Machine Learning and Intelligent Systems

Linear Models for Regression - Part 2

Maria A. Zuluaga

October 13, 2023

EURECOM - Data Science Department

Table of contents

- Maximum Likelihood Estimation
  - MLE HOWTO
  - Exercise: Coin Toss Example and MLE
- Solving Linear Regression with MLE
  - Estimate $\hat{\mathbf{w}}$ using MLE
  - Matrix Notation
  - Solving for $\mathbf{w}$
- Solution to the 100m Olympic Games Problem
- The role of σ^2
- Recap

Maximum Likelihood Estimation

Intuition: Simple Coin Toss

- Suppose you find a coin.
- You ask yourself, ''What is the probability that this coin comes up heads when I toss it?''

[1] Adapted and inspired from K. Weinberger's course (Cornell University)

Intuition: Simple Coin Toss

- Suppose you find a coin.
- You ask yourself, "What is the probability that this coin comes up heads when I toss it?" You toss it $n = 10$ times and obtain the following sequence of outcomes:

$$D = \{H, T, T, H, H, H, T, T, T, T\}.$$

- Based on these samples, how would you estimate $P(H)$?
- We observed $n_H = 4$ heads and $n_T = 6$ tails. So, intuitively,

$$P(H) \approx \frac{n_H}{n_H + n_T} = \frac{4}{10} = 0.4$$

[1] Adapted and inspired from K. Weinberger's course (Cornell University)

Intuition: Simple Coin Toss

- Suppose you find a coin.
- You ask yourself, "What is the probability that this coin comes up heads when I toss it?" You toss it $n = 10$ times and obtain the following sequence of outcomes:

$$D = \{H, T, T, H, H, H, T, T, T, T\}.$$

- Based on these samples, how would you estimate $P(H)$?
- We observed $n_H = 4$ heads and $n_T = 6$ tails. So, intuitively,

$$P(H) \approx \frac{n_H}{n_H + n_T} = \frac{4}{10} = 0.4$$

Can we derive this more formally?[1]

[1] Adapted and inspired from K. Weinberger's course (Cornell University)

The Maximum Likelihood Estimator (MLE)

The estimation process we just performed is nothing else than the Maximum Likelihood Estimate (MLE). For MLE, you typically proceed in two steps:

1. You make an explicit modeling assumption about what type of distribution your data was sampled from.
2. You set the parameters of this distribution so that the data you observed is as likely as possible.

The Maximum Likelihood Estimator (MLE)

The estimation process we just performed is nothing else than the Maximum Likelihood Estimate (MLE). For MLE, you typically proceed in two steps:

1. You make an explicit modeling assumption about what type of distribution your data was sampled from.
2. You set the parameters of this distribution so that the data you observed is as likely as possible.

Coin Toss example:

The Maximum Likelihood Estimator (MLE)

The estimation process we just performed is nothing else than the Maximum Likelihood Estimate (MLE). For MLE, you typically proceed in two steps:

1. You make an explicit modeling assumption about what type of distribution your data was sampled from.
2. You set the parameters of this distribution so that the data you observed is as likely as possible.

Coin Toss example:

1. The observed outcomes of a coin toss follow a binomial distribution. It has two parameters n and θ and it captures the distribution of n independent binary random events that have a positive outcome with probability θ. n is the number of tosses and θ the probability of having heads $P(H) = \theta$

The Maximum Likelihood Estimator (MLE)

The estimation process we just performed is nothing else than the Maximum Likelihood Estimate (MLE). For MLE, you typically proceed in two steps:

1. You make an explicit modeling assumption about what type of distribution your data was sampled from.
2. You set the parameters of this distribution so that the data you observed is as likely as possible.

Coin Toss example:

1. The observed outcomes of a coin toss follow a binomial distribution. It has two parameters n and θ and it captures the distribution of n independent binary random events that have a positive outcome with probability θ. n is the number of tosses and θ the probability of having heads $P(H) = \theta$
2. **We need to find** $\hat{\theta}$ given our observed data D

The Maximum Likelihood Estimator (MLE)

- Let $Z_1, Z_2, \ldots, Z_N$ be a random sample of *iid* random variables with joint PDF $p_\theta(z)$ that depends on θ.

