Machine Learning
Logistic Regression

DR. BHARGAVI R

SCOPE

VIT CHENNAI

Classification - Applications

Binary Classification

- Online transactions – Fraudulent / Not Fraudulent
- Email – Spam/ Not spam ?
- Tumor classification – Malignant/Benign

Multi-class Classification

- Optical Character Recognition
- Face classification

Multi-Label Classification

A variant of the classification problem where multiple nonexclusive labels may be assigned to each instance.

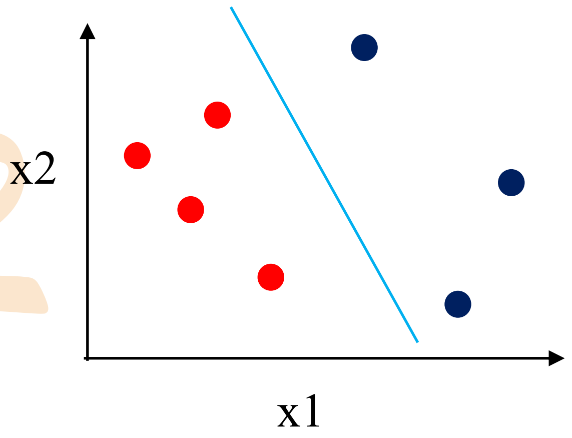
Binary classification

$y \in \{0,1\}$: 0 - Negative class (Not spam)

: 1 – Positive class (Spam)

Logistic Regression - Introduction

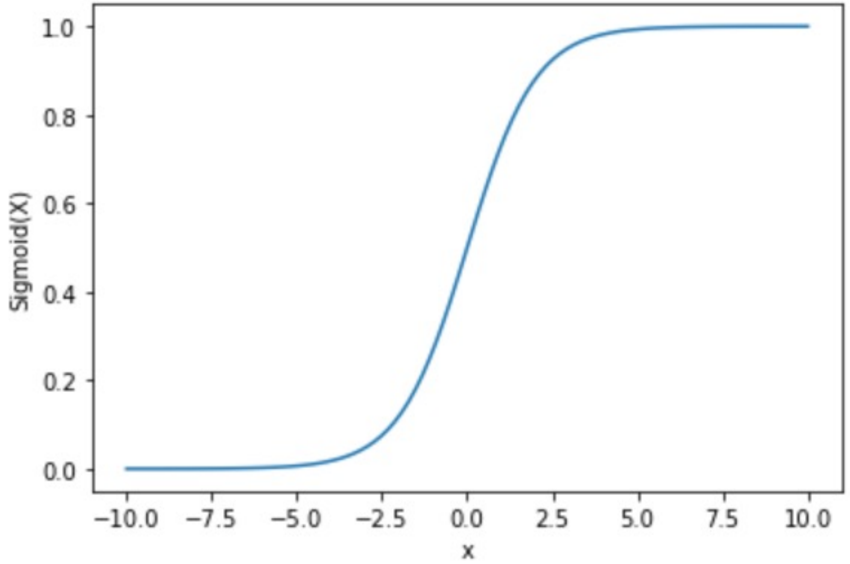
- Linear model.
- Used for binary classification
- Can be extended to handle multiclass as well
- Computationally inexpensive
- Easy to implement.



- Logistic Regression models the response/prediction as probability that y (output variable) belongs to a particular category.

