Maximum Likelihood Estimation for Linear and Logistic Regression

Nipun Batra

August 23, 2023

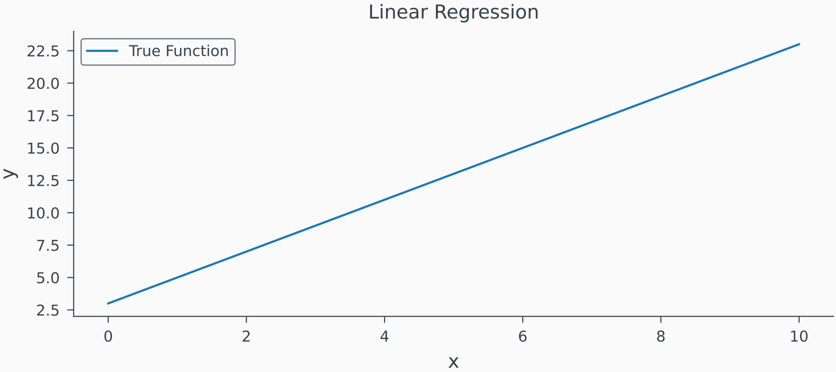
IIT Gandhinagar

Agenda

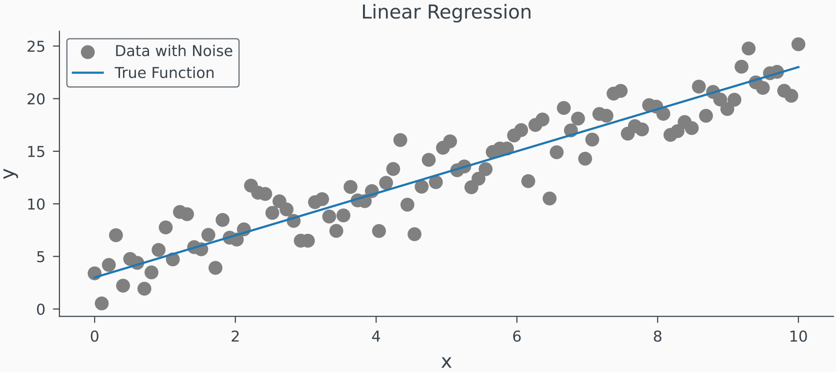
MLE for Linear Regression

MLE for Linear Regression

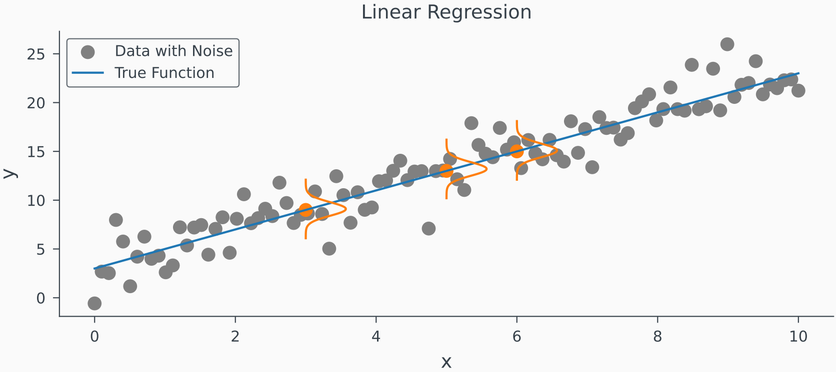
MLE for Linear Regression



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
$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, where $x_i \in R^d, y_i \in R$.

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

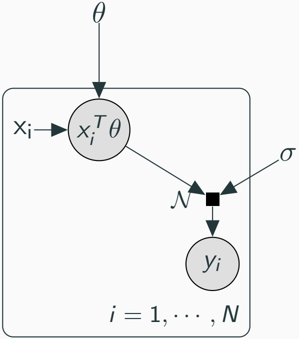
The functional relationship between x and y is given as $y = f(\mathrm{x}) + \epsilon$ where $\epsilon \sim \mathbb{N}(0, \sigma^2)$.

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:

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

Notebook: log-likelihood-linreg.ipynb

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.

Notebook: log-likelihood-linreg.ipynb

Multi-class classification

Self Study Notebook on Categorical distribution:
distributions.ipynb