Performance of Logistic Regression on fertility dataset

Shrikant patro,

School of Computer Science and Engineering, VIT Chennai, Tamil Nadu, India 600127

Email:shrikantjagannath.2018@vitstudent.ac.in

Abstract – The classification of the data point into binary or multiple classes is the very common task in machine learning. Perceptron learning algorithm and pocket algorithm are basic classification algorithm. But, the perceptron learning algorithm faces grave problem that it will never converge in case the data point is not linearly separable. So, to have better linear classification model without major disadvantage logistic regression is introduced .This paper will provide complete analysis of logistic regression in classification task. The dataset used here is woman fertility dataset from UCI repository. Towards end, we will check the improvement achieved by the logistic regression over the perceptron algorithm on the csv file belonging to the woman fertility dataset containing records of patient donated blood and other detail. This model will not only predict the person can be possible blood donor or not but based on input attribute it will give probability or chances of being donor.

Introduction

In Machine learning, Logistic regression is very popular model for classification of the data point for the linearly separable data point. It is basically a binary classifier but the libraries being used in the implementation is robust and support the multiclass classification of the data point. The need for the logistic regression felt because the perceptron learning algorithm which is very basic algorithm for classification faced very serious disadvantage during the classification of collection of non-linearly separable data point. The real life example or data collected from some real world experiment will generate random data point. As per the principle of the perceptron learning algorithm it will stop its classification task once it had successfully classified the data point correctly in two dimensional plane But, the situation as discussed above for the classification of non-linearly separable data point the Perceptron learning algorithm will fail badly. For such cases, the perceptron learning algorithm will never converge and this will lead to running for the infinite duration. To stop this process of classification we need to define the maximum iteration after which system will stop further classification process. The classification achieved by this process is still not good because it may produce good, average or bad result based on the last running instance or last iteration. So, Logistic regression is introduced to solve this problem and also enhance the classification by adding several feature that were missing in the case of basic perceptron learning algorithm.

Methodology

The Probability is the basis of the logistic regression. The odds ratio is given by P/ (1-P), where the P represent the success in the probability. The success can be successful identification of the disease, obtaining head on tossing the coin and so on. For binary classification the output label class label is Y=1 in favor of the event and Y=0 for unfavorable event. Then, the logit function is

than defined on top of the odds ratio. The logarithm of the odd ratio.

$$logit(p) = log \frac{p}{(1-p)}$$

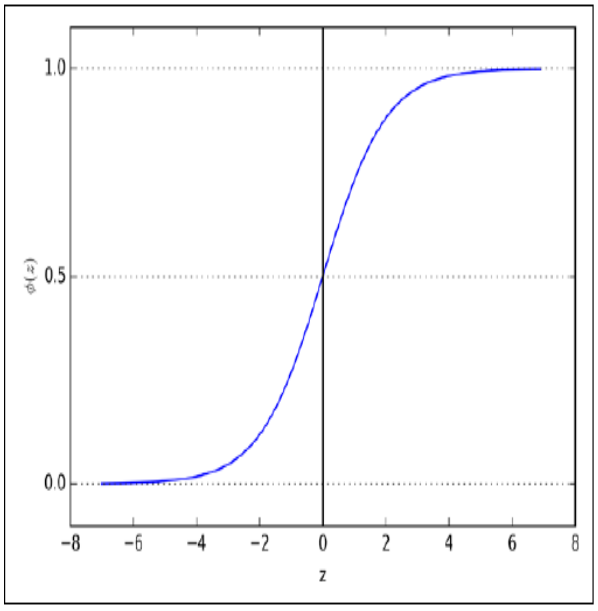
This logit function take as input the value in

range of 0 and 1 and transform them to which we can use to express a linear relationship between feature values and the log-odds:

$$logit\left(p\left(y=1|x\right)\right) = w_0x_0 + w_1x_1 + w_mx_m = \sum_{i=0}^{n} w_mx_m = w^Tx$$

