Kernel Methods Challenge
African Masters of Machine Intelligence

Elie N. Mulamba, Regis Konan Marcel Djaha

Abstract

The data challenge represents the pinnacle of our exploration of kernel methods in the machine learning course at AMMI. Our primary objective was to implement and comprehend machine learning algorithms while adapting them to handle structural data, specifically Protein sequences data. Through this report, we outline our approach to the Kaggle-hosted challenge, where we were tasked with predicting the binding potential of DNA sequence regions to specific transcription factors. Our most successful model achieved an impressive 6th place on the private leaderboard, boasting a score of 67.00%. This accomplishment highlights our proficiency in the domain of sequence classification and demonstrates the practical application of machine learning in genomics research.

Keywords: machine learning, kernel methods, DNA sequence, sequence classification,

1. Introduction

Transcription factors play a critical role in regulating gene expression by binding to specific sequences in the genome. These binding events can either activate or repress the transcription of target genes. To study these interactions, experimental techniques are used to create genome-wide protein-DNA binding maps. These maps help classify the entire genome into two classes: bound or unbound, based on the presence or absence of the TF of interest.

The information obtained from these binding maps is crucial for understanding the regulatory mechanisms underlying gene expression. By identifying the specific genomic regions where a TF binds, researchers can gain insights into the genes and pathways that are regulated by that TF. This knowledge can further aid in understanding various biological processes and diseases.

2. Objetives

The main objective is to build separate predictive models for each dataset using the training data and labels, and then predict the labels for the test sequences.

3. Data description

The data challenge presents three distinct datasets (k=0, k=1, and k=2) containing training sequences and binary labels for each dataset. The training sequences, stored in Xtrk.csv files, consist of 2000 sequences represented by rows, and Xtek.csv contains 1000 test sequences with Id starting at 1000k. Additionally, the challenge includes test sequences (Xtek.csv) for each dataset, comprising 1000 sequences with unique IDs starting from 1000k.

4. Experiments

4.1. Algorithms

- Hard margin SVM : We experimented with the Hard-margin SVM algorithm discussed in class on our data. After submission, the performance on the public leaderboard was 54.61%, and on the private leaderboard was 54.61%.
- Soft margin SVM: We then experimented with soft margin and obtained the following results for the public and private leaderboards respectively 53.73% and 55.46%

Not satisfied with these results obtained with algorithmic methods, we decided to explore some kernel methods in an attempt to classify them better.

4.2. Kernels

During our experimentation, we explored various kernels, most of which were the ones introduced during our class. The following section will provide definitions for all the kernels we tested.

- Kernel Logistic Regression (KLR), where the type of kernels (linear, polynomial, and the Gaussian radial basis function/rbf), sigma, the regularizer and the degree were considered as the main hyperparameter, with this kernel we got 53.46% on public leaderboard and 53.73 % on private leadernoard.
- Kernel Ridge Regression (KRR): We've been experimenting with kernel ridge regression, which is very efficient

[*] Team: Overfitting

Email addresses: emulamba@aimsammi.org (Elie N. Mulamba), rkmdjaha@aimsammi.org (Regis Konan Marcel Djaha)

compared with algorithmic methods. The results we've obtained on public leaderboard 63,8% and 62.9% on private leaderboard.

- Multiple Spectrum Kernel : Applying a mismatch penalty to compare two sequences. the score obtained on the public leaderboard is 68.13% and on the private leaderboard 67.00%.

We explored a couple of kernels like Gaussian, String kernel, and mismatch kernel but were unable to have results on them. especially string mismatch kernel is very computationally intensive.

The study focuses on binary classification of sequential data using support vector machines (SVM) with multiple spectrum kernels. The k-spectrum kernel, considering all possible k-mers in sequences, is tested on both numerical data representations and the original sequences, achieving up to 67% accuracy. The SVM maps input sequences into a high-dimensional vector space, finding a linear decision boundary to classify them as positive or negative. By combining three k-spectrum kernels with different k values (12, 13, and 15) into mismatch kernels, the authors optimize the weights, allowing them to capture information about sub-strings of various lengths, which proves more effective than single-length sub-strings. This approach presents a promising option for binary classification of sequential data and enhances the predictive capability of the model.

5. Results and discussion

Our primary objective was to predict whether a DNA sequence region is the binding site to a specific transcription factor or not. Inasmuch as several simple strategies were tried at the beginning which made us gain a better understanding of the problem at hand.

The following table summarizes the different results obtained after our experiment.

Table 1: Table of results.

Model	Public score (%)	Private score (%)
Hard-margin SVM	54.61	54.61
Soft-Margin SVM	53.73	55.46
KLR	53.46	53.73
KRR	63.80	62.90
MKL	**68.13**	**67.00**

Our best accuracy score on private score was achieved by implementing multiple kernel learning (MKL) using a simple gradient descent approach. It iteratively optimizes the kernel weights to combine different kernels for better performance.

6. Conclusion

The performance of the kernel SVM with multiple spectrum on our dataset in this challenge presents the classifier as a promising algorithm that is effective in predicting whether a DNA sequence region is a binding site for a specific transcription factor or not. Our final result was an accuracy of 67.00% in the private leaderboard, demonstrating the model's ability to make accurate predictions.

One of the significant advantages of using multiple spectrum kernels is their ability to capture different features and patterns present in the DNA sequences. By combining the k-spectrum and (k, m)-mismatch embeddings, we leverage the strengths of both representations, leading to a more robust and expressive model.

For future work beyond the challenge, we propose to conduct extensive hyperparameter tuning to further improve the performance of the model. Additionally, we will explore other strategies, such as nested cross-validation, to ensure the robustness of our hyperparameter selection process.

In conclusion, the kernel SVM with multiple spectrum has demonstrated promising results in this DNA sequence binding site prediction challenge. With further optimization and fine-tuning of hyperparameters, we believe this approach can achieve even better accuracy and generalization to new data, making it a valuable tool in computational biology and genomics research.

References

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

[2] Juliette Marrie. Kernel methods ammi-2023, 2023. URL `https://kaggle.com/competitions/kernel-methods-ammi-2023`.

[3] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[4] Daniel Usvyat, Narayan Jayaram, and Andrew C. R. Martin. Evaluating tools for transcription factor binding site prediction. *Briefings in Bioinformatics*, 17(2):368–381, 2016. doi: 10.1093/bib/bbv081.