Maximum Likelihood Estimation for Linear and Logistic Regression

Nipun Batra

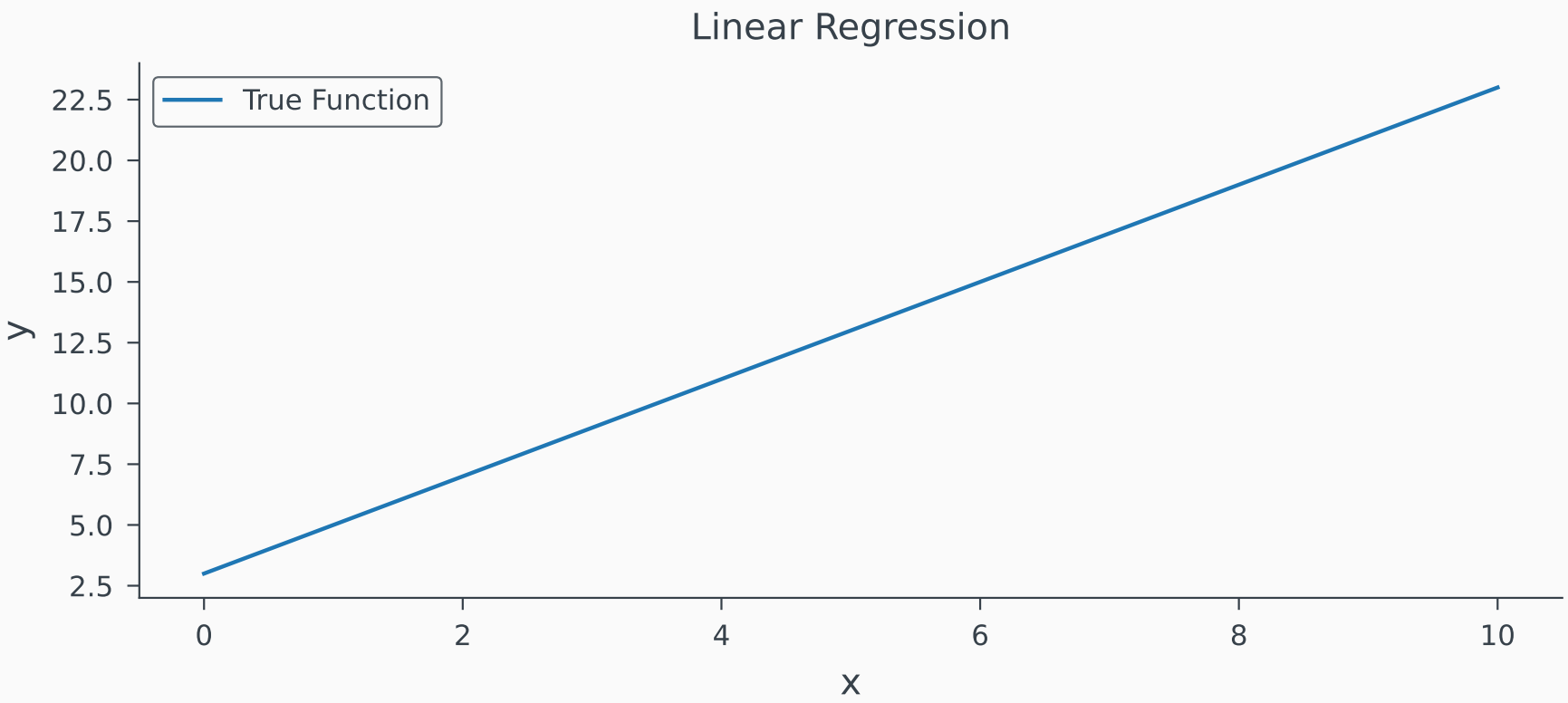
August 17, 2023

IIT Gandhinagar

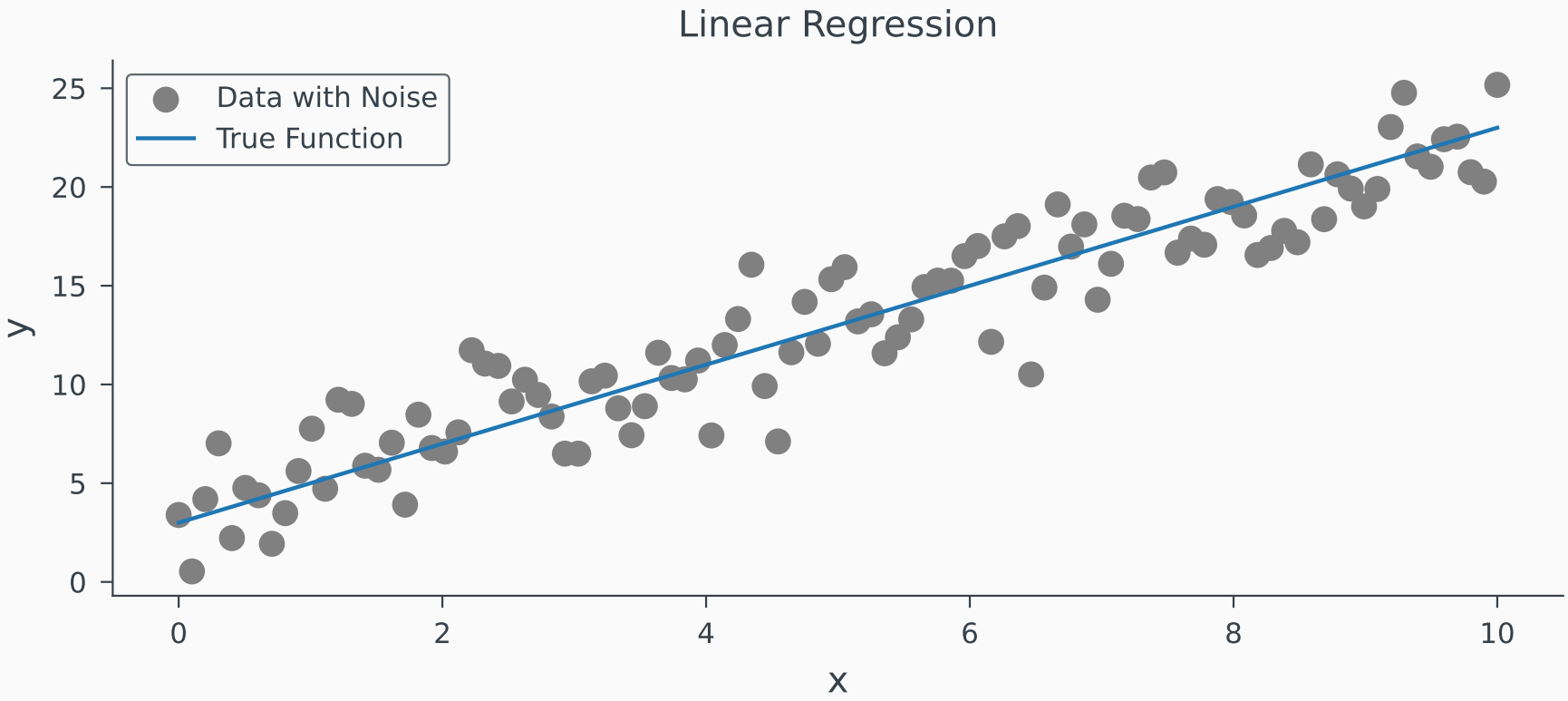
Agenda

MLE for Linear Regression

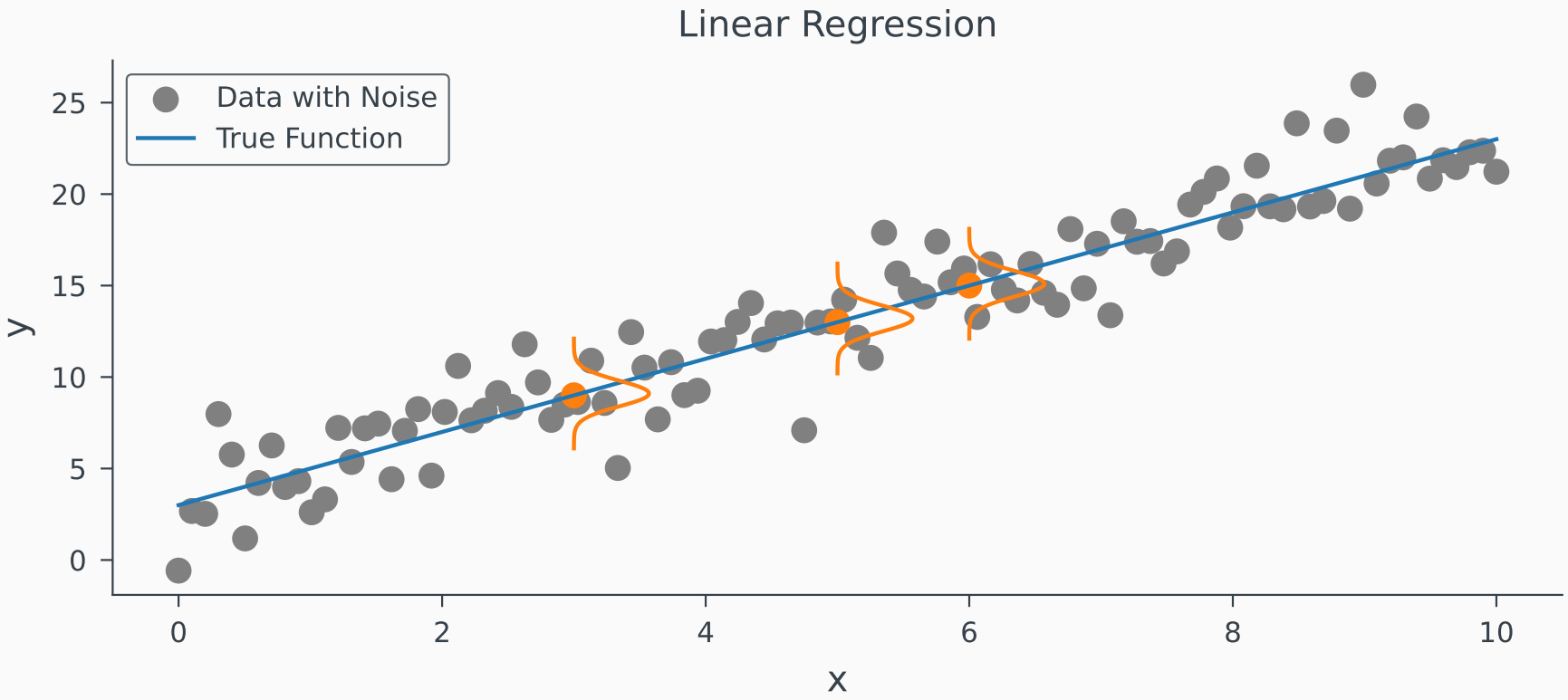
MLE for Linear Regression



MLE for Linear Regression



MLE for Linear Regression



Let us assume we have a dataset
$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where $x_i \in R^d, y_i \in R$.

Let us assume we have a dataset
$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where $x_i \in R^d, y_i \in R$.

We consider a regression problem with the likelihood function:
$p(y|x) = \mathbb{N}(y|f(x), \sigma^2)$.

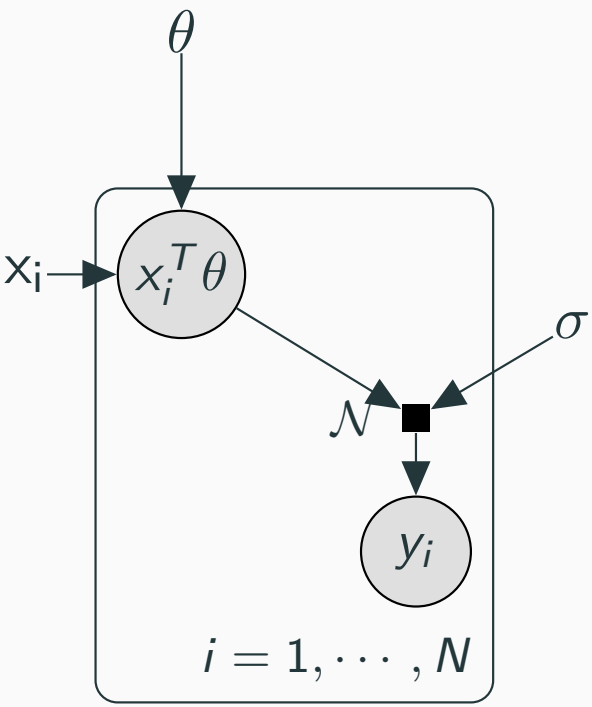
The functional relationship between x and y is given as $y = f(x) + \epsilon$ where $\epsilon \sim \mathbb{N}(0, \sigma^2)$.

The functional relationship between x and y is given as $y = f(\mathrm{x}) + \epsilon$ where $\epsilon \sim \mathbb{N}(0, \sigma^2)$.

where $f(\mathrm{x}) = x^T\theta$ for linear regression

The functional relationship between x and y is given as $y = f(\mathrm{x}) + \epsilon$ where $\epsilon \sim \mathbb{N}(0, \sigma^2)$.

where $f(\mathrm{x}) = x^T\theta$ for linear regression



Likelihood for linear regression

Likelihood is generally given as:

$$P(D|\theta) \tag{1}$$

Likelihood for linear regression

Likelihood is generally given as:

$$P(D|\theta) \tag{1}$$

Our data is: $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

Likelihood for linear regression

Likelihood is generally given as:

$$P(D|\theta) \tag{1}$$

Our data is: $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

Note: For purposes of computing likelihood, we assume that the input (x) is fixed and variation is only in the output (y).

Likelihood for linear regression

Likelihood is generally given as:

$$P(D|\theta) \tag{1}$$

Our data is: $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

Note: For purposes of computing likelihood, we assume that the input (x) is fixed and variation is only in the output (y).

