AIMS African Institute for Mathematical Sciences SENEGAL

Kernel Methods

# Project: Data Challenge
# Team Name : A_O

ADOU Kouame Mathurin

Pauline Ornela MEGNE CHOUDJA

July 5, 2022

## 1 Introduction

Data challenge is a way to learn how to implement machine learning algorithms, gain understanding about them and adapt them to structural data. With this in mind, we will study one of the machine learning tasks which is the classification problem (predicting whether a DNA sequence belongs to the Covid-19). In this challenge our main goal is to implement these algorithms from scratch and this short report aims at summarizing our approaches.

## 2 Data description

The data we used for this project are sets of DNA sequencing reads: short DNA fragments ( $\sim$ 100 to $\sim$ 300 bp long), which come from sequencing experiments, or have been simulated from complete genomes. Some of these fragments come from Covid-19 genomes, others from humans or random bacteria.

Note that the data provided to us had already been vectorised using the Spectrum kernel [1] approach. The Fig1 below shows the distribution of target, where the labels are either 1 if the fragment is identified as Covid-19, and 0 otherwise.

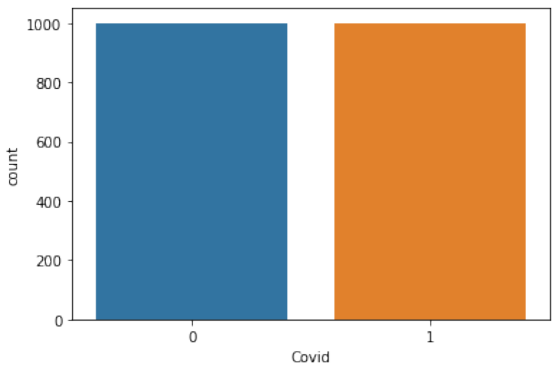


Figure 1: Labels distribution

## 3 Methods

The data were classified using multiple approaches mostly based on our course notes and tutorial sessions.

- Logistic Ridge Regression (LRR): the training was performed with the gradient descent (GD) and Newton-Raphson (NR) method;
- Kernel Logistic Regression (KLR), where the type of kernels (linear, polynomial and the Gaussian radial basis function/rbf), sigma, the regularizer $\lambda$ and the degree were considered as the main hyperparameters;
- Kernel SVM, here the main hyperparameter was the value of C, tuned between (1, 10,100 and 1000).

Our best model is the KLR obtained with the following hyperparameters: rbf kernel, $\sigma = 0.46$ (obtain using the median heuristic approach), and $\lambda = 0.0001$. The table below is a summary of the scores obtained for each model respectively for the training and validation sets.

| | **Accuracy (%)** | |
|---|---|---|
| | Training set | Validation set |
| LRR with GD | 78.43 | 80.00 |
| LRR with NR | 94.03 | 95.15 |
| **KLR** ($\sigma = 0.46, \lambda = 0.0001$,method=NR) | **97.46** | **96.21** |
| Kernel SVM, C=100 | 100.00 | 95.76 |

Table 1: Recap of observations

**Note:** The script file only contains the implementation of the Kernel Logistic Regression,which give us the best score among all the models we used. The entire notebook with all the different approaches can be find here.

## 4 Conclusion and Perspectives

Our final result when consisering the Kernel Logistic Regression give an accuracy of **96.40%** in the public leaderboard and **97.40%** in the private one.

In order to improve the performance of the model we are looking forward to reconsider the vectorization process using the mismatch kernel, and also tune more the hyperparameters.

## References

[1] C. Leslie, E. Eskin, and W. S. Noble. The spectrum kernel: A string kernel for svm protein classification. In *Biocomputing 2002*, pages 564–575. World Scientific, 2001.