit Data

Due: June 1st, 2023 *** **Late Accepted Until:** June 4, 2023 w. 5pts off every day ***

Goal

The goal of the CS412-Machine Learning project is to gain hands-on experience on the ML project pipeline.

In this project, you will have the opportunity to apply several classification algorithms to the real-world problem of credit risk assessment using the widely used benchmark: *German Credit (Statlog) Dataset*. You will gain practical experience in selecting and evaluating classification algorithms and determining the most important features for predicting credit risk. Through this practical experience, you will gain a better understanding of how machine learning can be applied in real-world scenarios and the importance of feature selection in building effective models for credit risk assessment.

Dataset

The German Credit (Statlog) Dataset is a widely used benchmark dataset for credit risk assessment. It contains $1,000$ instances, each represented by 9 features such as age, sex, credit history, and job stability. The dataset includes 700 instances labeled as "good credit" and 300 instances labeled as "bad credit". The goal is to predict whether a given loan applicant is likely to have good or bad credit based on the available features.

You will be provided with the German Credit Dataset in a comma-separated values (CSV) format, which will be available on SUCourse. The whole dataset will be provided in a single file format, with $1,000$ rows and 10 columns. The last column represents the class label, while the remaining 9 columns correspond to the 9 features of the dataset.

It is important to note that some of the features in the dataset are categorical, such as "Housing" and "Checking account". For these features, you may need to use one-hot encoding or ordinal encoding depending on the nature of the feature and the algorithm used. Additionally, the dataset contains several features where there are NaN values, which need to be addressed through data imputation techniques such as mean imputation or KNN imputation.

Therefore, before building any models, it is necessary to pre-process the dataset appropriately by handling missing values and encoding categorical features.

Methodology

Here is a list of the classification algorithms that you can apply to the problem (you should choose at least 3 that you deem suitable for the project):

- K-Nearest Neighbors (KNN)
- Decision Trees (DT)
- Logistic Regression (LR)
- Multilayer Perceptrons (MLP)
- Support Vector Machines (SVM)

For each method above, you are expected to:

- Choose hyperparameters (at least 1, up to 3) using 5-folds
- Train the final model using all folds with the best hyper-parameters
- Evaluate the final model's performance on your test data

Some of the methods above may or may not require a data preprocessing step (one-hot encoding, feature selection, scaling, normalization), we left you to decide what kind of data preprocessing strategy might be used for a given method.

Use a seed at the first cell in your notebook (before any random number generation) so that your results are comnparable/reproducable. We can say seed=42.

Bonus: Up to 10pts

Once you have completed training and evaluation of 3 baseline techniques discussed in the previous section, you may want to explore more advanced techniques to further improve your results. For instance, you could try a well-founded feature selection or extraction technique, or use an ensemble of some of the base systems developed in the main part. Data exploration techniques can also help you gain a better understanding of the dataset and its underlying patterns.

Report

Your report should include (with the appropriate section titles in this order):

- **Summary/Abstract**: A 1-2 paragraph summary of your work. E.g. "We tried .... algorithms for the German Credit Data problem and observed their performances after hyper-parameter optimization for each one separately. The best one was .... with %x.... accuracy."
- **Introduction**: Briefly explain the task and its importance (mention classification, briefly mention dataset and number of classes...)
- **Dataset**: Explain the dataset in a bit more detail (number of samples in each subset, train-test split, ...) and include examples from both classes, so that your report is self-contained.
- **Methodology**: Summarize the algorithms you have worked on, mention if any preprocessing is applied, and which hyperparameters are tuned and why/how. If a method was unsuccessful, you can summarize the finding saying that "we have also tried ...., but it diverged/took very long/...." without going into more detail.
- **Experiments**: Include tables, figures for all hyperparameter tuning phases. You should have one big table where you present the final performances of all algorithms you tried. If applicable (e.g. in NN based approaches), you should also include a training curve (train and validation loss per epoch).
- **Discussion**: Here you can briefly discuss the best algorithm and do a brief error analysis (you can give confusion matrix here along with the type of most seen errors...).
- **Bonus**: Describe if you have done anything above the basic requirements, such as those discussed in the Bonus section above.
- **Conclusion**: Briefly explain your final thoughts, did the experiments yield the same results as in your intuition, were you surprised in any of your findings? etc.

Important: Do not give a chronological account of your work (e.g. "we first tried this, and then ...". Instead, give how those algorithms works without the chronology of events. (e.g. "we have tried these 3 approaches and ... worked best with an accuracy of ....")

Grading

Your project will be graded based on:

- **Work**: Amount of work and whether have you followed the right steps in evaluation, model selection...: 40pts
- **Report**: 40pts
- **Results compared to others**: 20pts (best group will have 20pts and grade will drop 2pts for approximately 0.1% drop in accuracy difference, down to the minimum of 10pts).
- **Bonus**: Up to 10pts

Submission

Person with the smallest ID number in the group will submit the **PDF report** and **ipynb files** to SUCourse. For each deliverable mentioned, please include your names and student IDs.