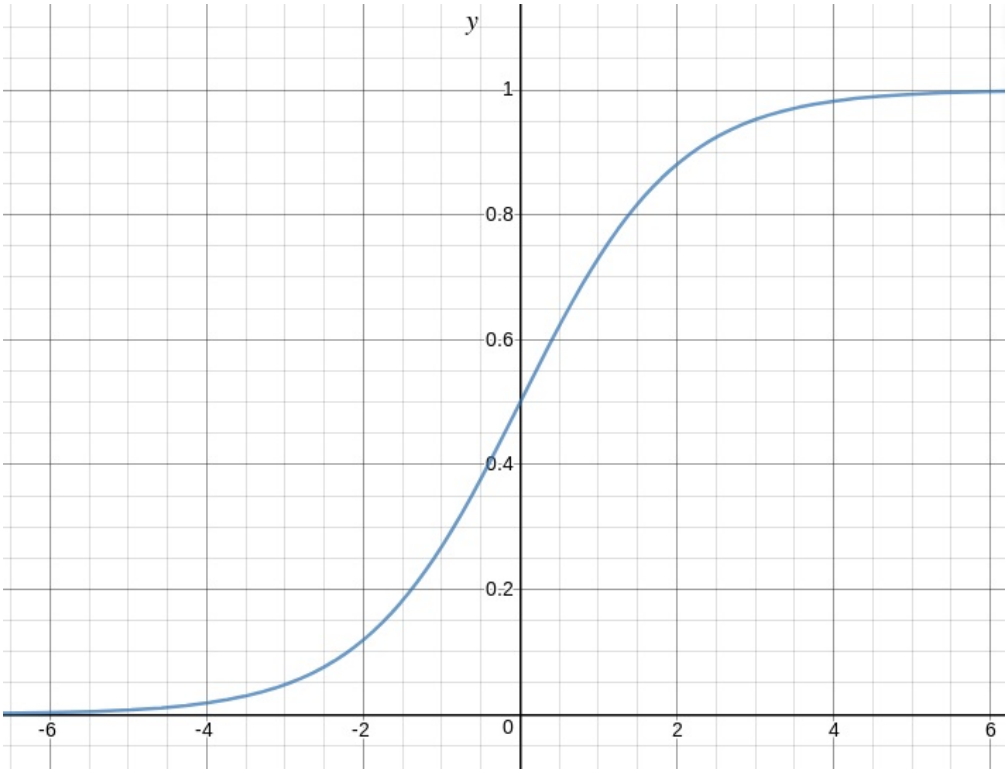
Manisha Chawla

Manisha Chawla
ID 12210370
Homework-2(Problem-IX)

Logistic Regression:-

Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable(or output), y, can take only discrete values for a given set of features(or inputs), X. Logistic regression is a regression model. The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as “1”. Just like Linear regression assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function.

Sigmoid Function:



Sigmoid Function$= \sigma(z) = \frac{1}{1+e^{-z}}$

$\sigma(z)$ tends towards 1 as $z \to \infty$
$\sigma(z)$ tends towards 0 as $z \to -\infty$
$\sigma(z)$ is always bounded between 0 and 1

Logistic regression becomes a classification technique only when a decision threshold is brought into the picture. The setting of the threshold value is a very important aspect of

$p = \sigma(W^T x_i) = \frac{1}{1 + e^{-W^T x_i}}$ and $L(W) = \sum_{i=1}^{\mathbb{N}} [y_i log(p_i) + (1 - y_i) log(1 - p_i)]$**

As we know the hypothesis of linear regression is:

$$h_w(x) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 .... + w_i x_i = W^T X$$

For logistic regression, focusing on binary classification here, we have class 0 and class 1 for a given x.To compare with the target, we want to constrain predictions to some values between 0 and 1. That's why **Sigmoid Function** is applied on the raw model output and provides the ability to predict with probability.

As we know $SigmoidFunction = \sigma(z) = \frac{1}{1+e^{-z}}$

So our hypothesis function will become,

$$p = HypothesisFunction = \sigma(W^T x) = \frac{1}{1+e^(-W^T x)}$$

What hypothesis function returns is the probability that y = 1, given x, parameterized by W, written as: p = P(y = 1|x; W). Decision boundary can be described as:
Predict 1, if $W^T x \geq 0 \rightarrow p(x) \geq 0.5$;
Predict 0, if $W^T x < 0 \rightarrow p(x) < 0.5$.