Our likelihood function (Normal distribution) is given by:

$$P(\mathcal{Y}|\mathcal{X}, \theta) = p(y_1, \ldots, y_n|x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} p(y_i|x_i, \theta) \tag{2}$$

Likelihood for linear regression

The MLE equation is given by:

$$\theta_{MLE} \in \arg_\theta \max p(Y|X, \theta) \tag{3}$$

Likelihood for linear regression

The MLE equation is given by:

$$\theta_{MLE} \in \arg_\theta \max p(Y|X, \theta) \tag{3}$$

Maximizing the likelihood $\equiv$ Maximizing the log likelihood $\equiv$ Minimizing the negative log likelihood.

Likelihood for linear regression

The MLE equation is given by:

$$\theta_{MLE} \in \arg_\theta \max p(Y|X, \theta) \tag{3}$$

Maximizing the likelihood $\equiv$ Maximizing the log likelihood $\equiv$ Minimizing the negative log likelihood.

Taking the negative log, we get:

Likelihood for linear regression

The MLE equation is given by:

$$\theta_{MLE} \in \arg_\theta \max p(Y|X,\theta) \tag{3}$$

Maximizing the likelihood $\equiv$ Maximizing the log likelihood $\equiv$ Minimizing the negative log likelihood.

Taking the negative log, we get:

$$\begin{aligned} -\log p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) &= -\log \prod_{i=1}^{N} p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) \\ &= -\sum_{i=1}^{N} \log p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) \end{aligned}$$

Likelihood for linear regression

The MLE equation is given by:

$$\theta_{MLE} \in \arg_\theta \max p(Y|X,\theta) \tag{3}$$

Maximizing the likelihood $\equiv$ Maximizing the log likelihood $\equiv$ Minimizing the negative log likelihood.

Taking the negative log, we get:

$$\begin{aligned} -\log p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) &= -\log \prod_{i=1}^{N} p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) \\ &= -\sum_{i=1}^{N} \log p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) \end{aligned}$$

For a given point (x_i, y_i),

Likelihood for linear regression

The MLE equation is given by:

$$\theta_{MLE} \in \arg_\theta \max p(Y|X,\theta) \tag{3}$$

Maximizing the likelihood $\equiv$ Maximizing the log likelihood $\equiv$ Minimizing the negative log likelihood.

Taking the negative log, we get:

$$\begin{aligned} -\log p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) &= -\log \prod_{i=1}^{N} p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) \\ &= -\sum_{i=1}^{N} \log p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) \end{aligned}$$

For a given point (x_i, y_i),

$$-\log p(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) = \frac{1}{2\sigma^2}\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\theta}\right)^2 + \text{const}$$

Likelihood for linear regression

Thus the negative log likelihood is simplified to:

$$\begin{aligned}\mathcal{NLL}(\boldsymbol{\theta}) &:= \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - \boldsymbol{x}_i^\top\boldsymbol{\theta}\right)^2 \\ &= \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}) = \frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\|^2\end{aligned}$$

Likelihood for linear regression

Thus the negative log likelihood is simplified to:

$$\begin{aligned}\mathcal{NLL}(\boldsymbol{\theta}) &:= \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\theta}\right)^2 \\ &= \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) = \frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2\end{aligned}$$

Negative Log Likelihood for Linear Regression

NLL is proportional to:

$$\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2$$

Likelihood for linear regression

Thus the negative log likelihood is simplified to:

$$\begin{aligned}\mathcal{NLL}(\boldsymbol{\theta}) &:= \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(y_i - \boldsymbol{x}_i^\top\boldsymbol{\theta}\right)^2 \\ &= \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}) = \frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\|^2\end{aligned}$$

Negative Log Likelihood for Linear Regression

NLL is proportional to:

$$\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\|^2$$

This is the same as the squared error loss.

To minimize $NLL(\theta)$, we differentiate with respect to θ.

To minimize $NLL(\theta)$, we differentiate with respect to θ.

$$\theta = (X^T X)^{-1} X^T y \tag{4}$$

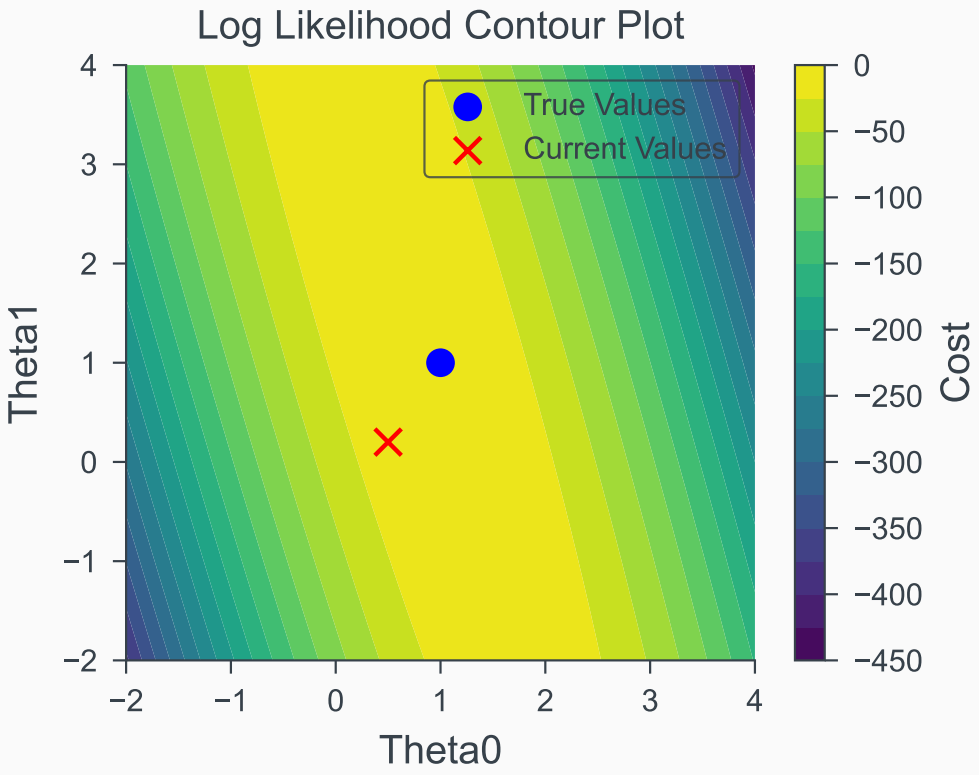
To minimize $NLL(\theta)$, we differentiate with respect to θ.