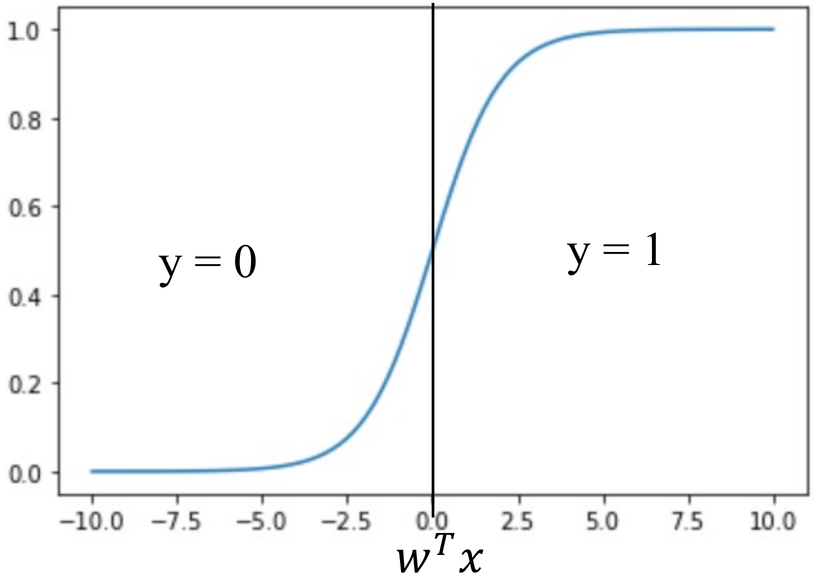
Hypothesis Function

- Can the hypothesis function $g(x, w) = \sum_{i=1}^{m} w_i x_i$ be used for classification?
- $g(x, w)$ results in a real value (–∞ <g(w,x)< +∞).
- For classification problems we need the result to be finite discrete values representing different classes.
- $Sigmoid(x) = \frac{1}{1+e^{-x}}$
- $\boxed{h_w(x) = \frac{1}{1+ e^{-g(x,w)}} = \frac{1}{1+ e^{-\sum_{i=1}^{m} w_i x_i}}}$
- $h_w(x)$ can also be written as $\frac{1}{1+ e^{-w^T X}}$
- $h_w(x)$ is called as Sigmoid/ logistic function
- $0 \le h_w(x) \le 1$



Hypothesis Function (cont...)

- Value of $h_w(x)$ is the estimated probability that y =1, for input x with given w's
- $h_w(x)$ = P(y=1 /x; w) i.e probability that y =1, for input x with given w's
- And P(y=0 /x; w) = 1 - P(y=1 /x; w) (since P(y=1 /x; w) + P(y=0 /x; w) = 1)
- If $h_w(x) \geq 0.5$ i.e $\boxed{w^T x \geq 0 \text{ then } y = 1}$
- If $h_w(x) < 0.5$ i.e $\boxed{w^T x < 0 \text{ then } y = 0}$
- For fixed w's, $\boldsymbol{w^T x}$ represent a linear decision boundary
- x_i's that results in $w^T x \geq 0$ are predicted as 1
- x_i's that results in $w^T x < 0$ are predicted as 0

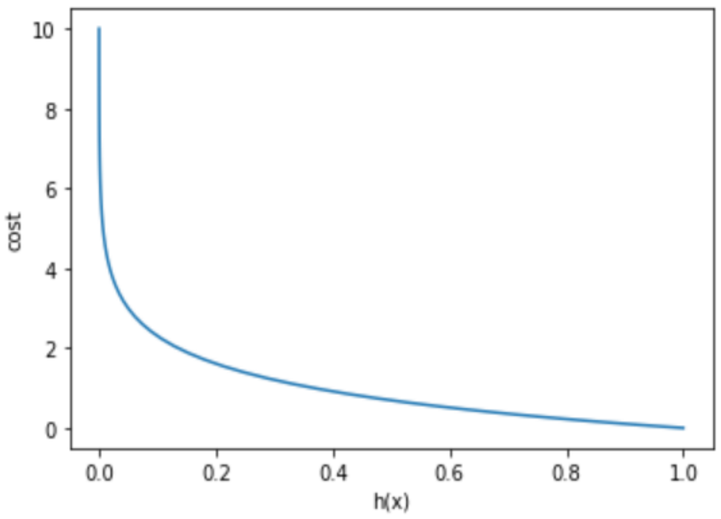


Cost Function

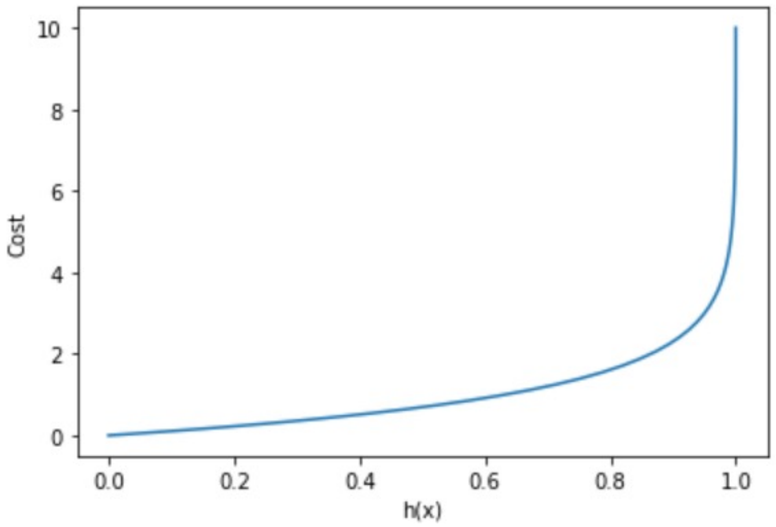
- Cost should be minimum (≈ 0) for the correct predictions and maximum(≈ ∞ or very high value) for the wrong predictions.

