Performance Of Logistic Regression On Student Performance Data

Shubham Barudwale
School of Computing Sciences and Engineering, VIT Chennai, Tamilnadu, India 600127
Email: barudwaleshubham.dinesh2015@vit.ac.in

Abstract-

In Logistic Regression we fit the data into N different classes and gives the probabilities of the points towards each classes. Logistic regression uses sigmoid function to give probabilities.

In this experiment we are dividing the points and fitting them into different classes and predicting the probabilities.The accuracy was about 73% for the data.

1. Introduction

In logistic regression we need to restrict the hard threshold of linear classification because linear regression does not use threshold.

After confining the output values we find the probabilistic value of the function using conditional probability. Which gives us the probability of a point belonging to one class.

After this also there can be misclassified points which will increase the probability of having the wrong group then we will apply most likelihood methods and gradient descent to inprove the corret probability by taking the smaller steps.

2. Methadology

Linear classification uses a hard threshold on the signal s = w^{T}x,

$$h(x) = sign(w^T x),$$

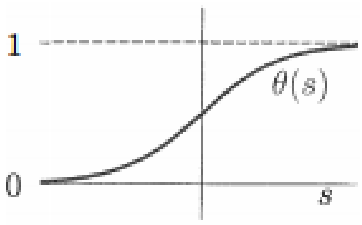
while linear regression uses no threshold at all,

$$h(x)=w^T x$$

In Logistic regression, we need something in between these two cases that smoothly restricts the output to the probability range [O, l]. One choice that accomplishes this goal is the logistic regression model

$$h(x)=\theta(w^T x)$$

where 8 is the so-called logistic function $\theta(s) =e^s/1+e^s$ whose output is between 0 and 1



We are trying to learn the target function $f(x) = P[y = +1 \mid x]$.
The data does not give us the value off explicitly. Rather, it gives us samples generated by this probability. Therefore, the data is in fact generated by a noisy target P(y | x) .

$$p(y|x)=\{f(x) \quad ; \text{ for } y=+1$$
$$\{1-f(x) \quad ; \text{ for } y=-1$$

The standard error measure e(h(x), y) used in logistic regression is based on the notion of likelihood; how 'likely' is it that we would get this output y from the input x if the target distribution P(y I x) was indeed captured by our hypothesis h(x)

$$P(x|y)-\{h(x) \quad ; \text{ for } y=+1$$
$$\{1-h(x) \ ; \text{ for } y=-1$$

We substitute for h(x) by its value B(wTx)

$$P(y \mid x) = \theta(y\ w^T x)$$

After substituting equivalent equations we get

$$E_{in}(w)=(1/N)\Sigma ln(1+e^{ynwTxn})$$

3. Dataset – Student Performance Dataset

Student performance dataset is the dataset from which we are trying to gain some information from 650 student’s academic performance data and various attributes which are affecting the performance. Attributes include sex, age, study time, free time, family information etc.

We will be using the Logistic regerssion method to learn and predict the performance. Out of 650 I have taken 455 data entries for lerning and rest for prediction i.e testing the performance of the algorithm.(70:30). The labels or target functions are mentioned in the ‘new’ field which is added manually. bad(0-7), medium(8-14), good(15-20) .

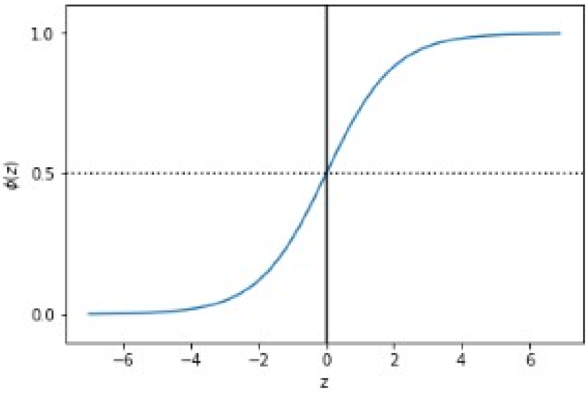
4. Algorithm

```
1: Initialize the weights at time step t = 0 to w(O)
2: for t = 0, 1, 2, . . . do
3: Compute the gradient
        $g_t=-(1/N)\Sigma(y_n x_n/1+^{tnwt(t)xn})$
4: Set the direction to move, Vt = -gt .
5: Update the weights: w(t + 1)= w(t) + TJVt .
6: Iterate to the next step until it is time to stop
7: Return the final weights w.
```

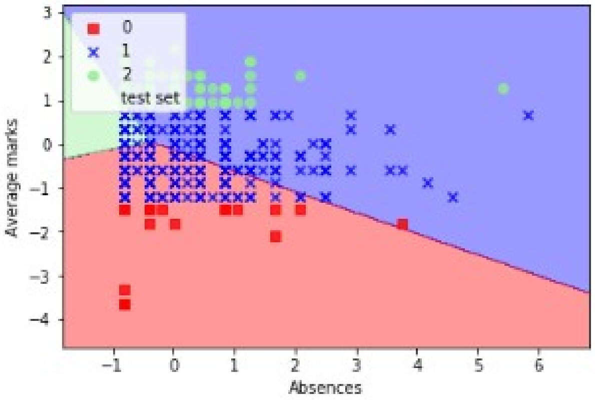
5. Experiments

Initially we have visualized the sigmoid function which is the core part of this experiment. We can visualize the how the Logistic regression gives the probability of each point according to the attributes between 0 and 1 for each class.

Then we took our data and splitted it into two parts as train and test data in ratio 7:3. Then we defined the logistic regression function and fitted the data. Which gave the accuracy of about 73.8 % on test data. Then we visualized the logistic regression by splitting it into three classes. Then we predicted the probabilities of the test data set for each class.



Sigmoid function



Logistic fitting regions

6. Conclusion

We used the Logistic Regression to seperate the Student performance data and classify them according to target function into classes. We can see the three classes wchich shows three different classes. But due to some algorithmic issues all the points were not classified correctly.

In sklearn logistic regression algorithm the accuracy was low as 73.8% which was comparitively lower than other algithms we have seen earlier.

When we see the probabilities of a point towards different classes we can observe that all the probabilities are almost same. Which shows that points are not very well distinguishable.

7. Referances

1)http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

2) Python Machine Learning by Sebastian Raschka

3) Learning From Data, Abu Musafa