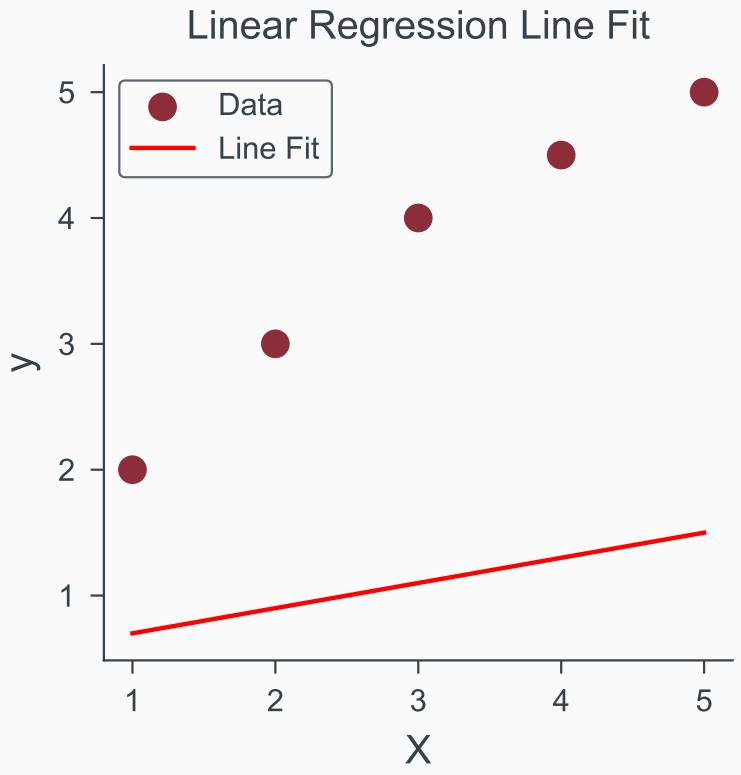
$$\theta = (X^T X)^{-1} X^T y \tag{4}$$

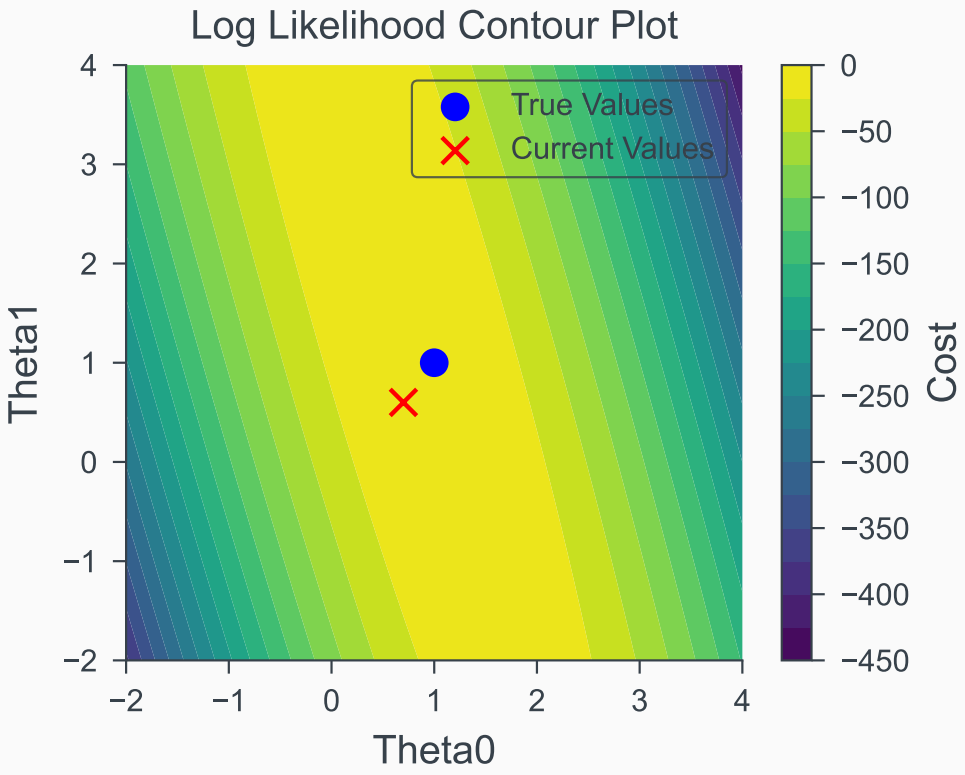
Maximum Likelihood Estimate for θ

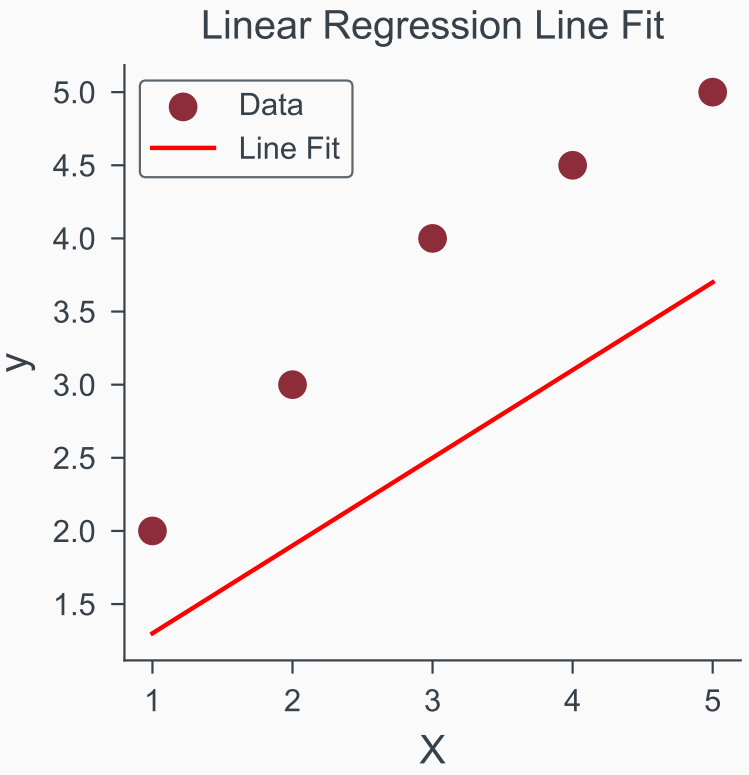
MLE of θ, denoted as $\hat{\theta}_{\text{MLE}}$, is given by:

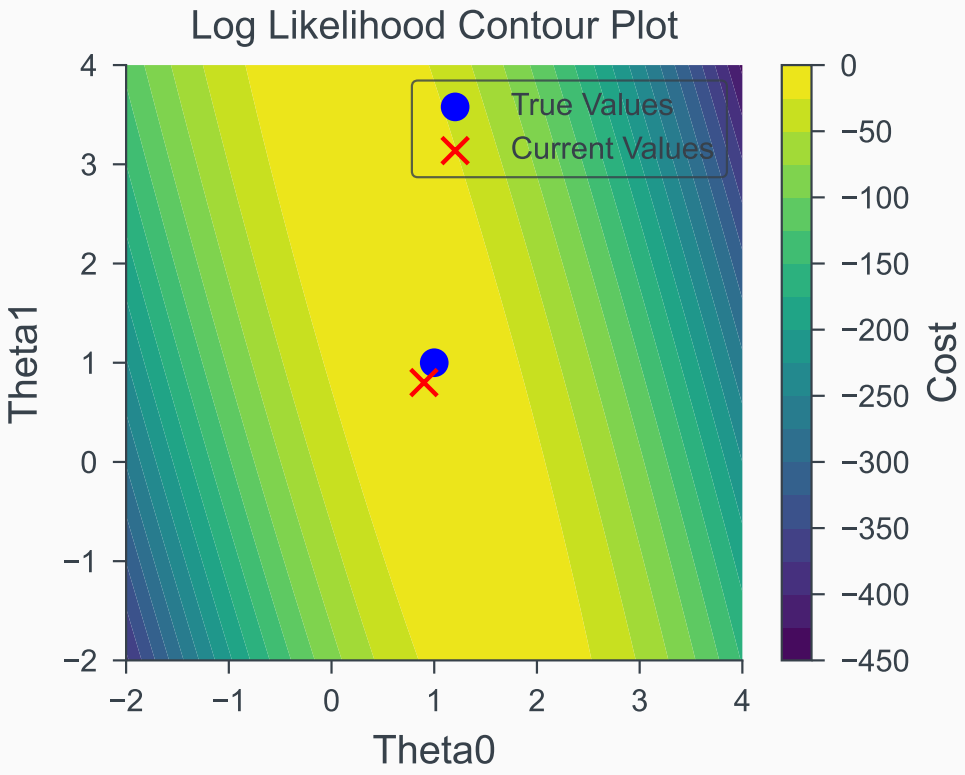
$$\hat{\theta}_{\text{MLE}} = (X^T X)^{-1} X^T y$$

