S21CSE471: Statistical Methods in AI **Spring 2021**

Logistic Regression

Lecture Date - 28th Jan 2021

Prepared by: Arpit Maheshwari, Krishna Shenvi, Shah Mohammad Kabeeruddin Maaz

1 Introduction

Logistic Regression is a classification technique which is used to predict a discrete outcome based on variables which may be discrete, continuous or mixed. Thus, when the dependent variable has two or more discrete outcomes, logistic regression is a commonly used technique.

Some simple binary examples where we can use Logistic Regression are:-
1. Email classification for predicting an email is spam or not.
2. Tumor Classification for classifying tumor types i.e. malign or benign.
3. To classify genuine and fraud transaction.

2 Why Linear Regression is not fit for classification problems?

The objective of linear regression model is to find the relationship between input variables with dependent continuous target variable. Linear regression predicts target variable as continuous value, which can be any real number ranging from negative infinity to infinity.Linear regression is suitable for regression problems such as predicting the price of a property or predicting the sales etc.

For classification problems, our target variable is divided into two or more classes. So, to use Linear Regression for classification we need to choose a threshold value. If our predicted continuous value is more than the threshold, the data point will be classified in one class and if not then we will classify it in other class.

Lets take an example of purchase of a product. Assume our data we have age of customer which is input variable and purchase on product is target variable. So, we will take purchase as a binary label where 1 denotes the product purchased and 1 denotes product not purchased by customer. If we try to fit a linear regression model on it as shown in Figure 1. The red line seems to be a good fit, considering purchase on Y-axis and Age on X-axis. Now, from this regression line we can predict according to any age value of y.

Considering 0.5 as threshold, if value of y is greater than 0.5 then this will predict customer purchased the product otherwise not. But the problem in linear regression is unbound i.e. the value can range between $(-\infty, \infty)$. So, in this case, regression line can predict a negative value of y as well, and even a very higher positive value which decreases performance of our model and gives high error score. Also, it makes our model highly sensitive to imbalance data.

To fit such a data, logistic regression comes into picture.Logistic regression uses a sigmoid function which maps a sigmoid curve as shown in Figure1 to our purchase data. As we know, probability can range between only [0,1], where the probability of something certainly to happen is 1, and unlikely to happen is 0, which helps in binary classification. Logistic regression is fit for these classification problems because it predicts a probability range between (0,1) i.e. it is bounded in nature. Eventually, logistic regression performs much better than linear regression in classification tasks and gives a low error score.

