Task – 1

Sub Task – 1

- Prakhar Tripathi

A detailed mathematical explanation of the Logistic Regression model describing the principles, assumptions, and equations involved and explaining how the model learns from the data and makes predictions.

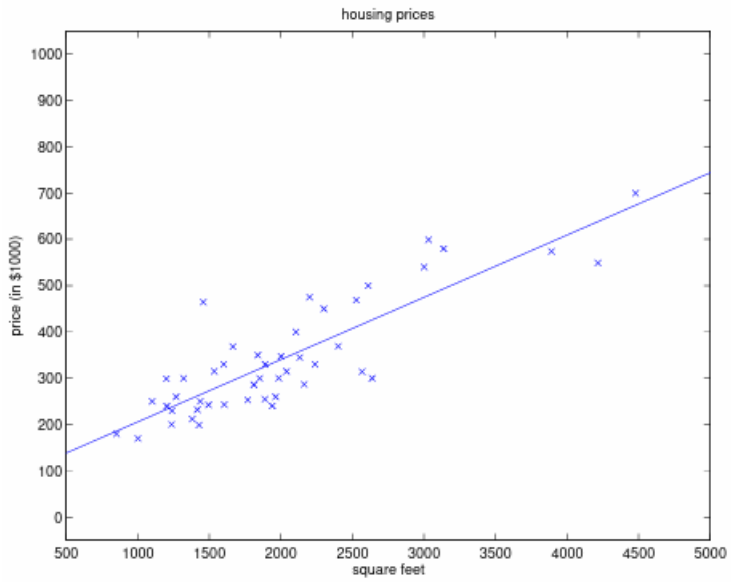
Contents

What is Regression?

In Machine Learning there are various models to perform different tasks. We may have datasets in which the data may be discrete, continuous-valued and even non-labelled.

If we have labelled data and the output obtained is continuous-valued, then such models may be said to be Regression model.

For example, I have a dataset with size of land in New Delhi and its housing prices. I want to make a model that would predict the price of houses with that certain land size. In such a case I will be getting a continuous distribution of label prices against the features land sizes.



In the above figure, a linear plot is obtained as a result of a certain model and is continuous valued. Hence such type of models are known as Regression.

There are mainly two types of Regression models

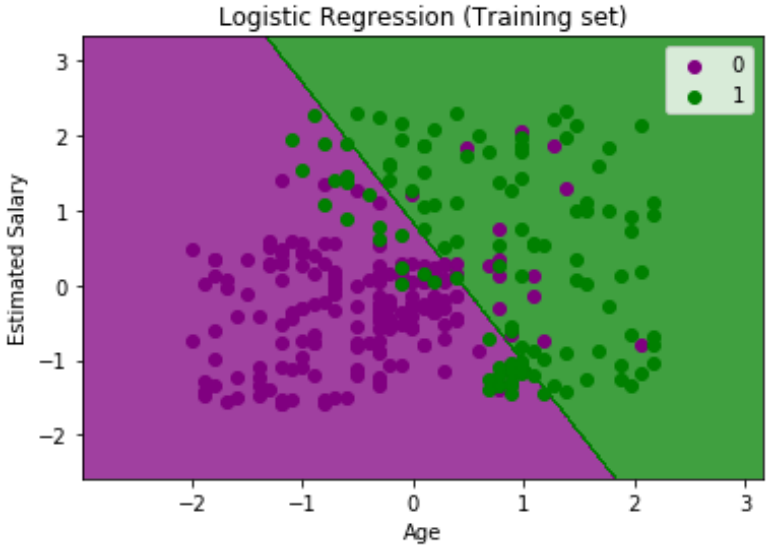
1) Linear Regression
2) Logistic Regression

Why is Logistic Regression not a Regression?

Here the name suggests that it should be a Regression model but Logistic Regression is based on the principle of classification. It is based such that it will classify the given data in various classes and label it.

For example, there are 100 students in a class. Each student is either fail or pass. So here Logistic Regression is such that it will classify the students into two classes pass and fail based on the threshold mark.

So Logistic Regression is not a Regression model but a classification model because it does not generate any continuous valued plot.



In the above figure the data is represented by the dots and the Logistic Regression model draws a Threshold line. According to the threshold line the data on the left side of the line belongs to class 0 and the data on the right side of the line belongs to class 1.