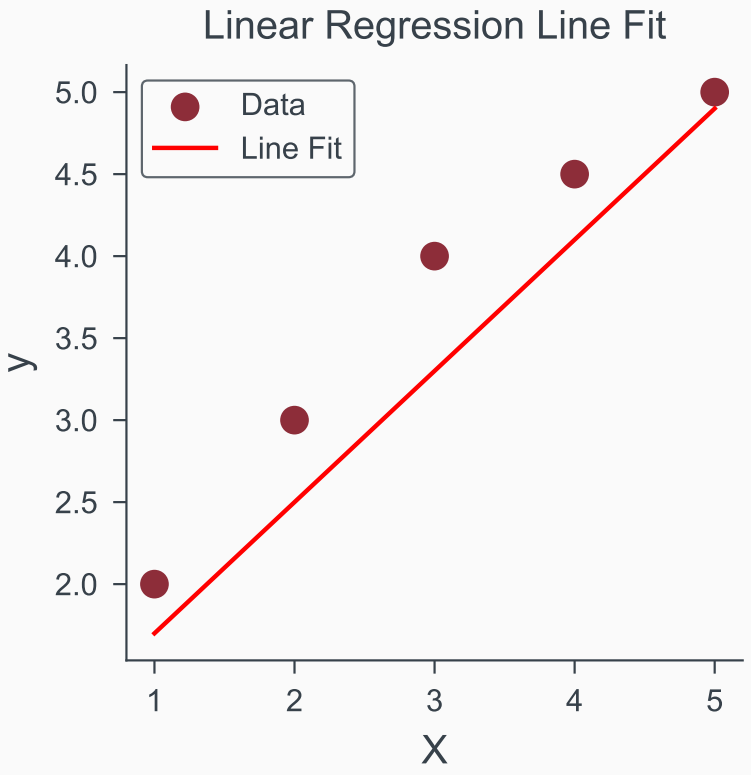

Log Likelihood Contour Plot
True Values
Current Values
Theta1
Theta0
Cost
4
3
2
1
0
−1
−2
−2
−1
0
1
2
3
4
0
−50
−100
−150
−200
−250
−300
−350
−400
−450


Linear Regression Line Fit
Data
Line Fit
y
X
5
4
3
2
1
1
2
3
4
5


Log Likelihood Contour Plot
True Values
Current Values
Theta1
Theta0
Cost
4
3
2
1
0
−1
−2
−2
−1
0
1
2
3
4
0
−50
−100
−150
−200
−250
−300
−350
−400
−450


Linear Regression Line Fit
Data
Line Fit
y
X
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1
2
3
4
5


Log Likelihood Contour Plot
True Values
Current Values
Theta1
Theta0
Cost
4
3
2
1
0
−1
−2
−2
−1
0
1
2
3
4
0
−50
−100
−150
−200
−250
−300
−350
−400
−450


Linear Regression Line Fit
Data
Line Fit
y
X
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1
2
3
4
5

MLE for Logistic Regression

Logistic Regression and Coin Toss

Coin Toss: We are given coin tosses: $D = \{y_1, y_2, \ldots, y_n\}$, where $y_i \in \{0, 1\}$.

Logistic Regression and Coin Toss

Coin Toss: We are given coin tosses: $D = \{y_1, y_2, \ldots, y_n\}$, where $y_i \in \{0, 1\}$.

Logistic regression: We are given a dataset:
$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where $x_i \in R^d, y_i \in \{0, 1\}$.

Logistic Regression and Coin Toss

Coin Toss: The probability of getting a head (class 1) is given by θ, i.e.

$$p(y = 1) = \theta$$

Logistic Regression and Coin Toss

Coin Toss: The probability of getting a head (class 1) is given by θ, i.e.

$$p(y = 1) = \theta$$

Logistic regression: The probability that a given input x belongs to class 1 is given by:

$$p(y = 1|x) = \sigma(x^T\theta)$$

Logistic Regression and Coin Toss

Coin Toss: We can say

$$y \sim \text{Bernoulli}(\theta)$$

Logistic Regression and Coin Toss

Coin Toss: We can say

$$y \sim \text{Bernoulli}(\theta)$$

Logistic regression: We can say

$$y \sim \text{Bernoulli}(\sigma(x^T\theta))$$

Logistic Regression and Coin Toss

Coin Toss: Likelihood is given by:

$$L(\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}$$

Logistic Regression and Coin Toss

Coin Toss: Likelihood is given by:

$$L(\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}$$

Logistic regression: Likewise, likelihood is given by:

$$L(\theta) = \prod_{i=1}^{n} \sigma(x_i^T\theta)^{y_i}(1-\sigma(x_i^T\theta))^{1-y_i}$$

Logistic Regression and Coin Toss

Coin Toss: Likelihood is given by:

$$L(\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}$$

Logistic Regression and Coin Toss

Coin Toss: Likelihood is given by:

$$L(\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}$$

Logistic regression: Likewise, likelihood is given by:

Logistic Regression and Coin Toss

Coin Toss: Likelihood is given by:

$$L(\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}$$

Logistic regression: Likewise, likelihood is given by: To simplify, we can write: $\hat{y}_i = \sigma(x_i^T \theta)$

Logistic Regression and Coin Toss

Coin Toss: Likelihood is given by:

$$L(\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}$$

Logistic regression: Likewise, likelihood is given by: To simplify, we can write: $\hat{y}_i = \sigma(x_i^T\theta)$ Thus, likelihood is given by:

$$L(\theta) = \prod_{i=1}^{n} \hat{y}_i^{y_i}(1-\hat{y}_i)^{1-y_i}$$

Logistic Regression and Coin Toss

Coin Toss: Log likelihood is given by:

$$\log(L(\theta)) = \sum_{i=1}^{n} y_i \log(\theta) + (1 - y_i) \log(1 - \theta)$$

Logistic Regression and Coin Toss

Coin Toss: Log likelihood is given by:

$$\log(L(\theta)) = \sum_{i=1}^{n} y_i \log(\theta) + (1 - y_i) \log(1 - \theta)$$

Logistic regression: Likewise, log likelihood is given by:

$$\log(L(\theta)) = \sum_{i=1}^{n} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)$$

Negative Log Likelihood for Logistic Regression

Negative Log Likelihood for Logistic Regression

NLL is proportional to:

$$-\sum_{i=1}^{n} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)$$

which is the same as the binary cross entropy loss function.