For a single observation : Cost $(h_w(x), y) =$ $-\log(h_w(x))$ for $y = 1$ (i.e for +ve sample)

$-\log(1 - h_w(x))$ for $y = 0$ (i.e for -ve sample)



$-\log(h_w(x))$ plot



$-\log(1-h_w(x))$ plot

Cost Function (cont…)

- Combining the cost for positive and negative predictions into a single equation

$$\text{Cost}\,(h_w(x), y) = -y\log\big(h_w(x)\big) - (1-y)\log\big(1 - h_w(x)\big)$$

- Cost function for n data points can be written as

$$\begin{aligned} J(w) &= \frac{1}{n}\sum_{i=1}^{n} cost(h_w(\boldsymbol{x}_i, y_i)) \\ &= \frac{1}{n}\sum_{i=1}^{n} -y_i log\big(h_w(x_i)\big) - (1-y_i)\log(1 - h_w(x_i)) \\ &= -\frac{1}{n}\sum_{i=1}^{n} y_i log\big(h_w(x_i)\big) + (1-y_i)\log(1 - h_w(x_i)) \end{aligned}$$

Where $h_w(\boldsymbol{x}_i) = \dfrac{1}{1+e^{-W^T x_i}}$

Gradient of Cost Function

Gradient of J(w):

Derivative of hypothesis function

$$\frac{dh_w}{dw} = \frac{-1}{\left(1+e^{-\boldsymbol{w}^T\boldsymbol{x}_i}\right)^2} \times e^{-\boldsymbol{w}^T\boldsymbol{x}_i} \times (-\boldsymbol{x}_i)$$

$$= \frac{1}{1+e^{-\boldsymbol{w}^T x_i}} \frac{e^{-\boldsymbol{w}^T x_i}}{1+e^{-\boldsymbol{w}^T x_i}} \boldsymbol{x}_i = h_w(1 - h_w)x_i$$

Substituting partial derivative of the hypothesis in the cost function and simplifying we get

Gradient of J(w) (say w.r.t w_j) = $\frac{1}{n}\sum_{i=1}^{n}(h_w(\boldsymbol{x}_i) - y_i)x_{ij}$

Gradient Descent

Do repeatedly

{

$$w_j = w_j - \alpha \frac{\partial}{\partial w_j}(J(W))$$

}

or

Do repeatedly

{

$$w_j = w_j - \alpha \sum_{i=1}^{n}(h_w(\boldsymbol{x}_i) - y_i)x_{ij}$$ (Simultaneously update all W_j's)

}

Here α is the learning rate

Prediction

The marketing department of a credit card company wants to organize a campaign to convince existing holders of the company's standard credit card to upgrade to the company's premium card for a nominal annual fee. The marketing department begins with the question "Which of the existing standard credit cardholders should be the target for the campaign?"

Dataset - 30 cardholders data that indicates whether the cardholder upgraded to a premium or not (y i.e response)

Two independent variables/features:

1. Total amount of credit card purchases in the prior year(x1)
2. Whether the cardholder ordered additional credit cards (at extra cost) for other members of the household (x2 : 0 no, 1 yes).

The regression coefficient vales are w_0 = -6.940, w_1 = 0.13947, w_2 = 2.774

Prediction (cont…)

Consider a cardholder who charged $36,000 last year and possesses additional cards for members of the household. What is the probability the cardholder will upgrade to the premium card during the marketing campaign.

Substitute the w_is and x_is in the function $h_w(\boldsymbol{x}_i) = \frac{1}{1+ e^{-W^T x_i}}$ to the predicted probability.