Now, we want to predict the probability that given data point as p (y = 1| x) , to evaluate the inverse of the logit function. This will produce the logistic function also stated as sigmoid function given by

$$\phi(z) = \frac{1}{1+e^{-z}}$$



Here, z is the net input, that is, the linear combination of weights and sample features and can be calculated as

$$z = \boldsymbol{w}^T\boldsymbol{x} = w_0 + w_1x_1 + \cdots + w_mx_m .$$

The sigmoid function appear with some magic that will reduce the number closer to one that is running toward the +ve infinity and the number is reduced close to zero that is closer to the –ve infinity. Thus, we conclude that this sigmoid function takes real number values as input and transforms them to values in the range [0, 1] with an intercept at $\phi(z) = 0.5$.

Now, learning weights in logistic regression is big challenge. Let us first define the likelihood L that we want to maximize, when we built logistic regression model. Assumption individual sample in our dataset is independent of other.

$$L(w) = P(y|x;w) = \prod_{i=1}^{n} P\left(y^{(i)}|x^{(i)};w\right) = \left(\phi\left(z^{(i)}\right)\right)^{y^{(i)}}\left(1-\phi\left(z^{(i)}\right)\right)^{1-y^{(i)}}$$

Taking log over the maximum likelihood function,

$$l(w) = \log L(w) = \sum_{i=1}^{n} \log\left(\phi\left(z^{(i)}\right)\right) + \left(1-y^{(i)}\right)\log\left(1-\phi\left(z^{(i)}\right)\right)$$

Now we could use an optimization algorithm such as gradient ascent to maximize this log-likelihood function. Alternatively, let's rewrite the log-likelihood as a cost function J that can be minimized using gradient descent as

$$J(w) = \sum_{i=1}^{n} -\log\left(\phi\left(z^{(i)}\right)\right) - \left(1-y^{(i)}\right)\log\left(1-\phi\left(z^{(i)}\right)\right)$$

Looking at the preceding equation, we can see that the first term become zero if y=0 , and second term will become zero if y= 1.

$$J\left(\phi\left(z\right), y; \boldsymbol{w}\right) = \begin{cases} -\log\left(\phi\left(z\right)\right) & \text{if } y = 1 \\ -\log\left(1-\phi\left(z\right)\right) & \text{if } y = 0 \end{cases}$$

Database- Woman's Fertility dataset

There are different CSV file for different phenomenon associated with woman fertility dataset. This dataset is collected by WHO with the help of 100 volunteers a semen sample analyzed according to WHO 2010 criteria. Sperm concentration are related to socio-demographic data, environmental factors, health status, and life habits. Dataset characteristics: Multivariate, Attribute characteristics: Real, Associated task: Classification, Regression, Number of instances: 100, Number of attribute: 10, Missing values: NA, Area: Life, Date Donated: 2013-01-17 and number of web hits: 142847.
The further description is specifically regarding the "Blood Transfusion Service Center". This include field like Recency (Months), Frequency (Times), Monetary (c.c. blood), Time (months) and donated. Donated is binary attributed stating in terms of 1 and 0 to represent whether donated or not.

Algorithm

In order to find the regression line we must follow the following steps:

Step 1: Import the pandas, numpy, sklearn , matplotlib Libraray. From sklearn import datasets, linear models logistic regression. From matplotllib import pyplot.

Step 2: Read dataset into pandas dataframe.

Step 3: Split training set and test set using train_test_split module imported from sklearn. Model _selection.

Step 4: Call logistic regression function from linear_model and fit the model using training set.

Step 5: perform prediction using test set.

Step 6: Finally plot the data point and regression using pyplot module in matplotlib.

Conclusion:

The logistic regression was introduced to overcome the disadvantage possessed by the earlier perceptron classification technique. It will stop the finding the best fitting line once it is stable and prevent the system from getting into infinite iteration in case non-linearly separable classification point.

References:

[1] Python Machine Learning – Sebastian Raschka.