Extending binary classification to K-class classification

For binary classification, we have a sigmoid function.
$p(y = 1|x) = \sigma(x^T\theta)$

Extending binary classification to K-class classification

For binary classification, we have a sigmoid function.

$p(y = 1|x) = \sigma(x^T\theta)$

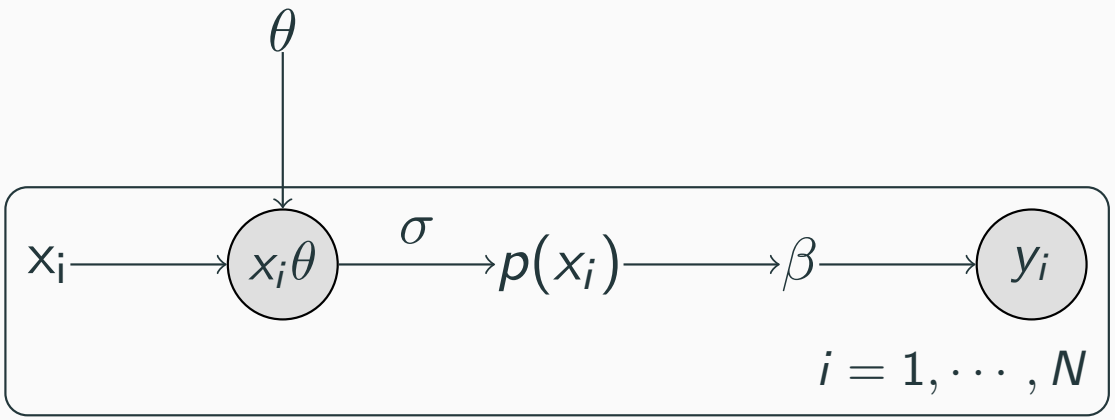
For multi-class classification, we have a softmax function.

$p(y = i|x) = \frac{e^{x^T\theta_i}}{\sum_{j=1}^{K} e^{x^T\theta_j}}$

Extending binary classification to multi-class classification

Self-Study: Derive the negative log likelihood for multi-class classification and show that it is the same as the cross entropy loss function.

MLE for Logistic Regression



Binary Classification:

The probability distribution in case of Logistic Regression considering two classes is Bernoulli distribution but there is a slight difference. The probability is now the output of the logistic function. Parameters are $\theta = [\theta_0, \theta_1]$.

$$p = P(Y = 1|X) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 X)}} \tag{5}$$

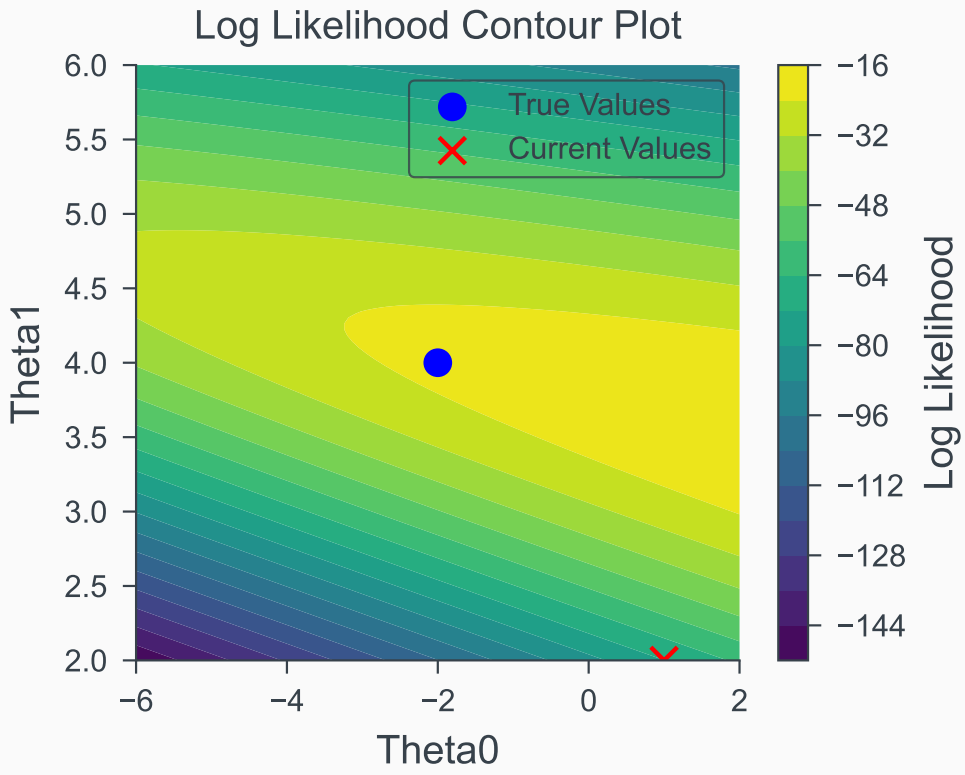
Rewriting the likelihood in this manner:

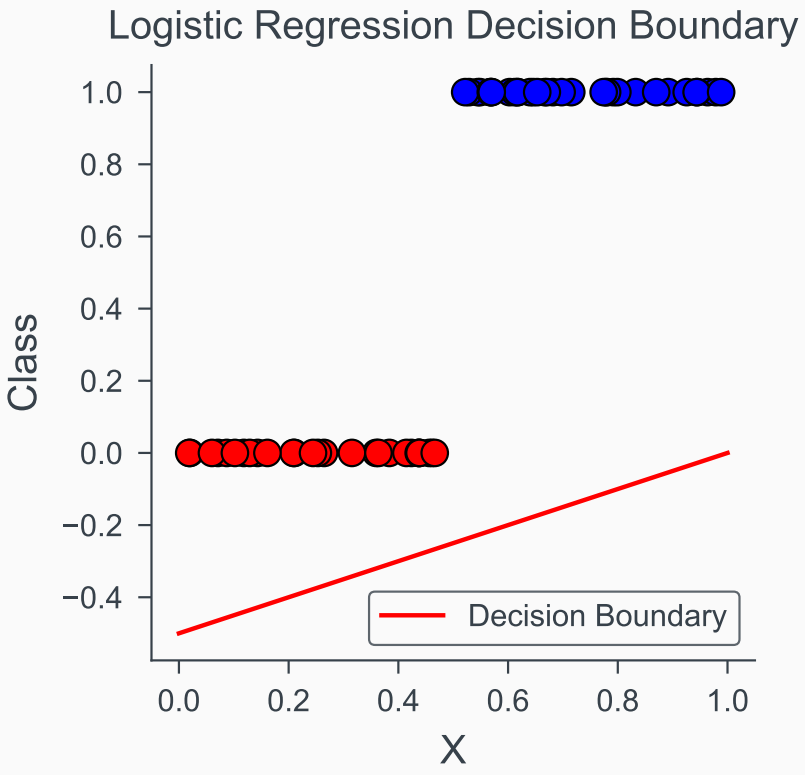
$$\begin{aligned} L(\theta) &= \prod_{y_i=1} p(x_i) \prod_{y_i=0} (1 - p(x_i)) \\ &= \prod \left(p(x_i)^{y_i} (1 - p(x_i))^{1-y_i} \right) \end{aligned}$$

