DNA Sequence Classification: SARS-CoV-2 (Covid-19)

Agisha Ntwali Albert (aagisha@aimsammi.org)
Mouhamadou Mansour Sow (msow@aimsammi.org)

Team: Albaro **Public score:** 0.994 - **Private score:** 0.990
https://github.com/mm230/Kernel_Methods_data_challenge.git

1 Introduction

This report on the Kaggle data challenge was completed as part of the "Kernel Methods in Machine Learning" course taught by Jean-Philippe Vert, Julien Mairal, Michael Arbel, and Romain Menegaux as part of the AMMI master's degree program. The project's goals include studying machine learning techniques, learning how to use them, and adapting them to structural data. Due of this, we have chosen the challenge of predicting whether a DNA sequence (or read) belongs to SARS-CoV-2 as our sequence classification task (Covid-19). Here, we outline the methods we used, our tests, and some of the outcomes. The best submissions were produced by using the kernel ridge regression (KRR) for the classification on vectorized dataset of the DNA Sequences. And curiously, we used the mismatch kernel method, that enabled to deal with the raw DNA sequences dataset.

2 Datasets

Short DNA fragments (between 100 and 300 bp long), either from sequencing experiments or simulating full genomes, make up the training and evaluation data sets. These fragments come from several sources, including human or random bacteria, as well as Covid-19 genomes.
The task is a binary classification problem since the goal is to identify among sequences the Covid-19 fragments. If the fragment is classified as Covid-19, the label will be 1, otherwise it will be 0.

3 Methods: Classifiers

The goal is to find a better performing classifier, ideally one that is as simple as possible as a baseline. This will begin with the implementation of a linear classifier, followed by a non-linear classifier, in this case one from the kernel family. Due to report length constraints, we will only present briefly the used classifier that allowed us to reach our final score.

Kernels that operate on vectors Vectorized versions of each dataset are provided as matrices in addition to the DNA sequences provided in the data challenge. This format was explored in the initial Kernels that we implemented. The most common being linear kernels, polynomial kernels, and gaussian kernels.

Kernel Ridge Regression (KRR) Ridge regression and classification (linear least squares with l2-norm regularization) are combined with the kernel trick in Kernel ridge regression (KRR). As a result, it learns a linear function in the space induced by the kernel and data. This corresponds to a non-linear function in the original space for non-linear kernels.

Kernels that operate on DNA sequences Following the use of the Polynomial and Gaussian Kernels with the given vetorized data, we were fascinated to use the string kernels that are appropriate for DNA Sequences. In our case, we used the Mismatch Kernel to improve our predictions. **Mismatch Kernel** measures the sequence similarity based on shared occurrences of k-length subsequences, counted with up to m mismatches, which is what we generally state as (k, m)-mismatch[1].

Classifiers As advised, we first implemented the Kernel Ridge Regression (KRR) using the **RBF** kernel and obtained 0.93 accuracy. We also try the Kernel Ridge with the **polynomial** kernel and get an accuracy of 0.90. As a result, we ultimately decided to keep the submission related to the kernel Ridge with the RBF kernel, in the first case. Also, the Mismmatch Kernel with KRR was used, and the result was 0.99.
The classifiers were implemented using mostly the course slides and labs, with no other library than `cvxopt`, `scipy` to deal with the data sparsity and `numpy`.

4 Results

For the competition, we used two approaches: one that dealt with vectorized data and another that used DNA sequences (raw data). The different outcomes of our trial are displayed in the figure 1. And the discussion will be made on the interesting public scores we had on the Leaderboard.

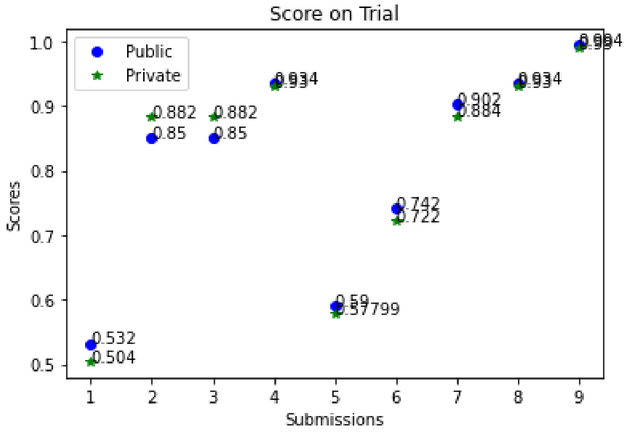

Figure 1: Submission Leader Board results and improvements

The best value we could obtain using the previously described Kernel Ridge Regression with $\lambda = 0.1$ and $\sigma = 0.4$ with the provided vectorized data was 0.93. Using the Mismatch Kernel with the DNA sequences and the Kernel Ridge Classifier (based KRR), we obtain 0.994 as public score and 0.990 as private score, putting us in ninth place. The performance was achieved by adjusting $\lambda = 0.001, 0.1$, the mismatches $k = 11, k = 12$, and performing cross validation (k-Fold).

5 Conclusion

In this challenge, we built Kernel methods from the ground up, allowing us to identify SARS-CoV-2 (covid 19) DNA sequences in a dataset of different DNA sequences. One of the keys to improving kernel ridge regression predictions is vectorisation methods (embeddings) based on raw sequence data. Finally, kernels designed for biological sequences were used, such as Mismatch Kernel [1], which improved our results.

References

[1] Christina S. Leslie, Eleazar Eskin, Adiel Cohen, Jason Weston, and William Stafford Noble. Mismatch string kernels for discriminative protein classification. *Bioinformatics*, 20(4):467–476, 01 2004.

[1] https://string-kernel.readthedocs.io/en/latest/mismatch.html