Assignment 1 - Covid ML Project

Anirudh Kalla

October 13, 2020

1 Data Extraction

The initial stage of any Machine Learning project is data extraction from reliable sources. In this project, the data we used was taken from the ChEMBL database (ChEMBL ID: 5118), *Replicase Protein 1ab.*

The extraction was carried out using the ChEMBL Web Resource Client[1].

2 Data Processing

The next part in the natural progression of the project was data processing. Having appropriate data tags and descriptors is extremely essential to the success of any ML application project, so a wide array of data cleaning techniques were used to ensure that the data we had was scientifically meaningful and would add relevant information to the model. The data was delimited using appropriate delimiters and converted to a .csv file for further insight. The .csv data was converted to a DataFrame and all the NaN values were dropped to prevent GIGO (Garbage In Garbage Out).

3 Data Analysis

Using RDKit, four of the most relevant decriptors were extracted by cross-referencing SMILES (Simplified Molecular-Input Line-Entry System) data present in our dataset. The descriptors used were as follows:-

1. Minimum Partial Charge
2. Maximum Partial Charge
3. Molecule Log P
4. Exact Molecular Weight

To prevent further errors, the NaN values were dropped again from the dataframe containing these extracted descriptors.
Now, since multi descriptor regression was performed, the aforementioned descriptors were scaled appropriately to prevent anomalous distribution of the weights learnt by the regression model. The descriptors were converted to a distribution from $0 - 1$ for further use. An indicative sketch of Vanilla Linear Regreesion is given below for representative purposes.
Note: Linear regression was not used for regression due to low accuracy. Other regression models were used, and are discussed in the section below.

$$y = W_1x_1 + W_2x_2 + W_3x_3 + W_4x_4 + ... + W_nx_n \tag{1}$$

Here,x_i are the descriptor values and W_i are the weights learnt by the model.

4 Data Visualization

To get an overall picture of the data description with respect of each of the descriptors and a general heatmap of correlation, matplotlib and seaborn were used. The following plots were used generated and provided deep insights into potential model building techniques.

1. Histogram PairPlot
2. Regression PairPlot
3. Kernel Density Pairplot
4. Correlation Heatmap

5 Model Building

To form an input vector of all the descriptors, the relevant (Normalized) dataframe columns were converted into a numpy array. Similarly the target column (BioActivity) was converted into another numpy array.
Post this, the Train and Test data was split in an appropriate ratio with optimum $random_state$ parameter, so as to have a coherent regression model. A wide assortment of Regression models were used maximize the regression score. The following models were tried:-

1. Linear Regression
2. Logistic Regression
3. Decision Tree Regressor
4. K Neighbors Regressor
5. Linear Discriminant Analysis
6. Gaussian N B
7. Support Vector Regression
8. Random Forest Regressor
9. Stochastic Gradient Descent Regressor
10. Gaussian Process Regressor (Dot Product: White Kernal)
11. Decision Tree Regressor
12. Gradient Boosting Regressor
13. Voting Regressor
14. Multi Output Regressor
15. Ridge Regressor

Out of these, the best results were obtained using Random Forest Regressor, Gradient Boosting Regressor and K Neighbor Regressor. They were then passed to a Voting Regressor to weigh out their negatives and deliver stronger and more robust results.
Note: To find the most appropriate $random_state$ values, the regressors were run on a loop with 100 different values of $random_state$ and after evaluation of the score after each iteration, the best $random_state$ was chosen.

6 Conclusion

The project gave a very deep understanding of many concepts, from Scientific Data Analysis to Experimental Biology. The following conclusions can be drawn from the Project :-

1. Data Cleaning is one of the most essential parts of any ML project. Having scientifically meaningful data and understanding the purpose of the data, along with its properties and labels is the cornerstone of any successful ML model.

2. Having a thorough knowledge of the subject in which the model will be deployed is imperative. In this case, understanding that bioactivity has many methods of measurement was very important. $IC50^2$ is one of these measuring methods for bioactivity. In this project, we have considered IC50 as our target vector as it had the most coherent data points.

3. Data Visualization is also a very good practice in ML models. Visualizing the data that we have, along with all all it's paramters is very important as it can give fundamental insights as to how the data is distributed, helping make informed choices about the further steps that have to be taken during model selection and deployment. Knowing how the data points correlate with each other helps select the most appropriate input/target/feature vectors for model building.

4. Last but not the least, it is very important to be alert and open to new ideas, as all the problem that come up during an ML project have most likely been addressed in some way, shape or form on online discussion forums like :-

   (a) github.com

   (b) stackoverflow.com

   (c) datascience.stackexchange.com

   (d) Kaggle

   (e) Medium (Towards Data Science)
   These platforms help out a great deal and can be phenomenal sources of learning if one looks carefully enough

7 Appendix

[1]: ChEMBL Web Resource Client: The client handles interaction with the HTTPS protocol and caches all results in the local file system for faster retrieval. Abstracting away all network-related tasks, the client provides the end user with a convenient interface, giving the impression of working with a local resource. Design is based on the Django QuerySet interface.

[2]: Inhibition Constant50: Time it takes to reduce the bioactivity of the sample assay to 0.5 of it's original value