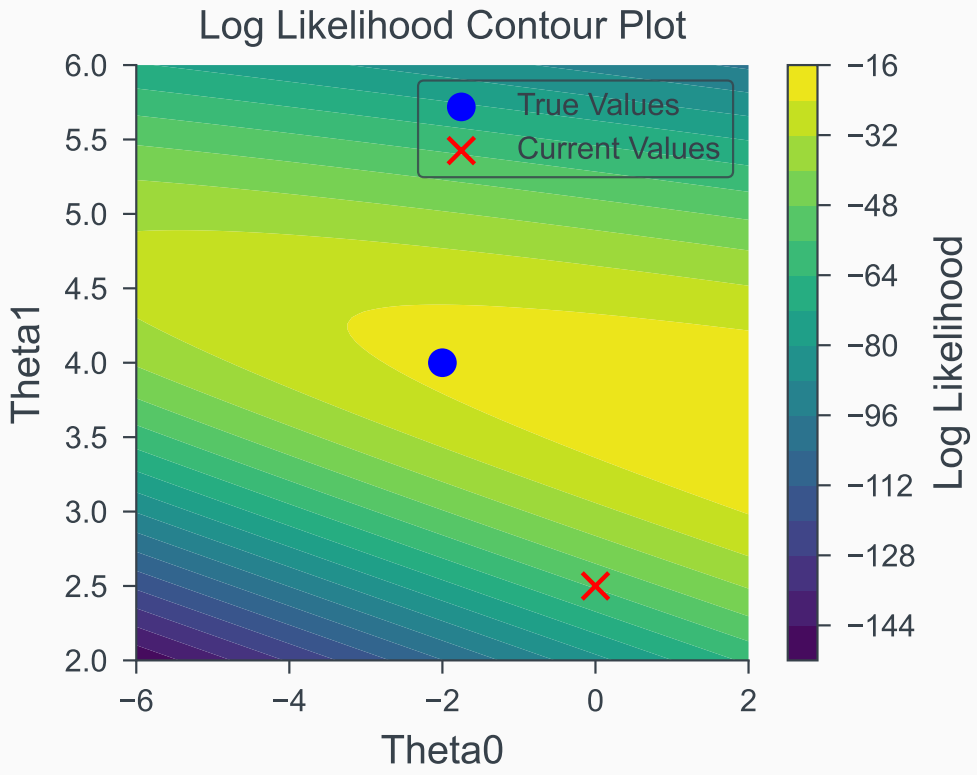
Taking log on both sides:

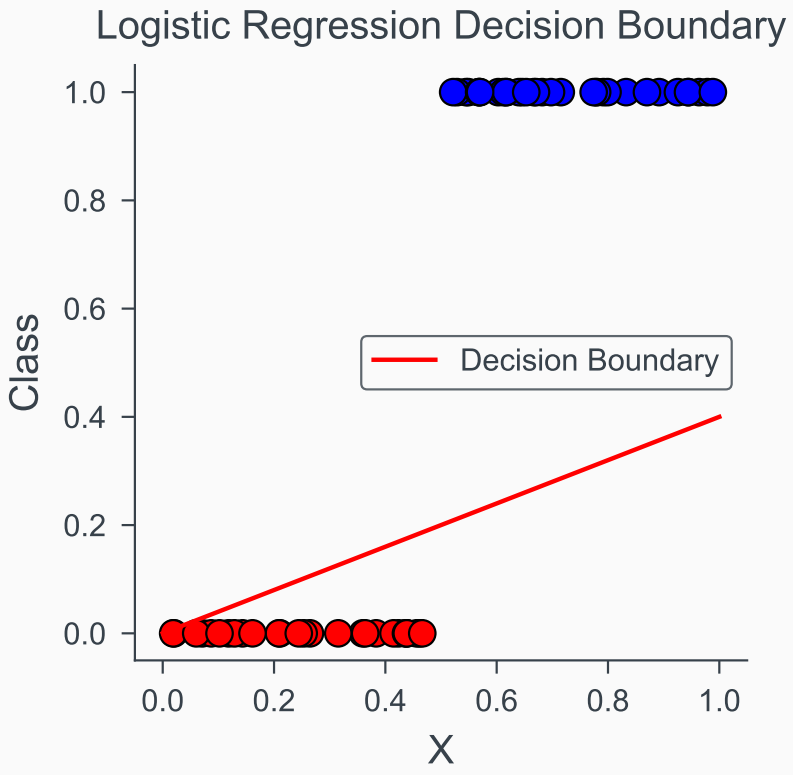
$$\log(L(\theta)) = \sum_{i=1}^{n} y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i))$$

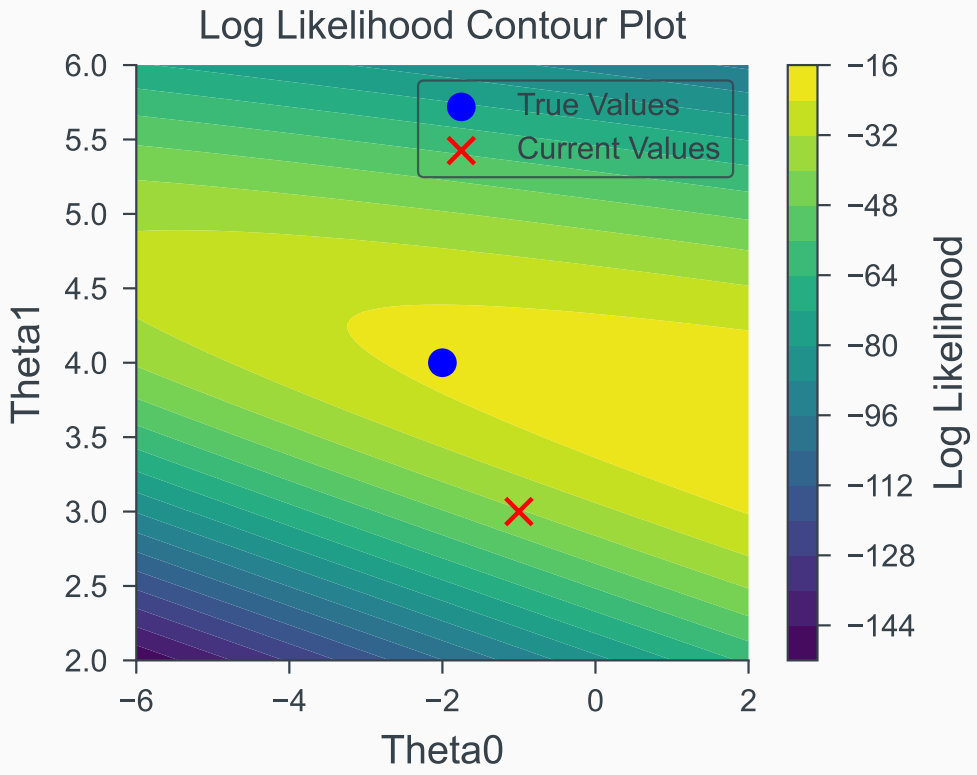
If we multiply this by $-\frac{1}{n}$,this is nothing but the binary cross entropy loss function!

