Epoch – Learning Phase

Machine Learning

Topic Report

Ishaan Jain

CO21BTECH11006

Logistic Regression:

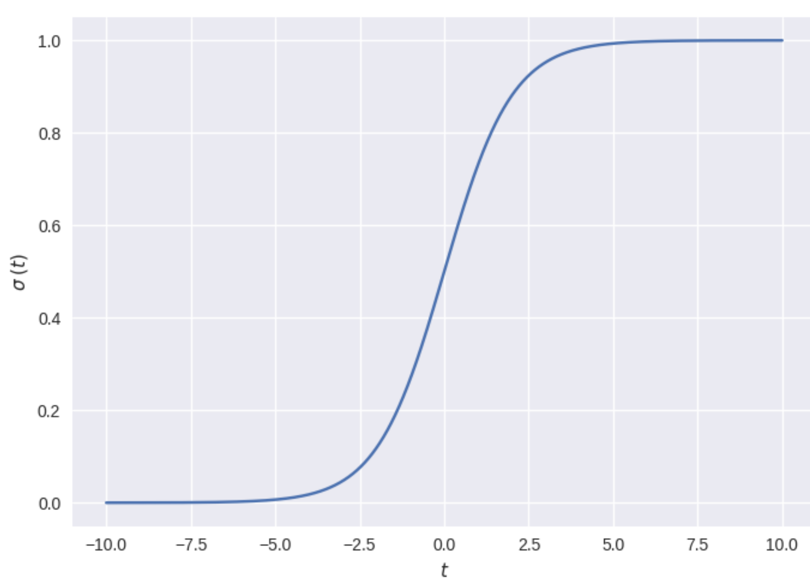
Logistic Regression is not a regression algorithm, it is a classification algorithm unlike its name. The hypothesis returns a real number in the range [0,1]. If the hypothesis function returns a value greater than 0.5, it is classified as it belongs to class 1, otherwise it is classified as class 0. Hence, logistic regression performs binary classification.

Hypothesis:

The hypothesis function is defined as the sigmoid function; it is defined as:

$h_\theta(x) = \sigma(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$ Where θ is the parameter being trained and x are the features of data.

$$\sigma(t) = \frac{1}{1+e^{-t}}$$



This function gives $\sigma(t) > 0.5 \; \forall \; t > 0$ and $\sigma(t) < 0.5 \; \forall \; t < 0$

$\theta^T x$ is the weighted sum of all the features, but instead of outputting the result, it outputs the sigmoid of that value to give it a face of probability.

Likelihood Function:

Let's take label $y = 1$ when the sample is in class 1 and $y = 0$ when the sample is in class 0. So,

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

$$P(y = 0 \mid x; \theta) = 1 - h_\theta(x)$$

This is the conditional probability of the sample data to fall in class 1 and class 0, respectively. In general, we can say that,

$$P(y \mid x; \theta) = \left(h_\theta(x)\right)^y \left(1 - h_\theta(x)\right)^{1-y}$$

Where y is the true label.

Assuming our dataset to be independently and identically distributed, we can define the likelihood as,

$$L(\theta) = P(\vec{y} \mid \vec{x}; \theta)$$

$$L(\theta) = \prod_{i=1}^{m} P\left(y^{(i)} \mid x^{(i)}; \theta\right)$$

We can take the log of likelihood,

$$l(\theta) = \log\left(L(\theta)\right) = \sum_{i=1}^{m} \log\left(P\left(y^{(i)} \mid x^{(i)}; \theta\right)\right)$$

$$l(\theta) = \sum_{i=1}^{m} \left[y^{(i)} \log\left(h_\theta\left(x^{(i)}\right)\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_\theta\left(x^{(i)}\right)\right)\right]$$

Training:

In order to maximize the likelihood, we should choose a θ such that $l(\theta)$ is maximum. We can use Gradient Ascent for this purpose. We will take a guess value of theta and keep it updating in the direction of gradient.

$$\theta_j \leftarrow \theta_j + \alpha \frac{\partial(l(\theta))}{\partial \theta_j} \tag{a}$$

The gradient, $\frac{\partial(l(\theta))}{\partial \theta_j}$ is calculated as

$$\frac{\partial\left(l(\theta)\right)}{\partial\theta_j} = \frac{\partial\left(\sum_{i=1}^{m}\left(y^{(i)} \log\left(h_\theta\left(x^{(i)}\right)\right) + \left(1 - y^{(i)}\right) \log\left(1 - h_\theta\left(x^{(i)}\right)\right)\right)\right)}{\partial\theta_j}$$

$$\frac{\partial\left(l(\theta)\right)}{\partial\theta_j} = \sum_{i=1}^{m} \left(\frac{\partial h_\theta\left(x^{(i)}\right)}{\partial\theta_j}\right)\left(\frac{y^{(i)}}{h_\theta\left(x^{(i)}\right)} + \frac{1 - y^{(i)}}{1 - h_\theta\left(x^{(i)}\right)}\right)$$

The derivative sigmoid function is $\frac{d\sigma(t)}{dt} = \sigma(t)\left(1 - \sigma(t)\right)$. So,

$$\frac{\partial\left(l(\theta)\right)}{\partial\theta_j} = \sum_{i=1}^{m} \left(h_\theta(x)\left(1 - h_\theta(x)\right)x_j^{(i)}\right)\left(\frac{y^{(i)}}{h_\theta\left(x^{(i)}\right)} + \frac{1 - y^{(i)}}{1 - h_\theta\left(x^{(i)}\right)}\right)$$

$$\frac{\partial(l(\theta))}{\partial\theta_j} = \sum_{i=1}^{m} \left(x_j^{(i)}\right)\left(y^{(i)} - h_\theta\left(x^{(i)}\right)\right) \qquad \text{...putting this in (a)}$$

$$\theta_j \longleftarrow \theta_j - \alpha \sum_{i=1}^{m} x_j^{(i)}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)$$

This equation looks similar to the gradient descent equation in linear regression, only $\frac{1}{2}$ factor isn't here and that too can be adjusted in the learning rate α. This way, we can get the optimal θ by performing several iterations until the solution converges.

Predictions:

As we found the optimal θ for our model, we can predict categories of the data set. Let us say we have a feature set x and we need to predict its class. We must find hypotheses function corresponding to that x.

If $h_\theta(x) > 0.5$, it belonged to class 1 otherwise it belonged to class 0

Here is a rough implementation of Logistic regression that I performed.