Kernel Methods for Machine Learning.
Team: We don't know what we don't know

Solafa Fadlallah **Ndeye Ngone Gueye**

July 30, 2023

1 Abstract

In this data challenge, we investigated DNA binary classification task. The task involves predicting whether a particular region of a DNA sequence is a binding site for a specific transcription factor(TF). This task is based on string sequences, where we were given a DNA sequence of 100 nucleotides and need to determine if it is a binding site for the TF.

To measure the performance of our classifier, we used the accuracy metric, which provides an indication of how well our model is able to correctly classify the binding sites.

2 Classifiers

For the binary classification task, we used two different algorithms: Kernel logistic regression and Kernel ridge regression.

2.1 Kernel Logistic Regression

In this method, the aim is to minimize the logistic loss:

$$\underset{f\in\mathcal{H}}{min}\frac{1}{n}\sum_{i=1}^{n}ln(1+e^{-y_i f(x_i)})+\frac{\lambda}{2}\|f\|^2 \tag{1}$$

By the representer theorem, any solution of the Kernel logistic regression can be expanded as

$$\hat{f}(x)=\sum_{i=1}^{n}\alpha_i K(x_i,x) \tag{2}$$

to find α we solve

$$\underset{\alpha\in\mathcal{R}^n}{min}\frac{1}{n}\sum_{i=1}^{n}ln(1+e_i^{-y_i[K\alpha]})+\frac{\lambda}{2}\alpha^T K\alpha \tag{3}$$

2.2 Kernel Ridge Regression

Kernel ridge regression is obtained by regularizing the mean square error criterion by the RKHS norm in which the solution is

$$\alpha=(K+\lambda nI)^{-1} \tag{4}$$

3 Mismatch Kernel

These kernels assess the similarity between sequences by considering shared occurrences of subsequences of length k, allowing for up to m mismatches. [1]
The (k,m)-mismatch kernel $K_{(k,m)}$ is the inner product in feature space of feature vectors:[2]

$$K_{(k,m)}(x,y)=\langle\Phi_{(k,m)}(x),\Phi_{(k,m)(y)}\rangle \tag{5}$$

4 Methodology

In this challenge, we utilized the training dataset containing DNA sequences along with their respective target values. Initially, we transformed the DNA sequences into feature vectors using the one-hot vector embedding technique. Subsequently, we applied logistic regression using these feature vectors as input. Despite our expectations of good performance, the model yielded an accuracy of only 57%.

Afterwards, we employed Kernel ridge regression with the mismatch kernel, which led to improved outcomes.

5 Results

The best submission scores were obtained using the Kernel ridge regression with the mismatch kernel. The results are as follows

Evaluation	Accuracy
Public Leaderboard	66.1%
Private Leaderboard	66.6%

6 Conclusion

Given more time, we would have explored and achieved improved results by utilizing various types of kernels suitable for sequence data, such as spectrum kernels and substring kernels. Additionally, we could have experimented with the best representation or a weighted sum of these representations. Furthermore, employing different algorithms like SVM with cross-validation techniques to fine-tune the parameters could have been beneficial. However, in this study, the mismatch kernel, after normalization, outperformed the one-hot vector method. The final results indicate that the model achieved a similar score on the entire test dataset with a slight overfitting.

References

[1] Eleazar Eskin, Jason Weston, William Noble, and Christina Leslie. Mismatch string kernels for svm protein classification. *Advances in neural information processing systems*, 15, 2002.

[2] 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.