The Maximum Likelihood Estimator (MLE)

- Let $Z_1, Z_2, \ldots, Z_N$ be a random sample of *iid* random variables with joint PDF $p_\theta(z)$ that depends on θ.
- The joint PDF is:

$$p_\theta(z_1, z_2, \ldots, z_N) = p(z_1, z_2, \ldots, z_N|\theta)$$

The Maximum Likelihood Estimator (MLE)

- Let $Z_1, Z_2, \ldots, Z_N$ be a random sample of *iid* random variables with joint PDF $p_\theta(z)$ that depends on θ.
- The joint PDF is:

$$\begin{aligned} p_\theta(z_1, z_2, \ldots, z_N) &= p(z_1, z_2, \ldots, z_N|\theta) \\ &= p(z_1;\theta)p(z_2;\theta)\ldots p(z_N|\theta) \end{aligned}$$

The Maximum Likelihood Estimator (MLE)

- Let $Z_1, Z_2, \ldots, Z_N$ be a random sample of *iid* random variables with joint PDF $p_\theta(z)$ that depends on θ.
- The joint PDF is:

$$\begin{aligned} p_\theta(z_1, z_2, \ldots, z_N) &= p(z_1, z_2, \ldots, z_N|\theta) \\ &= p(z_1; \theta)p(z_2; \theta) \ldots p(z_N|\theta) \\ &= \prod_{i=1}^{N} p(z_i|\theta) \end{aligned}$$

The Maximum Likelihood Estimator (MLE)

- Upon observing the data, $p(z_1, z_2, \ldots, z_N|\theta)$ becomes a function of θ alone.

The Maximum Likelihood Estimator (MLE)

- Upon observing the data, $p(z_1, z_2, \ldots, z_N|\theta)$ becomes a function of θ alone.
- Considering $p(z_1, z_2, \ldots, z_N|\theta)$as a function of θ, we write:

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} p(z_i|\theta)$$

The Maximum Likelihood Estimator (MLE)

- Upon observing the data, $p(z_1, z_2, \ldots, z_N|\theta)$ becomes a function of θ alone.
- Considering $p(z_1, z_2, \ldots, z_N|\theta)$as a function of θ, we write:

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} p(z_i|\theta)$$

- **In words:** $\mathcal{L}(\theta)$ is the likelihood of observing the given data given θ.
- We will denote $\mathcal{L}(\theta)$ the **likelihood** function.

The Maximum Likelihood Estimator (MLE)

- Upon observing the data, $p(z_1, z_2, \ldots, z_N|\theta)$ becomes a function of θ alone.
- Considering $p(z_1, z_2, \ldots, z_N|\theta)$as a function of θ, we write:

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} p(z_i|\theta)$$

- **In words:** $\mathcal{L}(\theta)$ is the likelihood of observing the given data given θ.
- We will denote $\mathcal{L}(\theta)$ the **likelihood** function.

Definition

The Maximum Likelihood Estimator of θ (MLE) is the value $\hat{\theta}$ that maximizes the likelihood. It is the value that makes the data the most "probable".

The Maximum Likelihood Estimator (MLE)

- Finding $\hat{\theta}$ that maximizes the likelihood of the data $p(Z|\theta)$ accounts to:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \prod_{i=1}^{N} p(z_i|\theta)$$

The Maximum Likelihood Estimator (MLE)

- Finding $\hat{\theta}$ that maximizes the likelihood of the data $p(Z|\theta)$ accounts to:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \prod_{i=1}^{N} p(z_i|\theta)$$

- Rather than maximizing this product, which can be complex, we can use the fact that the logarithm is a monotonic function so it will be equivalent to maximize the **log likelihood**:

The Maximum Likelihood Estimator (MLE)

- Finding $\hat{\theta}$ that maximizes the likelihood of the data $p(Z|\theta)$ accounts to:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \prod_{i=1}^{N} p(z_i|\theta)$$

- Rather than maximizing this product, which can be complex, we can use the fact that the logarithm is a monotonic function so it will be equivalent to maximize the **log likelihood**:

$$\ell(\theta) = \sum_{i=1}^{N} \log p(z_i|\theta)$$

The Maximum Likelihood Estimator (MLE)

