KERNEL METHODS PROJECT		
Student's Names		Deadline: 30.07.23, 23:59 PM
Binta Sow & Uriel Nguefack Yefou	AIMS African Institute for Mathematical Sciences SENEGAL	**July 29, 2023**
Kaggle Team Name: **The Hunters**		Ac. Year: 2022 - 2023
Lecturer & TA: "**Jean-Philippe Vert & Juliette Marrie**"		

Report Kernel Methods project: DNA Sequence Classification(Public score:68.13%, Private score:66.46%)

1 Introduction

The goal of this data challenge is to use kernel methods for a data classification task on DNA sequence data. We want to predict whether a DNA sequence is binding site to a specific transcription factor. For the training, we are giving 3 datasets of 2000 sequences each and the idea is to use the models designed to make predictions on three other datasets of 1000 sequences each, the predictions obtained will then be combined to have one prediction. This brief report summarises our efforts, with a focus on parameter optimisation and classification. Our best model was obtained using multiple Kernel mistmatch and gave us a public score of 68.13% and private score of 66.46%. In the next sections, we will first present the methodology we haved used going from the data preprocessing to the models designed, after that we will present the results obtained for each of the models.

2 Methodology

- Data preprocessing:
  - The label data y was given as either 1 if the sequence was identified as binded, and 0 otherwise. But with the kernel algorithms we used, we decided to work with a binary class of label -1 and 1. So the y label was converted to label -1 and 1.
  - The DNA fragment sequence were break up into subsequences using k-mer. The size of the k-mer(kmer_size) was found using Hyper parameter tuning techniques that we will present in the next part.
- Hyper parameter tuning:
  - **Optuna**: Optuna is an automatic hyperparameter optimization software framework, particularly designed for machine learning. The goal is to find out the optimal set of hyperparameter values for the models through multiple trials while maximizing the evaluation metric which in our case is the accuracy.
  - **Crossvalidation**: Combined with Optuna, crossvalidation is a technique in which we train our model using the subset of the data-set and then evaluate using the complementary subset of the data-set. We found the best value of the crossvalidation parameter k using Optuna.
- Models
  - **Kernel SVM**: Kernel SVM provides a powerful and flexible approach for handling non-linearly separable data. SVM problem:

$$\min_{\alpha,\xi} \frac{1}{n}\sum_{i=1}^{n}\xi_i + C\alpha^T K\alpha$$

st

$$\xi_i \leq 0 \quad \text{and} \quad y_i(K\alpha)_i + \xi_i - 1 \leq 0 \quad \text{for } i = 1,\ldots,n$$

, where $K_{ij} = K(x_i, x_j)$

  - **Kernel Ridge Regression**: is a regularized extension of linear regression that uses the kernel trick to handle non-linearly separable data:

$$\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}(y_i - f_\beta(x_i))^2 + \lambda\|\beta\|_2^2$$

Here, $f_\beta(x) = \beta^T\Phi(x)$ and $K(x, x')$ is the corresponding kernel function derived from Φ.

– **Kernel Logistic Regression**: is the following optimization problem:

$$\arg\min_{f\in H}\frac{1}{n}\sum_{i=1}^{n}\ln(1+e^{-y_i f(x_i)})+\lambda\|f\|_H^2$$

where x_i are the data, $y_i \in \{-1, 1\}$ are the labels, H is RKHS of some p.d. kernel.

– **Kernel Spectrum with SVM**[1]: This method extract all sub-sequences of a particular length k from sequences. As a result, two sequences are compared by counting the number of sub-sequences they share. If x and x' are two DNA sequences, and $\phi_u(x)$ is the number of occurrences of sub-sequence u in x, then the spectrum kernel assessed in x and x' is:

$$K_{SP}(x,x') = \sum_u \phi_k(x)\phi_k(x') = \langle \phi_k(x), \phi_k(x')\rangle_u$$

– **Kernel Mismatch with SVM** [2]: Compare sub-sequences with less mismatches than or equal to a given parameter m. If x and x' are two DNA sequences, and $\phi_{u,m}(x)$ is the number of occurrences of sub-sequences in x with at most m mismatches from u, then the mismatch kernel assessed in x and x' is given by:

$$K_{MS}(x,x') = \sum_{u\in A^k} \phi_k(x)\phi_k(x') = \langle \phi_k(x), \phi_k(x')\rangle_{u,m}$$

where $A = \{'A','T','C','G'\}$, and $\phi_{(k)}(x) = (\phi_u(x))_{u\in A}$.

3 Results

The models presented in Section 2 have been used to make predictions on our 3 datasets of 1000 sequences each, the results obtained are presented in Table 1.

Models	Public leaderboard accuracy(%)	Private leaderboard accuracy(%)
KernelSVM + Optuna + Crossvalidation	66.133	64.666
Kernel Ridge Regression + Optuna + Crossvalidation	54.66	52.66
Kernel Logistic Regression + Optuna + Crossvalidation	53.53	51.86
Kernel Spectrum with SVM	63	62.20
Kernel Mismatch with SVM	**68.13**	**66.46**

Table 1: Performance of our different models on the leaderboard

From the table above, we can observe that we have tried several models. After careful consideration, we decided to focus on the Kernel Mismatch with SVM, which yielded the best results for classifying DNA. The scores obtained on both the public leaderboard and private leaderboard were higher compared to the other models. The SVM method is employed to optimize the SVM objective function and determine the optimal weights for the given input data and kernel matrix with (lambd=1). The Kernel Mismatch is trained on each of the three datasets. Our most successful submission consisted of the sum of multiple Kernels Mismatch , specifically for $(k, m) \in$ (12, 2), (13, 2), (15, 3). This submission achieved a score of 66.46% on the private leaderboard and 68.13% on the public leaderboard.

4 Conclusion

The goal of this task, which consisted of employing kernel methods for DNA sequence categorization, was to design and test multiple kernels from scratch. After experimenting with several kernels, the powerful kernel mismatch paired with SVM excelled other approaches in terms of outcomes. The final accuracy achieved on the private leaderboard was 66.46% which could be further improved with more hyperparameter tuning on the value of lambda, we choose to randomly select the value of "lambda" due to the time-consuming nature of training the kernel mismatch, which would slow down the cross-validation process. Overall, the challenge was exciting and provided a great opportunity for our team to gain hands-on experience with coding kernel methods from scratch.

References

[1] Christina Leslie, Eleazar Eskin, and William Stafford Noble. The spectrum kernel: A string kernel for svm protein classification. In *Biocomputing 2002*, pages 564–575. World Scientific, 2001.

[2] Christina Leslie, Eleazar Eskin, Jason Weston, and William Stafford Noble. Mismatch string kernels for svm protein classification. *Advances in neural information processing systems*, pages 1441–1448, 2003.