Predicting whether a DNA sequence belongs to the SARS-CoV-2 (Covid-19) [Github][Colab]

by **Kaningini Lutala Netho Junior**
African Masters Of Machine Intelligence

1. Introduction

In the context of the course ”Machine Learning with Kernel Methods”taught by Julien Mairal & Jean-Philippe Vert and Roumain Ménégaux (T.A.) to the students of the African Master of Machine Intelligence (2021–2022), we were asked to solve a Kaggle data challenge whose objective was to implement kernel-based classification methods to predict whether a DNA fragment taken during a DNA sequencing is a Covid-19 fragment or not.
Here we present the method that we used, our experiments, and some results. The best submissions were obtained using the kernel applied directly to the DNA sequences, combined into a single RBF kernel, and then used in an SVM as a binary classifier.

2. Dataset

The predictions are made on one data set. In this dataset, we have 2000 training sequences and 1000 test sequences. We have to point out that this dataset gathers data that are labeled .

3. Kernels & Classifiers

- **Kernels** :

  We know that in machine learning, a kernel refers to a method that allows us to apply linear classifiers to nonlinear problems by mapping nonlinear data to a higher-dimensional space without needing to visit or understand that higher-dimensional space. In other words, in classification problems, where the task is to classify different classes based on known input labels (supervised learning), we have different methods, one of which is the SVM (Support Vector Machine). Kernel methods (kernel tricks) are used in SVM. Kernel in machine learning is used to deal with the non-linearity present in the data set. A user-specified kernel function (similarity function) adds another dimension to the dataset, which allows the dataset to be classified using a linear hyperplane.

  In our challenge, we opted for kernels that operate on vectors. In addition to the DNA sequences provided in the data challenge, there are vectorized versions given as matrices, for the dataset. The kernel we have implemented handles this format. These are the linear kernel, the polynomial kernel and the Gaussian kernel.

- **Classifiers :**

  In SVM classification, the data can be linear or non-linear. There are different kernels that can be defined in an SVM classifier. For a linear dataset, we can define the kernel as **”linear”**. On the other hand, for a non-linear dataset, there are other kernels, such as **rbf**, **polynomial**, etc..

In this challenge, we had to use these three classifiers with our model: rbf, linear and polynomial. And this allowed us to compare the different performance results obtained by each classifier in our model.

4. Rusult

As mentioned above, we had to experiment our SVM model with the different classifiers. An SVM with a Radial basis function (rbf) kernel with C = 200 and gamma = 1 gave us a good baseline since we obtained 0.972 on the public ranking and 0.974 on the private ranking and this by using the original dataset. However, most of our experiments on the vectorized data we were given could not do better than 0.94.

Classifier	kmer_size	C	Gamma	Degree	Accuracy
rbf	4	200	1	2	0.9720
polynomial	4	50	1	3	0.9600
linear	3	50	1	2	0.9460
rbf	(vector)	1	1	2	0.7160

Tab 1 : Comparative table of performance of classifiers with respect to different values of C and gamma

Conclusion : We implemented the Kernel methods from scratch as requested for this challenge. We obtained a score of 0.972 for the public ranking and 0.974 for the private ranking, which is a satisfaction for us. We think that this score could be further improved by using for example Mismatch Kernels and also making an optimal choice of the value of C in the SVM.