- Finding $\hat{\theta}$ that maximizes the likelihood of the data $p(Z|\theta)$ accounts to:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \prod_{i=1}^{N} p(z_i|\theta)$$

- Rather than maximizing this product, which can be complex, we can use the fact that the logarithm is a monotonic function so it will be equivalent to maximize the **log likelihood**:

$$\ell(\theta) = \sum_{i=1}^{N} \log p(z_i|\theta)$$

- Replacing accordingly:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \sum_{i=1}^{N} \log p(z_i|\theta) \tag{1}$$

Estimating $\hat{\theta}$: HOWTO

1. Plug in all the terms for the distribution in Eq. 1

Estimating $\hat{\theta}$: HOWTO

1. Plug in all the terms for the distribution in Eq. 1
2. Take the log of the function.

Estimating $\hat{\theta}$: HOWTO

1. Plug in all the terms for the distribution in Eq. 1
2. Take the log of the function.
3. Compute its derivative, and equate it with zero to find an extreme point.

Estimating $\hat{\theta}$: HOWTO

1. Plug in all the terms for the distribution in Eq. 1
2. Take the log of the function.
3. Compute its derivative, and equate it with zero to find an extreme point.
4. (Optional) To be precise, verify that it is a maximum and not a minimum, by verifying that the second derivative is negative.

Exercise: Coin Toss Example and MLE

Given that the binomial distribution is denoted as:

$$p(z;\theta) = \begin{pmatrix} n_H + n_T \\ n_H \end{pmatrix} \theta^{n_H} (1-\theta)^{n_T}$$

apply the MLE to find an expression for $\hat{\theta}$.

Solving Linear Regression with MLE

Estimating $\hat{\mathbf{w}}$ using MLE

Let us recall our assumption about the distribution of $\mathbf{y}$:

$$p(y_i | \mathbf{x}_i; \mathbf{w}, \boldsymbol{\sigma}^2) = \frac{1}{\boldsymbol{\sigma}\sqrt{2\pi}} \exp\left\{ -\frac{\left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2}{2\boldsymbol{\sigma}^2} \right\}$$

We can use the MLE to estimate $\hat{\mathbf{w}}$:

Estimating $\hat{\mathbf{w}}$ using MLE

Let us recall our assumption about the distribution of $\mathbf{y}$:

$$p(y_i | \mathbf{x}_i; \mathbf{w}, \boldsymbol{\sigma}^2) = \frac{1}{\boldsymbol{\sigma}\sqrt{2\pi}} \exp\left\{ -\frac{\left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2}{2\boldsymbol{\sigma}^2} \right\}$$

We can use the MLE to estimate $\hat{\mathbf{w}}$:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \prod_{i=1}^{N} p(y_i | \mathbf{x}_i; \mathbf{w}, \boldsymbol{\sigma}^2)$$

Derivation

Following the steps of the MLE HOWTO (slide 7):

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \prod_{i=1}^{N} p(y_i | \mathbf{x}_i; \mathbf{w}, \sigma)$$

Derivation

Following the steps of the MLE HOWTO (slide 7):

$$\begin{aligned}\hat{\mathbf{w}} &= \arg\max_{\mathbf{w}} \prod_{i=1}^{N} p(y_i|\mathbf{x}_i; \mathbf{w}, \boldsymbol{\sigma}) \\ &= \arg\max_{\mathbf{w}} \prod_{i=1}^{N} \frac{1}{\boldsymbol{\sigma}\sqrt{2\pi}} \exp\left\{-\frac{\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2}{2\boldsymbol{\sigma}^2}\right\}\end{aligned}$$

Derivation

Following the steps of the MLE HOWTO (slide 7):

$$\begin{aligned}
\hat{\mathbf{w}} &= \arg\max_{\mathbf{w}} \prod_{i=1}^{N} p(y_i | \mathbf{x}_i; \mathbf{w}, \sigma) \\
&= \arg\max_{\mathbf{w}} \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2}{2\sigma^2}\right\} \\
&= \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \log\left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2}{2\sigma^2}\right\}\right)
\end{aligned}$$

