Kernel Methods Project 'SafaZe Group'

Safa Mohamed, Zeinab Haroon

1 Background Overview

The goal of the data challenge is to learn how to implement machine learning algorithms from scratch and to acquire a better understanding of these techniques by adapting them to structural data.

- **Project objective:**
  We consider a sequence classification task: predicting whether or not aDNA sequence region is a binding site to a specific transcription factor.

2 Methodology

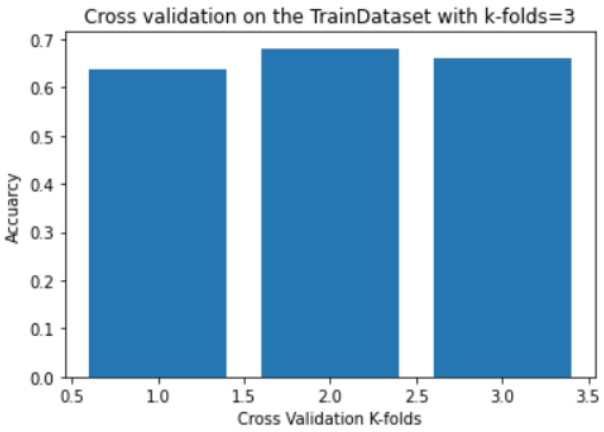
In this section entails the RidgeClassifier and Linear Kernel approaches which were considered during the model training and Parameters search. The following steps were investigated during our work implementing from scratch:

- **Data Preprocessing:**
  The dataset used 2000 training sequences, separated into train, test and the label data (Xtr.csv, Xte.csv and Ytr.csv). The sequences of the training and testing datasets have been extracted and encoded into different indices using 'bag of words transform function' with subsequence of length 7 that extends the input vector dimension into 16384 for both rows of sequences in the train and test-Data followed by data splitting into two-parts (trainset and validation set), in order to evaluate the performance of the model.

- **Architecture:**
  Ridge classifier model has been defined with lamdb regularizer and radial basis function kernel with it is regularizer (sigma).

  Cross-validation parameter and Random Search has later implemented seeking the best parameters (lambd and sigma) accuracy performance that used for model training that ended up with 67 percent score in the validation set.

  Finally, we applied Random search with Cross-validation on the train dataset with (lamdb=3.25e-09, sigma=6.08) and 3 k-fold that improved the mean score of the train data. The scores have shown an equal distribution of the mean score across the data in range (1,6) as illustrated below.



3 Experiment

In our experiment, we were presented with an optional numerical dataset that is similar to the main categorical ones which are; Xtr, Xte and Ytr for sequence training, testing and training labels that we run the Ridge classifier on, while Cross-validation parameter and Random search helped with finding the right parameters for best mean accuracy that found to be 66.79. Followed by the model training on the best parameters which showed a 99.10 train accuracy and 67.12 on Validation.

4 Results

Coming to the end we observed good outstanding scores when applied cross-validation on the full datasets following the same model and best-chosen parameters (sigma and lamdb), The data k-folds = 3 that showed 0.63, 0.68 and 0.66 scores-values, with a mean score and standard deviation around 66.45 and 2.28 respectively. The accuracy on the full train dataset is equal to 100.

5 Conclusion

The observation made upon the previous results demonstrated the better accuracy that can be obtained through random cross-validation splitting of the Full Data as well as training on Ridge classifier and using radial basis function kernel. Through the random search and cv Parameter, we were able to detect the best parameters for better accuracy boosting which later helped with project objective on predicting whether or not a DNA sequence region is a binding site to a specific transcription factor successfully.