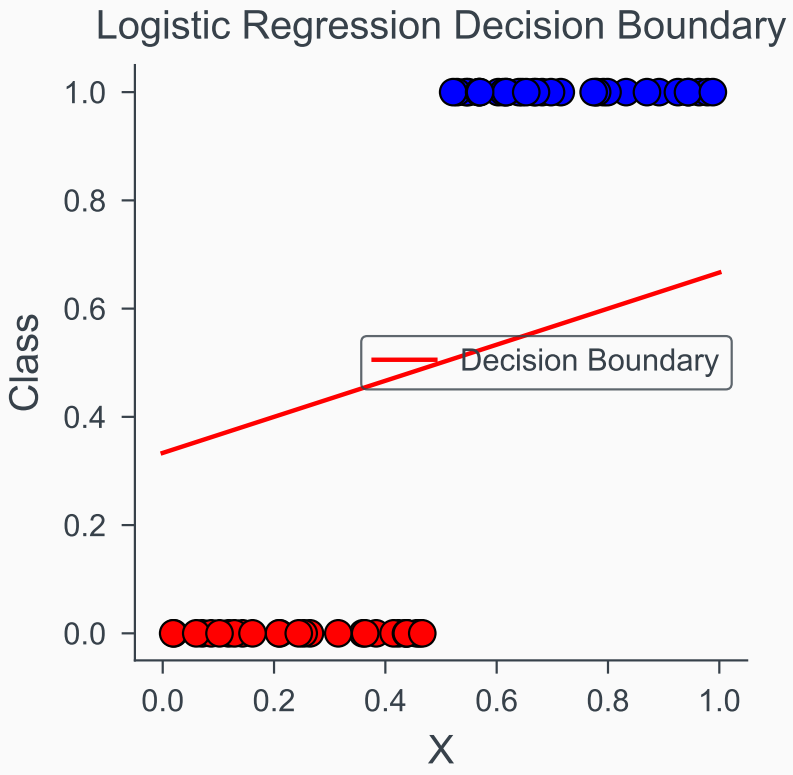

Log Likelihood Contour Plot
True Values
Current Values
Theta1
Theta0
Log Likelihood
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
−6
−4
−2
0
2
−16
−32
−48
−64
−80
−96
−112
−128
−144


Logistic Regression Decision Boundary
Class
X
Decision Boundary
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0.0
0.2
0.4
0.6
0.8
1.0


Log Likelihood Contour Plot
True Values
Current Values
Theta1
Theta0
Log Likelihood
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
−6
−4
−2
0
2
−16
−32
−48
−64
−80
−96
−112
−128
−144


Logistic Regression Decision Boundary
Decision Boundary
Class
X
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


Log Likelihood Contour Plot
True Values
Current Values
Theta1
Theta0
Log Likelihood
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
−6
−4
−2
0
2
−16
−32
−48
−64
−80
−96
−112
−128
−144


Logistic Regression Decision Boundary
Decision Boundary
Class
X
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

Coin toss V/s Binary Logistic Regression

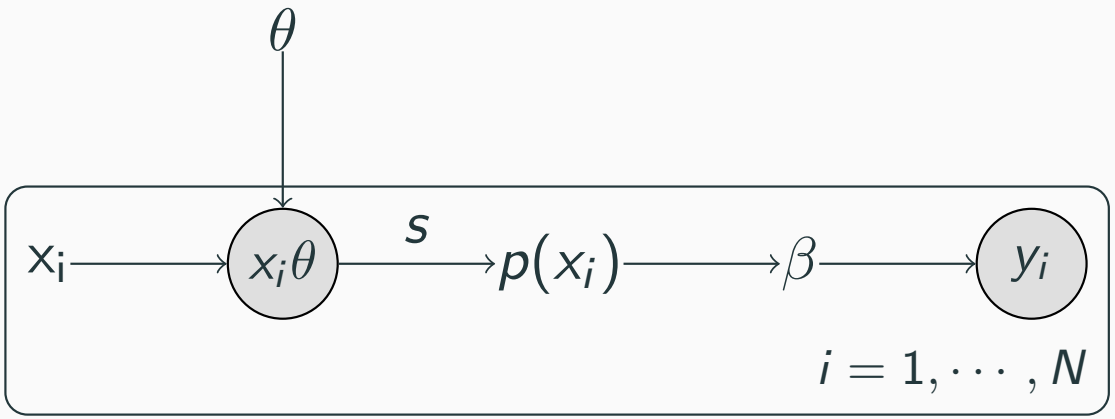
Coin toss	Binary Logistic Regression
Likelihood$=\theta^t(1-\theta)^{1-t}$	Likelihood$=(\sigma(x^T\theta))^t(1-\sigma(x^T\theta))^{1-t}$

Coin toss V/s Binary Logistic Regression

Coin toss	Binary Logistic Regression
Likelihood$=\theta^t(1-\theta)^{1-t}$	Likelihood$=(\sigma(x^T\theta))^t(1-\sigma(x^T\theta))^{1-t}$
Outcome is Head/Tail	Outcome is out of two possible classes

Coin toss V/s Binary Logistic Regression

Coin toss	Binary Logistic Regression
Likelihood$=\theta^t(1-\theta)^{1-t}$	Likelihood$=(\sigma(x^T\theta))^t(1-\sigma(x^T\theta))^{1-t}$
Outcome is Head/Tail	Outcome is out of two possible classes
Parameter is scalar	Parameter is vector with two values



Multi-class Classification:

The probability distribution in case of Logistic Regression considering more than two classes is Categorical distribution. The probability is now the output of the softmax function. Parameters are $\theta = [\theta_0, \theta_1, \ldots, \theta_k]$.

$$p = P(Y = i|X) = \frac{e^{\theta x_i}}{\sum_{j=1}^{n} e^{\theta x_j}} \tag{6}$$

Now:

$$L(\theta) = \prod_{i=1}^{n} \prod_{j=1}^{K} p^j(x_i)$$

Taking log on both sides:

$$\log(L(\theta)) = \sum_{i=1}^{n} \sum_{j=1}^{K} y_i^k \log(p^k(x_i))$$

If we multiply this by $-\frac{1}{n}$,this is nothing but the cross entropy loss function!

Now if we differentiate this wrt θ, it is difficult to find a analytical solution with it. Thus in order to solve for MLE for logistic regression, methods like Gradient Descent, Newton-Raphson, etc. are used. For example through Gradient descent, the below decision boundary i.e. θ has been calculated.

Binary V/s Multiclass Logistic Regression

Binary Logistic Regression	Multiclass Logistic Regression
Binary Cross Entropy Loss	Cross Entropy Loss

Binary V/s Multiclass Logistic Regression

Binary Logistic Regression	Multiclass Logistic Regression
Binary Cross Entropy Loss	Cross Entropy Loss
$p(x) = \sigma(x^T\theta)$	$p(x) = s(x^T\theta)$

Binary V/s Multiclass Logistic Regression

Binary Logistic Regression	**Multiclass Logistic Regression**
Binary Cross Entropy Loss	Cross Entropy Loss
$p(x) = \sigma(x^T\theta)$	$p(x) = s(x^T\theta)$
Bernoulli Likelihood	Categorical Likelihood