Machine Learning Assignement

Observation Report

For Credicxo Tech Private Limited.

Report By-

Chirag Goyal

9910570397

goyalchirag98@gmail.com

About the dataset

The dataset has 6598 rows and 170 columns categorizing the elements into MUSK and NON-MUSK.

Approach

1. I found out that out of the 170 columns three of them were not necessary [ID, molecule_name, conformation_name], so I dropped those colums.
2. Out of the remaining 167 columns the column titled class was our target data.
3. The remaing 166 columns had a varying range so I used the MinMax Scaler to scale the data down to a range of -1 to 1 since the values were also negative.
4. I split the data into 80 : 20 ratio (train:test) as directed in the instruction.
5. I performed Logistic Regression on the data and found the accuracy to be not satisfactory enough.

	precision	recall	f1-score	support
0	0.99	0.95	0.97	1162
1	0.72	0.91	0.80	158
accuracy			0.95	1320
macro avg	0.85	0.93	0.89	1320
weighted avg	0.96	0.95	0.95	1320

6. I further split the training data into training and validation data (80:20).
7. The ANN consists of:
    1. 4 dense layers.
    2. 2 dropout layers to encounter overfitting.

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 16)                2672
_________________________________________________________________
dropout_1 (Dropout)          (None, 16)                0
_________________________________________________________________
dense_2 (Dense)              (None, 32)                544
_________________________________________________________________
dropout_2 (Dropout)          (None, 32)                0
_________________________________________________________________
dense_3 (Dense)              (None, 64)                2112
_________________________________________________________________
dense_4 (Dense)              (None, 64)                4160
_________________________________________________________________
dense_5 (Dense)              (None, 2)                 130
=================================================================
Total params: 9,618
Trainable params: 9,618
Non-trainable params: 0
_________________________________________________________________
```

8. Number of epochs = 100, batch size = 64.
9. The model gave an accuracy in the range of 98-99%.

	precision	recall	f1-score	support
0	0.99	0.99	0.99	1119
1	0.97	0.96	0.96	201
micro avg	0.99	0.99	0.99	1320
macro avg	0.98	0.98	0.98	1320
weighted avg	0.99	0.99	0.99	1320
samples avg	0.99	0.99	0.99	1320

Accuracy and Loss curves