Logistic Regression

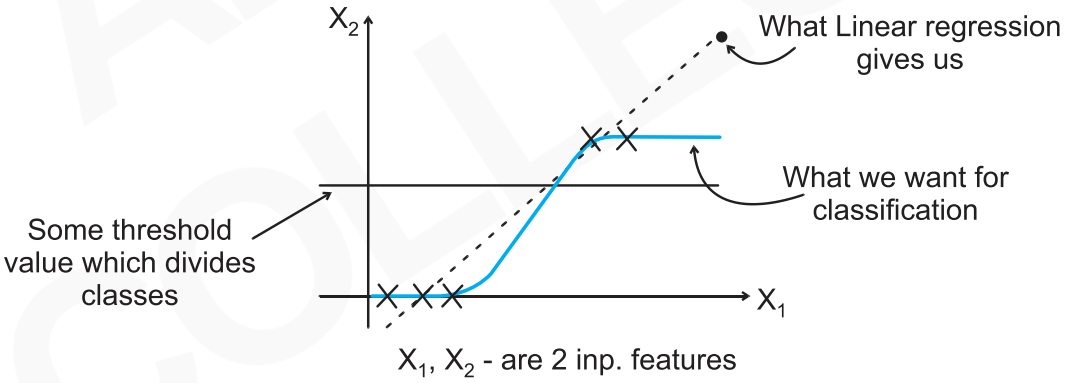
Intuition & Logic

Logistic Regression is a supervised ML algorithm for classification problems (not regression ones).

Unlike Linear Regression which predicts continuous values, we use Logistic Regression to predict the probability that an input belongs to a specific class or category.

For classification problems we want to predict a specific class, not continuous values. Let's consider an example of binary classification where we have 2 possible values in output - 0 or 1.

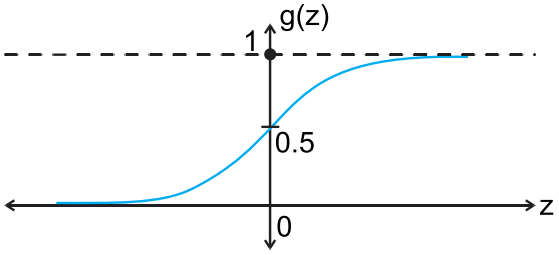
Linear regression can't predict a class for us so logistic regression uses SIGMOID function to map linear model values to a specific range of 0 to 1 forming a " S " shaped curve (sigmoid/logistic curve).



Because probabilities lie between 0 to 1 sigmoid function is great to calculate those probabilities.

What is Sigmoid function?

$$g(z) = \frac{1}{1 + e^{-z}} \text{ with range } (0, 1)$$



Where

$g(z) \to 1$ as $z \to \infty$

$g(z) \to 0$ as $z \to -\infty$

$g(z) = 0.5$ at $z = 0$

So, we use $g(z)$ as our hypothesis function where $z = \theta_0 + \theta_1 x_1$

So,

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1)$$

$$h_\theta(x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1)}}$$

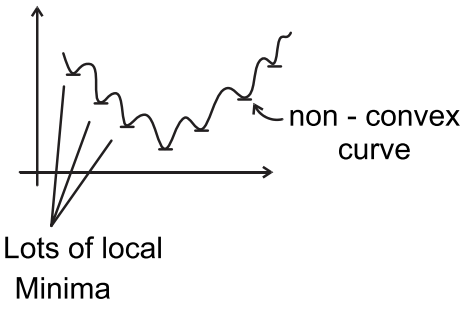
$\theta_0 + \theta_1 x_1 = \theta^T x$

so,

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

This formula represents the probability of the input belonging to class 1.

Now, how do we compute out θ_0 & θ_1, by minimizing the cost function. But we use MSE ($J(\theta)$ of linear regression) it gives us a non-convex function for which it's not easy to find the minima.



So instead of MSE, we use a different cost function i.e.