$$p(X) = p(Y = 1|X) = \frac{1}{1+e^(-W^T x)} = \frac{e^(W^T x)}{1+e^(W^T x)}$$

$$1 - p(X) = p(Y = 1|X) = 1 - \frac{1}{1+e^(-W^T x)} = \frac{1}{1+e^(W^T x)}$$

$$\frac{p(X)}{1-p(X)} = e^{W^T x}$$

taking log both sides,

$$log(\frac{p(X)}{1-p(X)}) = W^T x$$

This is $logit$ of $P(X)$

Given samples $\{x_i, y_i\} \in \mathbb{R}^p X \{0, 1\} \forall i = 1, 2, 3..n$

from $logit$ of $P(x_i)$

$$log(\frac{p(x_i)}{1-p(x_i)}) = W^T x_i$$

We need to estimate $\{w_0, w_1, w_2, w_3 .... w_i = \hat{W}\}$

Cost Function:
Intuitively, we want to assign more punishment when predicting 1 while the actual is 0 and when predict

0 while the actual is 1. The loss function of logistic regression is doing this exactly which is called Logistic Loss.

We can use Maximum Likelyhood Estimation of that, in MLE we take all the data and break it into two groups based on their labels.

MLE can be given as:
For smaple labeled "1" estimate W such that $p(x)$ is as close to 1 as possible. i.e. $\pi_{i,y_i=1}p(x_i)$

For smaple labeled "0" estimate W such that $1-p(x)$ is as close to 1 as possible. i.e. $\pi_{i,y_i=0}(1-p(x_i))$

We can write it as

$$L(W) = \pi_{i,y_i=1}p(x_i).\pi_{i,y_i=0}(1-p(x_i))$$

In a generalised form,

$$L(W) = \pi_{i=1}^{n}p(x_i)^{y_i}.(1-p(x_i))^{1-y_i}$$

Our objective is to maximize this L(W).

Objective : $\max_w L(W)$ or $-\min_w L(W)$

Alll the $p(x_i)$ and $(1-p(x_i))$ are the probabilities and we have to maximize it.
So to minimizing this is equivalent to minimizing the log of that function. Because this function is always positive.

log likelihood:

$$l(W) = -logL(W) = -\sum_{i=1}^{n} y_i logp(x_i) + (1-y_i)log(1-p(x_i))$$

For minimizing $l(W)$,we need to do compute its gradient $\frac{\partial l}{\partial W}$

As we know that differentiation of sigmoid function will give :

$$\frac{d}{dz}\sigma(z) = \sigma(z)(1-\sigma(z))$$

$$\frac{\partial l}{\partial W} = -\sum_{i=1}^{n} y_i \frac{\partial}{\partial W} logp(x_i) + (1-y_i)\frac{\partial}{\partial W}log(1-p(x_i))$$

$$\frac{\partial l}{\partial W} = -\sum_{i=1}^{n} y_i \frac{\partial}{\partial W} log\sigma(W^T x_i) + (1-y_i)\frac{\partial}{\partial W}log(1-\sigma(W^T x_i))$$

$$\frac{\partial l}{\partial W} = -\sum_{i=1}^{n} y_i \frac{1}{\sigma(W^T x_i)}\frac{\partial}{\partial W}\sigma(W^T x_i) + (1-y_i)\frac{1}{1-\sigma(W^T x_i)}\frac{\partial}{\partial W}(1-\sigma(W^T x_i))$$

$$\frac{\partial l}{\partial W} = -\sum_{i=1}^{n} y_i \frac{1}{\sigma(W^T x_i)}\sigma(W^T x_i)(1-\sigma(W^T x_i))\frac{\partial}{\partial W}(W^T x_i)$$

$$-(1-y_i)\frac{1}{1-\sigma(W^T x_i)}\sigma(W^T x_i)(1-\sigma(W^T x_i))\frac{\partial}{\partial W}(W^T x_i)$$

$\frac{\partial l}{\partial W} = -\sum_{i=1}^{n} y_i(1-\sigma(W^T x_i))(x_i) - (1-y_i)\sigma(W^T x_i)(1-\sigma(W^T x_i))(x_i)$

$\frac{\partial l}{\partial W} = -\sum_{i=1}^{n} (y_i - \sigma(W^T x_i))(x_i)$

$\frac{\partial l}{\partial W} = -\sum_{i=1}^{n} (y_i - p(x_i))(x_i)$

$\frac{\partial l}{\partial W} = -\sum_{i=1}^{n} (y_i - \hat{y}_i)(x_i)$

Gradient = $\frac{\partial l}{\partial W} = \sum_{i=1}^{n} (\hat{y}_i - y_i)(x_i)$

.

We update W as $W^{t+1} = W^t - \eta \frac{\partial l}{\partial W}$ where t = iteration.