Introduction To GLM's

GLM stands for Generalized Linear Models. It represents a family of distributions in which different distributions follow the same equation pattern. There are many distributions like Gaussian, Bernoulli, Poisson etc.

The Exponential Family

According to GLM's, every distribution is a part of the exponential family.

Let's say a class of distributions follows exponential family equations.

According to it,

$$p(y;\ \eta) = b(y)exp\big(\eta^T T(y) - a(\eta)\big)$$

Here probability of y parametrized by parameter omega is according to the above formula.

Here, ω is the natural parameter, T(y) is often equal to y and a(η) is the log-partition function and b(y) is often a constant.

Assumptions of GLM

To derive a GLM for this problem, we will make the following three assumptions about the conditional distribution of y given x and about our model:

1) $y|x;\theta \sim ExponentialFamily(\eta)$
   It means given some parameter theta and x, the distribution of y follows some exponential family distribution.

2) $h(x) = E[y|x]$
   It simply states that our hypothesis function(will be described later) which is the output of our model is given to us by the estimated value of the GLM.

3) $\eta = \theta^T x$
   This simply states that the output variable of our GLM will be nothing but the required value of our machine learning model. There is no logical explanation for this as this is a design choice of the GLM

These are the assumptions to be made whenever we consider the distribution of our model to be a part of

Bernoulli Distribution

The Bernoulli distribution written as Bernoulli (ϕ) specifies a distribution over y ϵ (0,1).

Let the distribution be based on two different classes 0 and 1.

$$\text{So, } p(y = 1; \phi) = \phi$$

$$\text{Then, } p(y = 0; \phi) = 1 - \phi$$

Combining them, we can say

$$p(y; \phi) = \phi^{y}(1 - \phi)^{1 - y}$$

It states that if y=1; $p(y; \phi) = \phi$ and for y=0; $p(y; \phi) = 1 - \phi$. So it still means the same.

We now have to show that Bernoulli distribution is also a part of Exponential Family. In such a way we can say that Bernoulli is also a GLM.

Bernoulli Under GLM

As we know,

$$\begin{aligned} p(y;\ \phi) &= \phi^{y}(1-\phi)^{1-y} \\ &= exp(log((\phi^{y}(1-\phi)^{1-y})) \\ &= exp\big(ylog(\phi) + (1-y)log(1-\phi)\big) \\ &= exp(y\ log(\frac{\phi}{1-\phi}) + log(1 - \phi) \end{aligned}$$

So, the Bernoulli distribution can be expressed as the above expression.

as we are saying that Bernoulli is a part of Exponential family, then it should hold true the property equation of the Exponential Family.

Comparing the above equation with the equation of the Exponential Family, we get.

$$T(y) = y$$

$$a(\eta) = -log(1 - \phi)$$

$$= log(1 + e^{\eta})$$

b(y) = 1

$$\eta = log(\frac{\phi}{1 - \phi})$$

$$\phi = {}^{1}/_{(1 + e^{-\eta})}$$

Therefore, we can say that Bernoulli distribution is a part of Exponential family.

Logistic Regression

In Machine Learning, we are encountered with many types of datasets. The data that are used to be predicted is classified into two categories; features and labels. Features are the data that are the X's which are the input variables and the label is the output variable y. Since we have a dataset with input and output values, all we require is a curve fitting the model. We need a function that would be trained on the basis of given data points and would fit them as well. When we will have the function fitting the trained dataset, we can make predictions on the basis of the trained function. In terminology of Machine Learning, this function is known as hypothesis.

Hypothesis is a function which is assumed to fit the data points most accurately and would make correct predictions. When we train the model with the given data points, then our main logic is to anyhow find the most fit hypothesis function in order to make better predictions. Since, every feature will affect the output label, so we can say that the hypothesis function will depend on each of them.

Let's assume hypothesis function will depend on the features linearly.

In such a case,

$$h_{\theta}(x) = \theta_1 x_1 + \theta_2 x_2 + \ldots\ldots \theta_d x_d$$

if there are d features in total

But it is not always necessary that the hypothesis function will pass through the origin.

So there should be a constant term as well.