Log Loss (or Binary Cross Entropy)

All to log loss function

$$J(\hat{y}_i, y_i) = \begin{cases} -\log(\hat{y}_i) & \text{for } y = 1 \\ -\log(1-\hat{y}_i) & \text{for } y = 0 \end{cases}$$

predicted output, actual output

or

loss for each ith sample

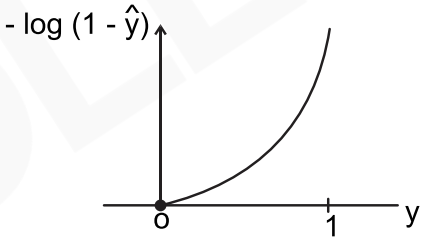
$$J(\hat{y}_i, \hat{y}_i) = -y_i\left(\log(\hat{y}_i)\right) - (1-y_i)\cdot\log(1-\hat{y}_i)$$

Why this specific cost function?

Because it penalizes severe wrong predictions.

if $y = 0$ & $\hat{y} = 0.9 \rightarrow$ we get very high loss

$y = 0$ & $\hat{y} = 0.1 \rightarrow$ how value of loss



So the cost function we use for logistic Regression is

$$J(\theta) = \frac{-1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log(\hat{y}^{(i)}) + \left(1-y^{(i)}\right).\log\left(1-\hat{y}^{(i)}\right)\right]$$

loss for entire data set

Now this cost function gives us a convex curve that guarantees us a global minimum.

In summary, we now have a approach very similar to Linear Regression with Gradient Descent.

1. Initialize parameters (θ_0, θ, etc.)
2. Use parameters to create model & make predictions.
3. Compute gradients $\frac{\partial J(\theta)}{\partial \theta_k}$
4. Update parameters
5. Repeat steps 2 to 4 until convergence.

EXTRA ADD ON

what will be $\frac{\partial J(\theta)}{\partial \theta_k}$ for $\theta_k = \theta_0, \theta_1$?

$$J(\theta) = \frac{-1}{m}\sum_{i=1}^{m} (y_i \cdot \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

& $z = \theta^T x$

using chain rule:-

$$\frac{\partial J(\theta)}{\partial \theta_1} = \frac{\partial J(\theta)}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial z}{\partial \theta_1}$$

① $$\frac{\partial J(\theta)}{\partial \hat{y}} = \frac{-1}{m}\left(\frac{y}{\hat{y}} - \frac{(1-y)}{(1-\hat{y})}\right)$$

because $\frac{\partial \log(x)}{\partial x} = \frac{1}{x}$

② $$\hat{y} = \frac{1}{1 + e^{-z}}$$

$$\frac{\partial \hat{y}}{\partial z} = \hat{y} \cdot (1 - \hat{y})$$

③ $$\frac{\partial z}{\partial \theta_1} = x \;\&\; \frac{\partial z}{\partial \theta_0} = 1$$

$$\therefore \frac{\partial J(\theta)}{\partial \theta_1} = \frac{-1}{m}(y \cdot (1 - \hat{y}) - \hat{y} \cdot (1 - y))x$$

$$\frac{\partial J(\theta)}{\partial \theta_0} = \frac{-1}{m}(y \cdot (1 - \hat{y}) - \hat{y} \cdot (1 - y))$$

$$\frac{\partial J(\theta)}{\partial \theta_1} = (1/m) * (\hat{y} - y)x$$

$$\frac{\partial J(\theta)}{\partial \theta_0} = (1/m) * (\hat{y} - y)$$

Different Types of Logistic Regression

1. **Binomial Logistic Regression**: When output has only two possible categories (binary classification). *Eg - Yes/No, Pass/Fail or 0/1*
2. **Multinomial Logistic Regression**: When output has 3 or more possible categories that are not ordered (multi-class classification). *Eg - categories for animals like "cat," "dog" or "rabbit."*
3. **Ordinal Logistic Regression**: When output has 3 or more categories with a natural order or ranking. *Eg - ratings like "low," "medium" and "high"*

Logistic Regression Assumptions

1. **Independent observations**: Each data point is assumed to be independent of the others means there should be no correlation or dependence between the input samples.
2. **Binary dependent variables**: It takes the assumption that the dependent variable must be binary, means it can take only two values.
   For more than two categories **SoftMax** functions are used.
3. **Linearity relationship between independent variables and log odds**: The model assumes a linear relationship between the independent variables and the log odds of the dependent variable which means the predictors affect the log odds in a linear way.
4. **No outliers**: The dataset should not contain extreme outliers as they can distort the estimation of the logistic regression coefficients.
5. **Large sample size**: It requires a sufficiently large sample size to produce reliable and stable results.

Keep Learning & Keep Exploring!