Derivation

Following the steps of the MLE HOWTO (slide 7):

$$
\begin{aligned}
\hat{\mathbf{w}} &= \arg\max_{\mathbf{w}} \prod_{i=1}^{N} p(y_i | \mathbf{x}_i; \mathbf{w}, \sigma) \\
&= \arg\max_{\mathbf{w}} \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{\left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2}{2\sigma^2}\right\} \\
&= \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \log\left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{\left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2}{2\sigma^2}\right\}\right) \\
&= \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \log\left(\exp\left\{-\frac{\left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2}{2\sigma^2}\right\}\right)
\end{aligned}
$$

Derivation

Following the steps of the MLE HOWTO (slide 7):

$$
\begin{aligned}
\hat{\mathbf{w}} &= \arg\max_{\mathbf{w}} \prod_{i=1}^{N} p(y_i | \mathbf{x}_i; \mathbf{w}, \sigma) \\
&= \arg\max_{\mathbf{w}} \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{\left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2}{2\sigma^2} \right\} \\
&= \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \log\left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{\left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2}{2\sigma^2} \right\} \right) \\
&= \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \log\left(\frac{1}{\sigma\sqrt{2\pi}} \right) + \log\left(\exp\left\{ -\frac{\left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2}{2\sigma^2} \right\} \right) \\
&= \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \xcancel{\log\left(\frac{1}{\sigma\sqrt{2\pi}} \right)} - \frac{\left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2}{2\sigma^2}
\end{aligned}
$$

Derivation (cont)

(cont)

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \cancel{\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)} - \frac{\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2}{2\sigma^2}$$

Derivation (cont)

(cont)

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \cancel{\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)} - \frac{\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2}{2\sigma^2}$$

$$= \arg\max_{\mathbf{w}} -\sum_{i=1}^{N} \frac{\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2}{2\sigma^2}$$

Derivation (cont)

(cont)

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \cancel{\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)} - \frac{\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2}{2\sigma^2}$$

$$= \arg\max_{\mathbf{w}} \; - \sum_{i=1}^{N} \frac{\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2}{2\sigma^2}$$

$$= \arg\min_{\mathbf{w}} \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2$$

Derivation (cont)

(cont)

$$
\begin{aligned}
\hat{\mathbf{w}} &= \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \cancel{\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)} - \frac{\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2}{2\sigma^2} \\
&= \arg\max_{\mathbf{w}} \; -\sum_{i=1}^{N} \frac{\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2}{2\sigma^2} \\
&= \arg\min_{\mathbf{w}} \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2 \\
&= \arg\min_{\mathbf{w}} \cancel{\frac{1}{2\sigma^2}} \sum_{i=1}^{N} \left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2
\end{aligned}
$$

Derivation (cont)

(cont)

$$\begin{aligned}
\hat{\mathbf{w}} &= \arg\max_{\mathbf{w}} \sum_{i=1}^{N} \cancel{\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)} - \frac{\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2}{2\sigma^2} \\
&= \arg\max_{\mathbf{w}} -\sum_{i=1}^{N} \frac{\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2}{2\sigma^2} \\
&= \arg\min_{\mathbf{w}} \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2 \\
&= \arg\min_{\mathbf{w}} \cancel{\frac{1}{2\sigma^2}} \sum_{i=1}^{N} \left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2 \\
&= \arg\min_{\mathbf{w}} \sum_{i=1}^{N} \left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2
\end{aligned}$$

Estimating $\hat{\mathbf{w}}$ using MLE

The final expression for $\hat{\mathbf{w}}$ is:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{i=1}^{N} \left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2$$

Does this look familiar?

Estimating $\hat{\mathbf{w}}$ using MLE

The final expression for $\hat{\mathbf{w}}$ is:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \sum_{i=1}^{N} \left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2$$

Does this look familiar?

