Project Assignment 1

CoViD Data Analysis using Machine Learning

Group:

Kuldeep Singh ; 16235 (BS-MS 5th year)
Nitin Kriplani ; 16129 (BS-MS 5th year)

Problem:

The objective of this project is to build a Machine Learning model which can predict the 'BioActivity' values of chemicals. These bio-activity values can be considered as the effectiveness of the drug against SARS-CoronaVirus.

To build the model, we have the data containing information of different drugs which can replicate the effects of protein. Obtained from ***ChEMBL Database*** .
This data has the chemical formula of each drug in "SMILES strings". The problem is to find out different properties of these drugs using these strings. And to calculate the bio-activity values of the drugs using RdKit descriptor function. Then compare these to actual bio-activity values to build our ML Model.

Approach / Strategy and Procedure :

- Retrieved data from ChEMBL website. Using ChEMBL API
- Data cleaning
- Selected SARS coronavirus 3C-like proteinase Bioactivity Data
- Created logarithmic pIC50 from IC50 after normalizing IC50. Logarithmic method was having a better r2 score.
- Preprocessing

    - Labeled compounds as active, inactive or intermediate according to their bioactivity values in IC50 units, to compare active and inactive separately.
    - Downloaded required libraries to create descriptors.
    - Four descriptors were created using rdkit. Out of which only three could be categorized differently using Mann-Whitney U test.
    - While using only these three descriptors, r-score observed was too low. Around 0.1-0.2. So later used bash script to create more descriptors.

- Trained a few models and chose the best performing one. (BayesianRidge method had the highest R2 score.)

Observations:

1. Number of training data is 133, which is too low to perform regression learning. Other coronavirus data-sets are also low in amount.
2. Descriptors created using rdkit had very low correlation, max was 0.35.
3. Three descriptors clearly show the difference between active and inactive and hence used as predictors.
4. Certain models are bad for small datasets like lassolars,RandomForestGenerator. While the BayesianRidge model worked a lot better than others.

Conclusions:

1. Due to low number of Data points, regression model training gave below average results. The BayesianRidge model gave a R2 score of 0.31 to 0.33.

2. More and better descriptors required for better accuracy of the model.
3. Available models gave R2 score as follows:
    - SVR = 0.2464
    - SGDRegressor = 0.1369
    - BayesianRidge = 0.3175
    - LinearRegression = -2.684e+17
    - RandomForestRegressor = 0.005313
4. Although the model is not robust, it can be said that given enough data set it can be improved to a great extent.