Readme file

Members :- rohit18259
shanu19104
jasdeep19047

What we have tried out:

We have tried out many models for classification in this assignment. Some of the notable ones are Naive Bayes, Categorical Naive Bayes, Support Vector Machines, Logistic Regression, Decision Trees, Random Forest, k neighbors classifier, multilayer perceptron classifier, XGBclassifier. Finally we used a Bagging classifier on top of all these base classifiers to increase their accuracy.

At first, we saw that the length of the amino acid sequences were variable in nature, and that the longest amino acid sequence had length of 25. So, we filled all the remaining characters of amino acid sequences with length less than 25 with the character ‘A’. Then we converted the original dataset to a 1 Hot encoded dataframe. In this dataframe, we had 650 columns (25*26 = 650).
Then we applied all the classifiers mentioned above (excluding bagging classifiers). But we saw that we couldn't get more than 76% accuracy using these classifiers and the features. After that, we tried to use a bagging classifier on top of all these classifiers as bagging is a method of ensembling and is known to increase accuracy. After applying bagging, we increased our accuracy to 77%. Then we tried to use lots of combinations using these features and these classifiers, but we couldn't get much better result.

As we saw that using different models is not giving much better results, we felt that we might need to change the features to achieve better results. We learned of various composition features corresponding to amino acid sequences such as length, count_amino_acids, molecular_weight, aromaticity, isoelectric_point, charge_at_pH, etc. We used the library Bio.SeqUtils.ProtParam to extract these features from the amino acid sequences. Once we did that, we got a total of 53 features in our dataframe. After that, we trained a random forest classifier on this dataframe. With this model, we were able to increase our accuracy to above 78%. Then we tried to improve our model by stacking a bagging classifier on top of the random forest classifier, with the number of estimators equal to 50 in both of them to fine tune them. With this, we were finally able to achieve an accuracy of 79.1% in our public score (which had an accuracy of 78.5% in the private score).

Techniques we used :

- Random forest classifier :- It is an ensemble method for decision trees where we combine multiple decision trees to give a better prediction answer.

- Bagging :- It is a method of ensembling which is used to improve the stability and accuracy of machine learning algorithms . In this assignment, we have used the inbuilt bagging classifier in python for this purpose.
- Amino acid composition features :- We have used the library Bio.SeqUtils.ProtParam to extract various composition features of the amino acid sequences such as length, count_amino_acids, molecular_weight, aromaticity, etc.
- predict_proba :- This is a function in sklearn which gives us the probability values for the target instead of binary values such as 0 or 1. It increases the accuracy of our test data results.