DNA SEQUENCE CLASSIFICATION WITH KERNEL METHODS

Data challenge report

AMMI SENEGAL 2022, Mbour

Team: AF Kernel **Public score:** 0.99800 (2^{th}) **Private score:** 0.99400 (6^{th})
Team member: Annine Duclaire Kenne Fenosoa Randrianjatovo

1 Introduction

The notion of kernel, recently introduced, has attracted a lot of interest because it allows obtaining nonlinear algorithms from linear algorithms in a simple and efficient manner. The main objective of this challenge is to implement the kernel-based methods able to tackle the classification problem i.e. predict whether a DNA sequence (or read) belongs to the SARS-CoV-2 (Covid-19). The different methods used, our experiments and the results obtained are presented. The best score was obtained using a linear kernel applied on the DNA sequence transformed using spectral embedding and then used into the SVM Machine Learning method as binary classifier. In the following sections, additional details about our experiments are provided.

2 Task & Datasets

This data challenge contains one dataset of 2000 labeled training sequences of DNA, as well as 1000 unlabeled test sequences that we want to classify. In order to do so, before submitting a model, we first measured the quality and performance of the model by using a cross validation technique on the labeled datasets. This method allows us to captures the highest possible accuracy of any Kernel Method such as Kernel Support Vector Machine, also it is efficient since it is very fast to implement. However, there is an optional training feature matrices of size 2000×64 and a test feature matrices of size 1000×64. These feature are based on bag of words representation. obviously, our accuracy from this optional dataset is not high as we want. Hence we have decided to create our own embedding. The training label was initialy 0 for negative or 1 for positive. Hence we have converted this label into -1 and 1 for simplicity since it will make the SVM model works better. But for Ridge Logistic regression we kept initial label 0 and 1.

3 Kernel Architecture & Methods

Kernels that operate on vectors: In Biology, applying directly the step of vectorization can lead of loses information, hence to capture more patterns, one can use the idea of Kernels. We have implemented linear Kernel which $K(x, y) = x^T y$, polynomial Kernel of degree d, $K(x, y) = (x^T y + 1)^d$ and radial basis function Kernel $K(x, y) = \exp\left[-\frac{||x-y||^2}{2\sigma^2}\right]$. These Kernels dealt with vectorized version of sequences of DNA provided in the data challenge.
Spectral Embedding: k-mer size of n means that the sequence of DNA were break up into n sub sequences for $n \in \mathbb{N}$. Spectral embeddings are the embeddings based on index of the one hot vector. The large size of k-mer may lead to a high accuracy but computationally expensive. However we reached our highest accuracy by using spectral embedding process for k-mer size of 10.

Kernel SVM

For this data challenge, we have solved this optimization probelm for Kernel SVM by using cvxopt library.

$$\arg\max_{\alpha\in\mathbb{R}^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j \Phi\left(x_i\right)^T \Phi\left(x_j\right) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j K\left(x_i, x_j\right)$$

under some constraints:

$$\begin{cases} 0 \leq \alpha_i \leq C, \text{ for i} = 1, \ldots, \text{n} \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

Also, we have implimented the Kenerl Ridge Logistic regression.

$$\arg\min_{f\in H} \frac{1}{n}\sum_{i=1}^{n} l\left(y_i - f\left(x_i\right)\right)^2 + C\|f\|_H^2$$

Parameter tunning : We were able to efficiently tune our parameters using Optuna library. As Optuna is an optimization framework that allows us to tune better, thus we found the best hyperparameters for our Kernel SVM.

4 Results

We did an implementation SVM with Linear, Polynomial and Radial Basis Kernel by using the vectorized data that we mentioned earlier. After tuning with Optuna, an SVM with a linear Kernel gave an accuracy of 99.80% of public leaderboar, which obviously outperforms the Ridge Logistic regression (93.20%) and the SVM with the other remaining Kernel available in our implementation but in the private leadboard is becoming 99.40%. However it is still a very good accuracy but a little bit less precision.

5 Conclusion

In this data challenge, we have implemented Kernel methods for the DNA classification task . The overall accuracy is 99.40% witch ii the sixth place in the private leaderboard. Certainly, this score could be enhanced by doing more hyper-parameter tuning and cross validation. For example, we did not give to many possible choice of kmer size as much as we could have, as it was a bit computationally expansive. Also, we should have tried to implement other alternative such as Gaussian Kernel, Sigmoid Kernel, Laplace RBF Kernel and so much more. Our source code is done from scratch and available on Github[1]

References

[1] Vasileios Apostolidis-Afentoulis *SVM Classification with Linear and RBF kernels* 2015.

[2] H. Fizazi Izabatene et al. *Contribution of Kernels on the SVM Performance* 2010.

[1] `https://github.com/FenosoaRandrianjatovo/DNA-sequence-classification-Kernel-methods`