Lab Course Machine Learning

Exercise Sheet 10

Prof. Dr. Dr. Lars Schmidt-Thieme
Kiran Madhusudhanan

HiWi: Basharat Mubashir Ahmed
Sophia Damilola Lawal

Submission deadline : January 29, 2023

General Instructions

1. Perform a data analysis, deal with missing values and any outliers.
2. Unless explicitly noted, you are not allowed to use scikit, sklearn or any other library for solving any part.
3. Data should be normalized.
4. Convert any non-numeric values to numeric values. For example you can replace a country name with an integer value or more appropriately use one-hot encoding.

1 Introduction to Natural Language Processing. (5 points)

In this lab, you will be working with the IMBD movie review dataset to perform various natural language processing tasks. Using the provided dataset, you will need to:

1. Perform tokenization on the review text.
2. Remove stop words from the tokenized text.
3. Use regular expressions to clean the text, removing any HTML tags, emails, and other unnecessary information.
4. Convert the cleaned data into a TF-IDF and BOW representation from scratch.

2 Support Vector Machines. (10 points)

In this assignment, you will be using the credit card fraud detection dataset from https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud to train and test a Support Vector Machine (SVM) classifier. Your task is to:

1. Split the dataset into training and testing sets.
2. Perform a stratified split to take care of the class imbalance.
3. Use sklearn to deal with the class imbalance and experiment the various ways we can deal with class imbalance, for e.g, Undersampling, Oversampling, SMOTE.
4. Implement the basic Pegasos Algorithm from the paper https://home.ttic.edu/~nati/Publications/PegasosMPB.pdf. This is in page 5, Fig 1.
5. Implement the mini-batch Pegasos algorithm from the paper https://home.ttic.edu/~nati/Publications/PegasosMPB.pdf. Do not forget the projection step. This is in page 6, Fig 2.
6. Implement the dual coordinate descent method for SVM's from the paper https://icml.cc/Conferences/2008/papers/166.pdf. This is Algorithm 1 in the paper.

   Report a final accuracy on the test set.

3 Optuna - A hyperparameter optimization framework

https://optuna.org/ **(5 points)**

In this problem, you will use sklearn's implementation of SVM. Explore all the hyperparameters supported, kernels and their importance. You are required to perform the hyperparameter tuning using the Optuna framework. The grade for this problem will be assigned based on the depth of analysis, that is, the desired plots of which hyperparameter is the most important, which areas in the search spaces were explored and so on. Use the following dataset for the task: https://www.kaggle.com/datasets/pcbreviglieri/smart-grid-stability.

Report a final accuracy on the test set with the detailed report on the parameters tuned and results obtained using the Optuna framework.

Reminder: Grade will be awarded based on the depth of the hyperparameter optimization done and the plots produced using Optuna! It is advisable to use Colab for this Question as it has GPU support.