African Masters in Machine Intelligence

Kernel Method Kaggle Challenge Report

Jean N'dah & Deborah D. Kanubala

June 1, 2020

1 Introduction

The objective of the Kaggle challenge was to implement machine learning algorithms from scratch and apply them to structural data. The task was a sequence classification problem which involved predicting if the DNA sequence region was binding site to a specific transcription factor.

2 Data Pre-processing

In creating our vocabulary of all sub-sequences in our data-set we tried different sub-sequences sizes ranging from 1 to 10. The vocabulary was created by combining both the train and test to avoid out-of-vocabulary issues but this did not have any impact in our data split. Our vocabulary was in the form of a python dictionary where which sub-sequence found was matched with a unique number. A snippet of this using a sub-sequence size of 5 is shown below:

```
{'GAGGG': 0,
 'AGGGG': 1,
 'GGGGC': 2,
 'GGGCT': 3,
 'GGCTG': 4,
 'GCTGG': 5,
 'CTGGG': 6,
 'TGGGG': 7,
 'GGGGA': 8,
 'GGGAG': 9,
 'GGAGG': 10,
 'GGGGG': 11,
 'CTGGC': 12,
 'TGGCC': 13,
 'GGCCC': 14,
 'GCCCA': 15,
```

Once this was completed, we worked on generating the one-hot vector for the data.

The original training data set was increased from $2000 \times [1+1]$ to $2000 \times [4096+1]$. The plus one included was for the label column. Details on how the one hot vector was done can be found in the function "transform1" in the class "Dataloader".

3 Model Development

We employed the use of the soft support vector machine (SVM) and the logistic regression in the development of our model. For the logistic regression, we used K-fold cross validation to find the best parameters. Using 80% of our data (training data) and a sub-sequence size of 6, the logistic regression model gave us an accuracy of 83.69%. Whiles on the validation data-set, the model gave an accuracy of 67%.
For the soft SVM, we used a combination of a Gaussian kernel and polynomial kernel (multiplication of 2 kernels) with a sub sequence size of 9. Our choice of polynomial degree was 2 with a $C = 10$, standard deviation of 35. The soft SVM gave us an accuracy of 99% on training and 67% on the validation. However, this model gave us a score of 66.6% on Kaggle public leader board and is our best score at the moment.

4 Conclusion / Lessons learnt

We observed from the work that the logistic regression was unable to handle big sub-sequence size. For instance, when we tried with a sub-sequence size of 8, we encountered a problem of unable to allocate memory to our array shape. This problem is ,however, resolved with any sub-sequence size less than 7. Also with a high sub-sequence size the accuracy of the SVM tends to drop. Additionally, we observe the effect of changing values such as the constant C and sigma had on the overall performance of the SVM model.
We also suspect that the SVM model probably is over-fitting, given that we had a close to perfect accuracy on the training data set and this accuracy decrease almost by 50% on the test and validation.
Ultimately, for this task, the SVM was the best model for this classification problem. However, results from the private leader-board showed the SVM model with sub-sequence size of 10 with a polynomial kernel was our best model with an accuracy of 64.4%. Finally, from the results of the competition, we agree and have learnt that "a wrong but simple model can work better than a more realistic but more complex model".