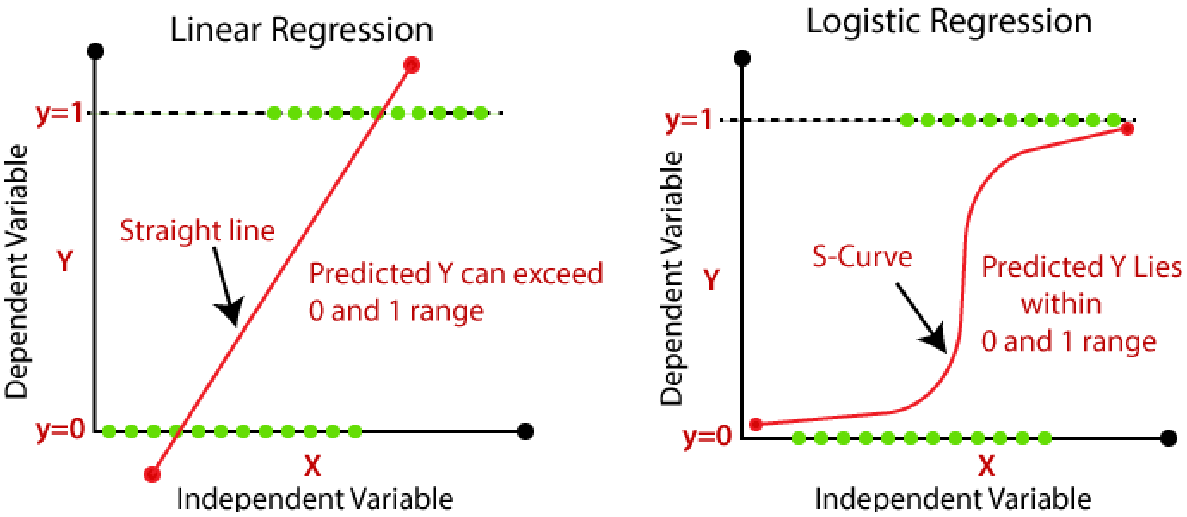

Figure 1: Comparison between Linear and Logistic Regression

3 Some properties of Sigmoid Function

1. The standard logistic function is known as sigmoid function which is defined as

$$S(x) = \frac{1}{(1 + e^{-x})} \tag{1}$$

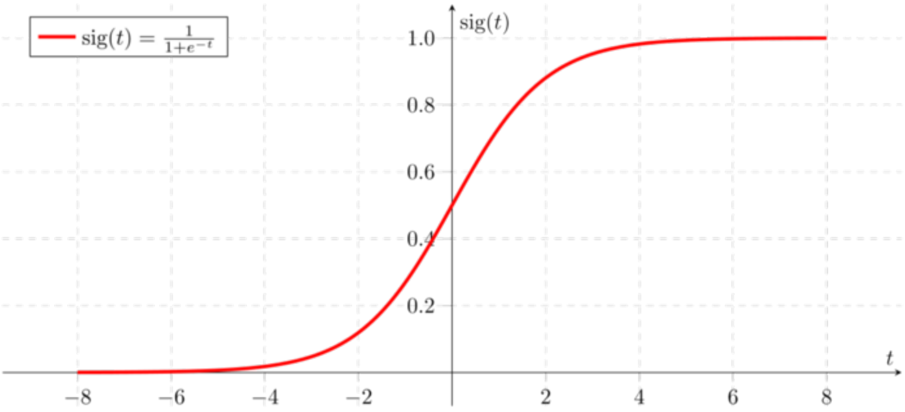

Figure 2: Sigmoid Function

2. As shown in figure, for all values of $x \in (-\infty, \infty)$ sigmoid function follows that $S(x)$ is increasing and bounded with value in $(0, 1)$.

3. We can find derivative of sigmoid function as shown in section 4 as:

$$S'(x) = \frac{e^x}{(1 + e^x)^2} = S(x)(1 - S(x)) \tag{2}$$

4. The sigmoid function satisfies the following properties :

   (a) $S(x) + S(-x) = 1$

   (b) $\lim_{x\to\infty} S(x) = 1$

   (c) $\lim_{x\to 0} S(x) = \frac{1}{2}$

   (d) $\lim_{x\to -\infty} S(x) = 0$

(e) $S'(x) = S(x) * S(-x)$

(f) $S'(x) = S'(-x)$

(g) $\lim_{x \to \pm\infty} S'(x) = 0$

5. Variation in sigmoid function with change in coefficient of x i.e. a.

$$S(x) = \frac{1}{(1 + e^{-ax})} \tag{3}$$

As shown in the graph below as a increases steepness of the curve increases and when a decreases the curve relaxes and steepness decreases.

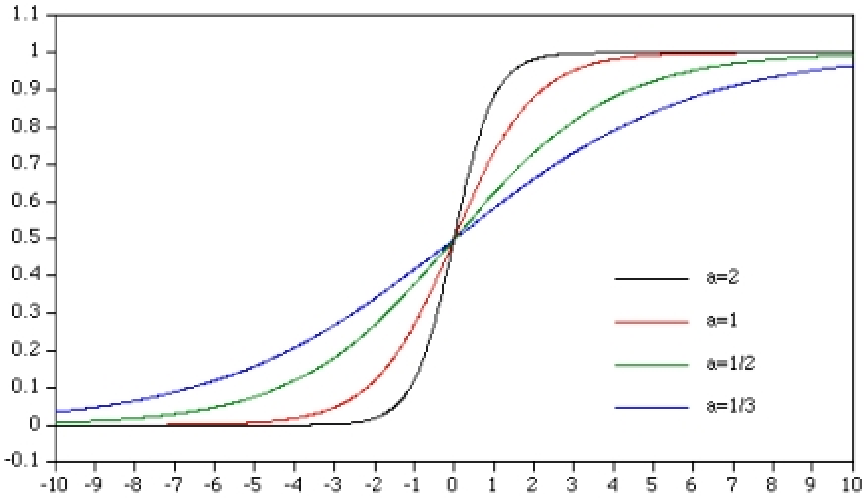

Figure 3: Sigmoid Function

6. Hypothesis function for logistic regression can be defined as:

$$H_\theta(x) = S(\theta_0 + \theta_1 x) \tag{4}$$

$$H_\theta(x) = \frac{1}{\left(1 + e^{-(\theta_0 + \theta_1 x)}\right)} \tag{5}$$

Generalized form of hypothesis with θ and X as vectors can be represented as

$$H_\theta(x) = \frac{1}{\left(1 + e^{-\theta^T X}\right)} \tag{6}$$

This is also equal to probability of y taking on a 1 value for a given parameters X and θ has to be determined.

