Advanced Learning Models Data Challenge Report

Nadia Yende, Amir Chenbeh, Ricardo Huarte

February 17, 2018

Abstract

Here is described how the data challenge for predicting bound of DNA sequence was approached.

1 Introduction

The given data challenge was comprised by three different datasets that described DNA sequences and the goal was to classify them and implement a model that predicts whether they belong to a specific transcription factor.

Different methods were considered and tested like Logistic regression with Stochastic Gradient Descent, Multi-layer Perceptron, Adaboost, and at the end, the one utilized for the challenge was a combination of Kernel method alongside SVM.

2 The Approach

After evaluating a couple of methods that proved to be not useful in the given case we decided to test Spectrum Kernel with SVM. In the modern literature [2] SVM is regarded as a viable option for binary classification of sequential data.

With this approach, the input sequences are mapped into a high-dimension vector space where the feature values provide the coordinates, then the SVM finds a linear decision boundary in the new space and then test weather the sequences are in the positive or negative side of the boundary. The features that we used in the spectrum kernel are the set of all possible subsequences of a fixed length k. If two sequences contain many of the k-length subsequences, their "inner product" computed by the k-spectrum kernel will be large.

We implemented a simple version of Spectrum Kernel. For each sequence we collect the set of k-length subsequences into an array. Now the inner product K(x,y) is computed by using the arrays of the two sequences x and y. This method is not efficient in terms of time consumption ($\mathcal{O}(n * \log n)$, n is the length of input sequences) but that was easy to implement.

We used Python in order to implement SVM with kernel, we made use of the libraries Numpy, Cvxopt and Pandas, thus fulfilling the requirement of no Machine Learning readily available libraries. Cvxopt library was used to solve the quadratic optimization problem using the kernel matrix as the input data.

3 Results

One draw-back of our implementation is that as we used the simple implementation of spectrum kernel, the time to compute all the kernel matrices for the three input and test data is one hour.

In the tests we performed, locally as well as directly in Kaggle, we found that the SVM implementation was providing better results than other approaches we tried, thus achieving a final score of 0.72799.

During submissions we noticed that the score is greatly influenced by the the size k of the spectrum (size of subsequences to consider). We try to tune this parameter by using a grid of values (3, 5, 8, 10). We got the best score for k= 5. We could have get a better score if we used a wide range of k but as we were limited to two submissions a day we wouldn't have got enough time to submit all predictions.

References

[1] John Shawe-Taylor, Nello Cristianini, *Kernel Methods for Pattern Analysis.* The Edinburgh Building, Cambridge cb2 2ru, UK

[2] Christina Leslie, Eleazar Eskin, William Stafford Noble, *The Spectrum Kernel: A String Kernel for SVM Protein Classification.* Department of Computer Science, Columbia University, New York, NY 10027

[3] Vladimir N. Vapnik, *The Nature of Statistical Learning Theory.* Springer New York