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots\ldots \theta_d x_d$$

It can also be written as:

$$h_\theta(x) = \theta^T X$$

where:

θ is a column matrix of all the thetas and **X** a column matrix of all the X's or the features.

θ is a column matrix of order **d** x 1 and **X** is also a column matrix same order. On transposing the θ matrix; it is converted into a row matrix with order 1 x **d**. On multiplying it with **X** we get the required hypothesis function as stated above.

Logistic Regression and Bernoulli

Logistic Regression is a machine learning model in which data is categorized into two classes; Class 0 and Class 1. Therefore, for any element there exists only two possibilities; either it belongs to Class 0 or Class 1.

It sounds similar to what Bernoulli says. According to Bernoulli, if the probability of any element belonging to a class is p then the probability of it not belonging to the class is 1 – p. So, by thinking it through the way that if it does not belong to the first class the it belongs to the other class. So, the probability of it belonging to the other class is 1 – p given there are only probable classes. This thus serves the purpose of Bernoulli.

Hence, we can say that elements following the Logistic Regression model follows the Bernoulli distribution. Bernoulli distribution is also a part of Exponential Family; the Logistic Regression model should abide by the conditions followed by the Bernoulli distribution as a part of the Exponential Family.

Relation between Logistic Regression and GLM

Logistic Regression works on Binary Classification, so y ϵ {0, 1}. Since it is a binary classification, it seems natural to choose Bernoulli Distribution to model

the conditional distribution of y given x. In the Bernoulli Distribution formulation we had $\phi = {}^{1}/_{(1 + e^{-\eta})}$. We also know

$$y|x;\theta \sim Bernoulli(\phi)$$

Then,

$$E[y|x;\ \theta] = \phi$$

As the expected value of the GLM in case of Bernoulli distribution is ϕ because that is only the unknown quantity in the Bernoulli Distribution.

From Assumption – 2 Of GLM,

$$\begin{aligned} h_\theta(x) &= E[y|x;\theta] \\ &= \phi \\ &= 1/(1 + e^{-\eta}) \\ &= 1/(1 + e^{-\theta^T x}) \end{aligned}$$

The second equation came from the fact explained above; the third equality follows from Assumption – 1 and is derived under the section Bernoulli Under GLM, and the last one came from Assumption – 3 of Assumptions of GLM.

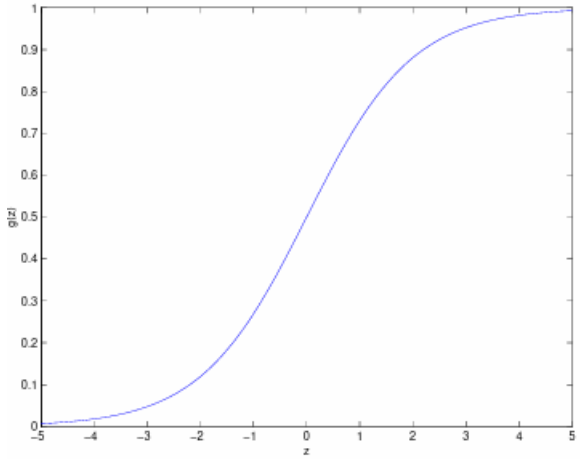
By the above transformations we have got our much needed hypothesis function which is the output value of our Logistic model.

Significance of Sigmoid Function

The hypothesis of the Logistic Regression is the sigmoid function. The speciality of this function is that it's range is limited in {0, 1} across the whole number scale.

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

Here g(z) is called the logistic or the sigmoid function.



This figure depicts the graph of sigmoid function.

In sigmoid function, as g(z)→1 as z→ ∞ and g(z)→0 as z→ $-\infty$.

Moreover, h(x) is always bounded between 0 and 1.

Derivations Under Sigmoid

As assumed before, $x_0 = 1$,

so that

$$\theta^T x = \theta_0 + \Sigma_{j=1}^{d} \theta_j x_j$$

Let's find out the derivative of the sigmoid function as it will be used further.

$$g'(z) = \frac{d}{dz}\left(\frac{1}{1+e^{-z}}\right)$$

$$= \frac{1}{(1+e^{-z})^2}(e^{-z})$$

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

$$= g(z)(1 - g(z))$$

Likelihood and Probability

- Probability refers to the chance that a particular outcome occurs based on the values of parameters in a model.
- Likelihood refers to how well a sample provides support for particular values of a parameter in a model.