4 Differentiation of sigmoid function:

We can write equation of the sigmoid function as follows:

$$S(x) = \frac{e^x}{1+e^x}$$

$$= (e^x)(1+e^x)^{-1}$$

By differentiating and applying chain rule, we will get,

$$\begin{aligned} S'(x) &= (e^x)(1+e^x)^{-1} + (e^x)(-1)(1+e^x)^{-2}(e^x) \\ &= \frac{(e^x)(1+e^x)}{(1+e^x)^2} - \frac{(e^x)^2}{(1+e^x)^2} \\ &= \frac{e^x}{(1+e^x)^2} \\ &= \frac{e^x}{1+e^x} \cdot \frac{1}{1+e^x} \\ &= S(x)(1-S(x)) \end{aligned} \tag{7}$$

Similarly, we can find the second derivative of sigmoid function as follows:

$$S''(x) = \frac{e^x\,(1-e^x)}{(1+e^x)^3} = S(x)(1-S(x))(1-2S(x)) \tag{8}$$

5 Derviation of maximum likelihood estimate for logistic regression

Logistic regression predicts probabilities, rather than just classes, so we can fit it using likelihood. For each training data-point, we have a vector of features, x_i, and an observed class, y_i. The probability of that class was either p, if $y_i = 1$, or $1 - p$, if $y_i = 0$.

$$P(Y_i = 1|X_i, \theta) = \sigma(\theta^T x_i)$$

$$P(Y_i = 0|X_i, \theta) = 1 - \sigma(\theta^T x_i)$$

To represent above probability equations in a single equation, we can write

$$P(Y_i|X_i, \theta) = (\sigma(\theta^T x_i))^{y_i} * (1 - \sigma(\theta^T x_i))^{1-y_i}$$

This is known as likelihood equation.Our goal is to maximize this likelihood for each of the samples on given data.

The likelihood is then

$$L(\theta) = \prod_{i=1}^{n} P(Y_i|X_i,\theta)$$

Taking log in both sides of above equation, we will get

$$\log L(\theta) = \log \prod_{i=1}^{n} P(Y_i|X_i,\theta)$$

Now, we can define this function as $G(\theta)$ and maximizing this is same as maximizing $L(\theta)$.

$$G(\theta) = \sum_{i=1}^{n} \log(P(Y_i|X_i,\theta))$$

Now, replacing value of $P(Y_i|X_i,\theta)$ in above equation.

$$G(\theta) = \sum_{i=1}^{n} \log[(\sigma(\theta^T x_i))^{y_i} * (1-\sigma(\theta^T x_i))^{1-y_i}]$$

$$G(\theta) = \sum_{i=1}^{n} y_i \log[(\sigma(\theta^T x_i))] + (1-y_i) log[(1-\sigma(\theta^T x_i))]$$

Now to maximize the equation we will take differentiation w.r.t. θ and set the derivatives equal to zero, and solve it.

$$G'(\theta) = \sum_{i=1}^{n} \frac{y_i}{(\sigma(\theta^T x_i))}(\sigma'(\theta^T x_i))x_i - \frac{(1-y_i)}{(1-\sigma(\theta^T x_i))}(\sigma'(\theta^T x_i))x_i$$

$$G'(\theta) = \sum_{i=1}^{n} \left(\frac{y_i}{\sigma(\theta^T x_i)} - \frac{(1-y_i)}{1-\sigma(\theta^T x_i)}\right)\sigma'(\theta^T x_i))x_i$$

By replacing value of $\sigma'(\theta^T x_i)) = \sigma(\theta^T x_i))(1-\sigma(\theta^T x_i)))$