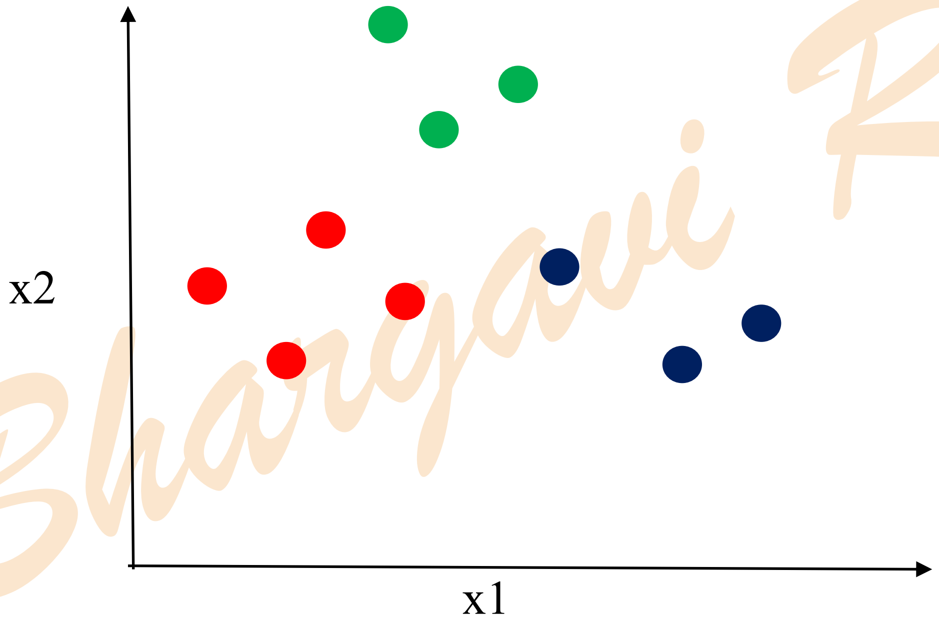
$-6.94 + (0.13947)(36) + (2.774)(1) = 0.85492$

$e^{-(0.85492)} = 0.423$

Estimated probability of purchasing premium card = $1 / (1 + 0.423) = 0.702$

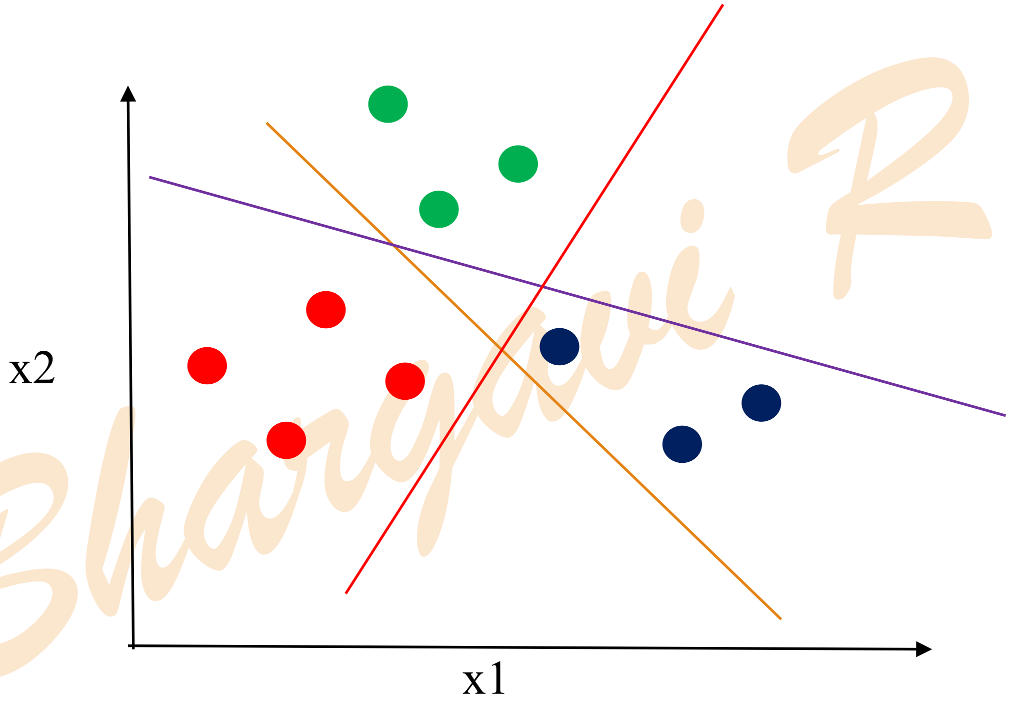
Multiclass Classification

$y \in \{1, 2, 3, \ldots.\}$



Multiclass Classification (cont…)

One-vs-all (or) one-vs-rest



Multiclass Classification (cont…)

One-vs-all (or) one-vs-rest :

Step1 : Modify the training data such that only one specific class has y=1 and rest all have y=0.

Step 2: Train the classifier.

Step3: Repeat Step 1 and 2 for remaining all classes each time making one class as y=1 and remaining all as y=0 and training individual models.

Step 4: Prediction : For a new test input x_t pick a class that maximizes the $h_w(x_t)$