DATA CHALLENGE REPORT: PREDICTION OF COVID IN DNA SEQUENCES

Ifeoma Veronica Nwabufo
African Masters in Machine Intelligence
African Institute for Mathematical Sciences, Senegal
inwabufo@aimsammi.org

1 Problem Formulation

The task involved the prediction of COVID virus given anonymized DNA sequences. The dataset consisted of 2000 DNA sequences embedding of training data and 1000 sequences of test data.

2 Preprocessing and Data Exploration

By data exploration, we saw that all data points were roughly on the same scale. The distribution of the classes were also balanced. We had to change the labels to be in the form -1 or 1 to be consistent with the models we were going to implement.

3 Models

We explored logistic regression, kernel linear regression and kernel SVM models.

1. **Logistic Regression**

   This was the baseline model. Here we implemented both the sklearn version (as a sanity check) and logistic regression from scratch. which gave accuracy scores of 95.25% and 87% respectively.

2. **Kernel Linear Regression**

   For the implementation of kernel linear regression we tried the Gaussian Radial Basis Function (rbf) kernel and the polynomial kernel. However, the rbf kernel gave better results over the polynomial kernel so we used the rbf kernel with sigma = 0.1. We however used two different approaches.

   - We checked for correlations amongst features to guard against multicollinearity since the backbone model was linear regression. We removed columns where correlation was greater than $|0.8|$. Specifically we removed columns 38,39, and 42. We got 0.001 as the optimal lambda by searching in a range of values. This model gave the best score on Kaggle with 0.93 on the private leaderboard and a cross-validation mean f1-score of 0.92.
   - In this second approach, we did not drop any correlated columns. Here, we used a lambda value of 0.01. The result from this approach gave 0.92 on Kaggle with a cross-validation mean f1-score of 0.91.

   In general, since the problem was a classification task, we converted the output using np.sign to change the predictions to -1 or 1.

3. **Kernel SVM**

   Here, we used the rbf kernel and searched through a range of values for the optimal C. The optimal C was C=1 with cross-validation mean accuracy score of 87.75.

Model	**Kernel**	σ	λ	**C**	**Score**
Logistic Regression	-	-	0.01	-	87%
Kernel Linear Regression	rbf	0.1	0.01	-	92%
Kernel Linear Regression	rbf	0.1	0.001	-	93%
SVM	rbf	0.1	-	1	87.75%

Table 1: Summery of Results

4 Cross Validation and Hyperparameter Tuning

In all models (except logistic regression), we performed 5-fold cross validation which proved to be useful in prevent overfitting or underfitting. This proved to be helpful as the score on the private and public leaderboard were roughly the same.

We also performed hyperparameter tuning to obtain the optimal C (in the case of SVM) and optimal lambda (in the case of kernel linear regression).

5 Codes

Codes for this implementation can be found here.

https://github.com/Success-Vera/Kaggle-Data-Challenge

6 Conclusion

In this work, we looked at different ways to classify DNA sequences using linear models with the help of kernel functions. This proved to be a very simple way to transform these linear models to powerful non-linear classifiers.