We will get,

$$G'(\theta) = \sum_{i=1}^{n} \left(\frac{y_i}{\sigma(\theta^T x_i)} - \frac{(1-y_i)}{1-\sigma(\theta^T x_i)}\right)\sigma(\theta^T x_i))(1-\sigma(\theta^T x_i)))x_i$$

$$G'(\theta) = \sum_{i=1}^{n} \left(y_i(1-\sigma(\theta^T x_i)) - (1-y_i)\sigma(\theta^T x_i)\right)x_i$$

By simplifying the equation finally we will get,

$$\boxed{G'(\theta) = \textstyle\sum_{i=1}^{n} \left(y_i - \sigma(\theta^T x_i))\right) x_i}$$

To solve this equation and find θ, we can use gradient descent algorithm. Then we substitute this value in sigmoid function $\sigma(\theta^T x_i))$ and predict the class by using threshold value.

6 An example of Logistic Regression and how does it handle the multiclass problem.

Let us consider the task of predicting genders (male/female) based on heights and weights.

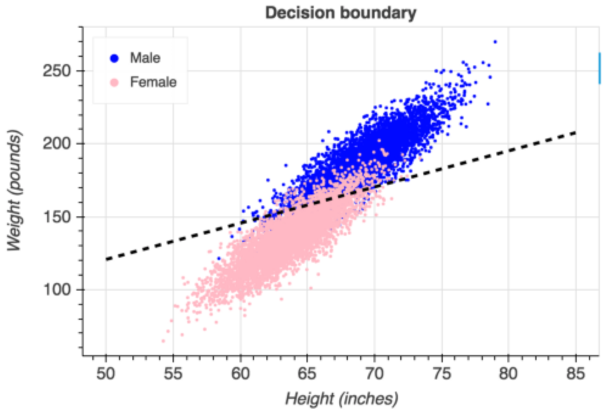

Figure 4: Mapping of male and female based on heights and weights

This example is taken from Conway and Myles Machine Learning for Hackers book, Chapter 2 containing a balanced data set of 10,000 samples of people's weight and height.

As per Logistic regression hypothesis, we know:

$$H_\theta(x) = \frac{1}{(1 + e^{-\theta^T x})} \tag{9}$$

Since our data set has two features: height and weight, the logistic regression hypothesis is the following:

$$H_\theta(x) = S(\theta_0 + \theta_1 x_1 + \theta_2 x_2) \tag{10}$$

This classifier will predict "Male" if:

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 \geq 0 \tag{11}$$

And classify female for less than 0. By using coefficients, a manual prediction would be to simply compute the vector product $\theta^T x$ and to check if the resulting scalar is bigger than or equal to zero i.e. Male or otherwise Female. Now, lets say if we have more than two classes, we can still use logistic regression in that case.There are basically two ways to do it:

1. One to One
2. One to Many

One to One:

One to One is one of the heuristic method for using binary classification algorithms for multi-class classification.It splits a multi-class classification dataset into binary classification problems.

For example, consider a multi-class classification problem with four classes: 'red,' 'blue,' and 'green,' 'yellow.' This could be divided into six binary classification datasets as follows:

Binary Classification Problem 1: red vs. blue
Binary Classification Problem 2: red vs. green
Binary Classification Problem 3: red vs. yellow
Binary Classification Problem 4: blue vs. green
Binary Classification Problem 5: blue vs. yellow
Binary Classification Problem 6: green vs. yellow

The formula for calculating the number of binary datasets, and in turn, models, is as follows: (NumClasses * (NumClasses – 1)) / 2. We can see that for four classes, this gives us the expected value of six binary classification problems. Each binary classification model may predict one class label and the model with the most predictions or votes is predicted by the one to one strategy.

One to Many:

One to many is another heuristic method for using binary classification algorithms for multi-class classification.

It also involves splitting the multi-class dataset into multiple binary classification problems. For example, given a multi-class classification problem with examples for each class 'red,' 'blue,' and 'green'. This could be divided into three binary classification datasets as follows:

Binary Classification Problem 1: red vs [blue, green] Binary Classification Problem 2: blue vs [red, green] Binary Classification Problem 3: green vs [red, blue]

A binary classifier is then trained on each binary classification problem and predictions are made using the model that is the most confident.

References

[1] https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

[2] https://www.analyticsvidhya.com/blog/2020/10/demystification-of-logistic-regression/

[3] https://www.stat.cmu.edu/ cshalizi/uADA/12/lectures/ch12.pdf

[4] Conway and Myles Machine Learning for Hackers book, Chapter 2