PREDICTION OF A BIOLOGICAL RESPONSE OF A MOLECULE

CSU 4032 Data Mining Assignment
Department of Computer Science and Engineering
National Institute of Technology Calicut

ABSTRACT

Problem Definition:

Our objective is to build a prediction model which can relate molecular information to an actual biological response[1]. Each molecule is represented by a number of characteristic features and these are represented as columns in the data set. Each molecule has a corresponding feature vector which is represented as rows in the data set.
We plan to implement a classifier using a supervised learning algorithm which could accurately predict if the molecule evokes the biological response or not.

Training Data Set:

The training data set[2] is in the comma separated values (CSV) format. Each row represents a molecule and there are 3752 such rows. The first column contains experimental data describing an actual biological response (label); the molecule was seen to evoke this response (1), or not (0). The remaining columns represent molecule properties/descriptors (d1 to d1776). These are calculated properties/features of a molecule that can capture some of the characteristics of the molecule - for example size, shape and elemental composition.

Test Data Set:

The test data set[2] is in the same format as the training set except for the first column (label). There are 2501 molecular representations in the test set.

Tools:

Python	(General Purpose Programming Language)
NumPy[3]	(Python Library for Numerical Computing)
SciPy [3]	(Scientific, Mathematical and Engineering Package)
Scikit-Learn[4]	(Python Machine Learning Library)
IPython	(Interactive Python Environment with Debugging support)
PyData[5]	(Python Data Analysis Library)

References:

[1] https://www.kaggle.com/c/bioresponse, Biological Response Problem Page.
[2] https://www.kaggle.com/c/bioresponse/data, Biological Response Data Set.
[3] http://docs.scipy.org/doc/numpy/user/, NumPy User Guide.
[4] http://scikit-learn.org/stable/user_guide.html, Scikit-Learn User Guide.
[5] http://pandas.pydata.org/, PyData Official Documentation.

Submitted By:-

Rahul P	B090195CS
Mohammed Hashir	B090167CS
Assim Ambadi	B090068CS
Delbin Thomas	B090109CS