In Machine Learning whenever the data of training set is considered as a fixed thing, it is likelihood whereas whenever the data is varied keeping the parameters fixes is known as probability.
So, firstly the likelihood of parameters is determined to optimize them and find out the best parameters. Then probability of the required data is calculated.

Likelihood: Introduction

As our model is Logistic Regression distributed over Bernoulli Distribution

Let's assume the probability of the label to be equal to 1 is as:

$$P(y = 1|x; \theta) = h_\theta(x)$$

Then by Bernoulli, the probability of the event not be 1 i.e. to be 0 will be:

$$P(y = 1|x; \theta) = 1 - h_\theta(x)$$

As discussed in the Bernoulli Introduction section, 1

$$p(y|x; \theta) = h_\theta(x)^y(1 - h_\theta(x))^{1-y}$$

Assuming that the n training examples were generated independently, we can write the likelihood of the parameters as:

$$\begin{aligned} L(\theta) &= p(\vec{y}|X; \theta) \\ &= \prod_{i=1}^{n} p\left(y^{(i)}|x^{(i)}; \theta\right) \\ &= \prod_{i=1}^{n} h_\theta\left(x^{(i)}\right)^{y^{(i)}} \left(1 - h_\theta\left(x^{(i)}\right)\right)^{1-y^{(i)}} \end{aligned}$$

It just multiplies each probability as each event is considered to be an independent event. In the above equation the index (i) goes from 1 to n means through all the dataset. The probability has a power of y just to ensure if the probability of the 'ith' term is calculated and y=1 then its value will be considered by the first term and if not then then will be considered by the second term.

Log Likelihood

It will be easier to maximize the log likelihood because log is a strictly increasing function. If we use log, then the terms in multiplication will be written in addition which will help us to maximise the likelihood.

$$\begin{aligned} l(\theta) &= log\mathrm{L}(\theta) \\ &= \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{y}^{(i)} log\left(h\left(\mathrm{x}^{(i)}\right)\right) + \left(1-\mathrm{y}^{(i)}\right) log\left(1-h\left(\mathrm{x}^{(i)}\right)\right) \end{aligned}$$

To maximise the likelihood, we can use gradient ascent. To define the gradient ascent function, we need to differentiate the function of log likelihood.

$$\begin{aligned} \frac{\partial}{\partial x} l(\theta) &= \left(y\frac{1}{g(\theta^{\mathrm{T}}x)} - (1-y)\frac{1}{1-g(\theta^{\mathrm{T}}x)}\right)\frac{\partial}{\partial \theta_{\mathrm{j}}} g(\theta^{\mathrm{T}}x) \\ &= \left(y\frac{1}{g(\theta^{\mathrm{T}}x)} - (1-y)\frac{1}{1-g(\theta^{\mathrm{T}}x)}\right) g(\theta^{\mathrm{T}}x)\left(1-g(\theta^{\mathrm{T}}x)\right)\frac{\partial}{\partial \theta_{\mathrm{j}}}\theta^{\mathrm{T}} \\ &= \left(y\left(1-g(\theta^{\mathrm{T}}x)\right) - (1-y)g(\theta^{\mathrm{T}}x)\right)\mathrm{x}_{\mathrm{j}} \\ &= (\mathrm{y}\text{-}\mathrm{h}_{\theta}(\mathrm{x}))\mathrm{x}_{\mathrm{j}} \end{aligned}$$

Here we used the fact that:

$$g'(z) = g(z)(1 - g(z))$$

Update Rule and Prediction

This gives us the stochastic gradient ascent rule:

$$\theta_{\mathrm{j}} := \theta_{\mathrm{j}} + \alpha(\mathrm{y}^{(\mathrm{i})} \text{-} \mathrm{h}_{\theta}(\mathrm{x}^{(\mathrm{i})})\mathrm{x}_{\mathrm{j}}^{(\mathrm{i})}$$

This runs through the whole data and finds the parameter that best fits the model. After calculating the theta we can calculate the hypothesis of each data and then we can divide them in two classes on the basis of a threshold. Like if the threshold=0.5, then the hypothesis values coming less than 0.5 will be classified as class 0 and the ones above the threshold will be classified in the class 1.