Logistic Regression

Aaryan
CO21BTECH11001

Let number of features of dataset = n

Let number of sets of features = m

Data consists of matrices X and y where i^{th} column of X represents the i^{th} feature of dataset and i^{th} element of y represents the value of variable dependent on set of features listed in i^{th} row of X.

Logistic regression is a type of classification algorithm where it assumes a linear relationship between dependent (X) and independent (y) variables.

In a binary classification problem, y consists of only two values, usually 0 and 1.

$$\text{Let} \quad X^{(i)} = \begin{bmatrix} 1 \\ X_1 \\ X_2 \\ . \\ . \\ . \\ X_n \end{bmatrix} \quad and \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ . \\ . \\ . \\ \theta_n \end{bmatrix}$$

where θ is known as parameter.

We define a hypothesis function $h_\theta(x)$ as follows –

$$h_\theta\left(X^{(i)}\right) = \frac{1}{1 + e^{-\theta^T X^{(i)}}}$$

where $X_0^{(i)} = 1$

We will calculate a value of θ which best fits the approximation –

$$If\ h_\theta\left(X^{(i)}\right) \geq 0.5\ \ then\ y_i = 1$$
$$else\ if\ h_\theta\left(X^{(i)}\right) < 0.5\ then\ y_i = 0$$

The above approximation is only for a binary classification problem. In any other classification problem, we can similarly fix landmarks for $h_\theta\left(X^{(i)}\right)$

Now, we will define a function which is a measure of probability of accuracy of hypothesis function, which is known as log-likelihood function.

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

Objective – Minimize or Converge the log-likelihood function.

There are two approaches to do this –

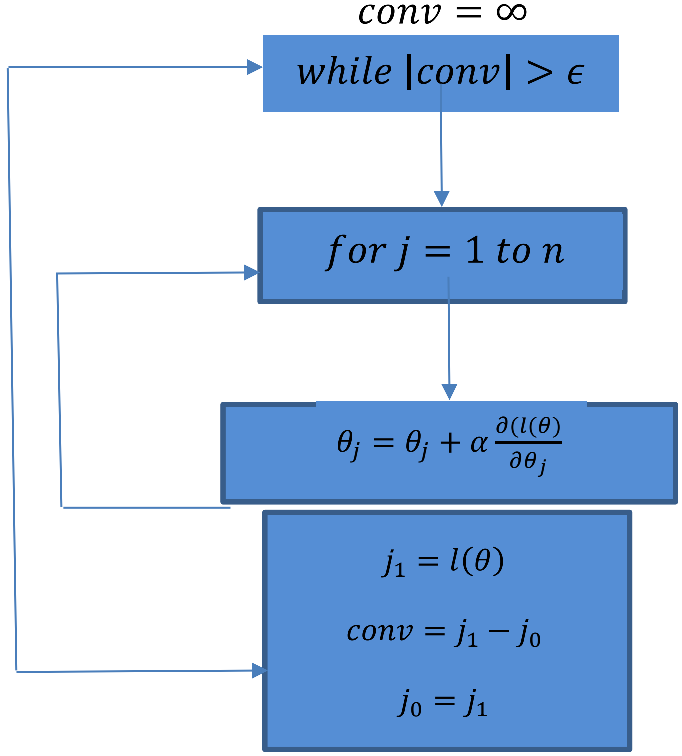
Gradient Ascent Algorithm –

$$Convergence\ limit = \epsilon = 10^{-10}$$

$$Initialize\ \theta = \vec{0}$$

$$j_0 = l(\theta)$$

$$conv = \infty$$



Newton’s Algorithm of Classification –

$$Let\ J(\theta) = -\frac{1}{m} l(\theta)$$

$$Hessian\ matrix -$$

$$H_{ij} = \frac{\partial^2 J}{\partial \theta_i \partial \theta_j}$$

$$Gradient\ vector - \qquad \nabla_\theta J = \frac{\partial J}{\partial \theta}$$

Algorithm –

$$Convergence\ limit = \epsilon = 10^{-12}$$

$$Initialize\ \theta = \vec{0}$$

$$j_0 = J(\theta)$$

$$conv = \infty$$

$$while\ |conv| > \epsilon$$

$$\theta = \theta - H^{+} \nabla_\theta J$$

$$j_1 = J(\theta)$$

$$conv = j_1 - j_0$$

$$j_0 = j_1$$

$$H^{+}\ represent\ pseudo$$

$$inverse\ of\ matrix\ H$$

$$Stop$$

After getting optimal θ, we can get the value corresponding to a new data D as

$$If\ h_\theta(D) \geq 0.5\ \ then\ val = 1$$
$$else\ if\ h_\theta(D) < 0.5\ then\ val = 0$$

Questions –

1. Is Logistic Regression a regression algorithm or classification algorithm?
   Ans. Regression algorithm
2. What is the type of decision surface in Logistic Regression algorithm?

   Ans. A linear curve (straight line)

3. Why do we need to take $X_0^{(i)} = 1\ \forall\ i$?
   Ans. Because in the hypothesis function there is a constant term apart from the linear combination of $X^{(i)}$ and θ, which is θ_0, so the multiplier of θ_0 can be any value. For simplicity, we take it as 1.
4. What is the range of values of hypothesis function?
   Ans. (0,1)
5. Name three methods by which we can increase the accuracy of logistic regression?
   Ans. Removal of incomplete dataset , Feature Scaling/ Normalization, Removal of outliers of sparse features.
6. What are the disadvantages of linear regression model?
   Ans. It constructs linear boundaries which is not as accurate in non-linear problems.