A/ It is the quadratic loss function, *aka* squared loss, (lecture 1):

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathcal{L} = \arg\min_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2$$

Matrix Notation

- $\mathbf{x}$, y and $\mathbf{w}$ can have large dimensions

Matrix Notation

- $\mathbf{x}$, y and $\mathbf{w}$ can have large dimensions
- The current notation can be cumbersome to handle

Matrix Notation

- $\mathbf{x}$, y and $\mathbf{w}$ can have large dimensions
- The current notation can be cumbersome to handle
- We will favor the use of matrix notation:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \cdots \\ \cdots \\ y_N \end{bmatrix} \qquad \mathbf{w} = \begin{bmatrix} w_0 \\ \cdots \\ w_D \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1D} \\ 1 & x_{21} & \cdots & x_{2D} \\ 1 & x_{31} & \cdots & x_{3D} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{ND} \end{bmatrix}$$

Estimating $\hat{\mathbf{w}}$ with Matrix Notation

Using the matrix notation, the expression we had obtained for $\hat{\mathbf{w}}$ becomes:

$$\arg\min_{\mathbf{w}} \frac{1}{N} (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w})$$

Following the MLE HOWTO, now we need to solve for:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} \left(\frac{1}{N} (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) \right)$$

Matrix Derivatives Cheat Sheet[2]

Matrix/vector manipulation

All bold capitals are matrices, bold lowercase are vectors.

Rule	Comments
$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$	order is reversed, everything is transposed
$(\mathbf{a}^T\mathbf{Bc})^T = \mathbf{c}^T\mathbf{B}^T\mathbf{a}$	as above
$\mathbf{a}^T\mathbf{b} = \mathbf{b}^T\mathbf{a}$	(the result is a scalar, and the transpose of a scalar is itself)
$(\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC}+\mathbf{BC}$	multiplication is distributive
$(\mathbf{a}+\mathbf{b})^T\mathbf{C} = \mathbf{a}^T\mathbf{C}+\mathbf{b}^T\mathbf{C}$	as above, with vectors
$\mathbf{AB} \neq \mathbf{BA}$	multiplication is **not** commutative

Common vector derivatives

In these examples, b is a constant scalar, and $\mathbf{B}$ is a constant matrix.

Scalar derivative $f(x) \rightarrow \frac{df}{dx}$	Vector derivative $f(\mathbf{x}) \rightarrow \frac{df}{d\mathbf{x}}$
$bx \rightarrow b$	$\mathbf{x}^T\mathbf{B} \rightarrow \mathbf{B}$
$bx \rightarrow b$	$\mathbf{x}^T\mathbf{b} \rightarrow \mathbf{b}$
$x^2 \rightarrow 2x$	$\mathbf{x}^T\mathbf{x} \rightarrow 2\mathbf{x}$
$bx^2 \rightarrow 2bx$	$\mathbf{x}^T\mathbf{Bx} \rightarrow 2\mathbf{Bx}$

[2]Adapted from: Kirsty McNaught - Matrix Derivatives Cheat Sheet

Solving the OLS

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} \left(\frac{1}{N} (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) \right) = 0$$

Cheat Sheet Notes

Manipulation:

$$(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T$$

$$(\mathbf{a} + \mathbf{b})^T \mathbf{C} = \mathbf{a}^T \mathbf{C} + \mathbf{b}^T \mathbf{C}$$

Derivatives:

$$\mathbf{x}^T \mathbf{B} \rightarrow \mathbf{B}$$

$$\mathbf{x}^T \mathbf{B} \mathbf{x} \rightarrow 2\mathbf{B}\mathbf{x}$$

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

Solution: Recap

Least Squares Solution

$$\hat{\mathbf{w}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{2}$$

$$\left(\mathbf{X}^T\mathbf{X}\right)\hat{\mathbf{w}} = \mathbf{X}^T\mathbf{y} \tag{3}$$

The expression we obtained is commonly know as the **ordinary least squares** (OLS).
We have found a general expression to obtain the unknown parameters of a linear regressor.

Solution: Recap

Least Squares Solution

$$\hat{\mathbf{w}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{2}$$

$$\left(\mathbf{X}^T\mathbf{X}\right)\hat{\mathbf{w}} = \mathbf{X}^T\mathbf{y} \tag{3}$$

The expression we obtained is commonly know as the **ordinary least squares** (OLS).
We have found a general expression to obtain the unknown parameters of a linear regressor.

In some cases Eq. 3 can be ill-posed.

- If the features are not linearly independent
- If $N \ll D$

leading to errors in the estimation of $\hat{\mathbf{w}}$.

Solution: Recap

Least Squares Solution

$$\hat{\mathbf{w}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{2}$$

$$\left(\mathbf{X}^T\mathbf{X}\right)\hat{\mathbf{w}} = \mathbf{X}^T\mathbf{y} \tag{3}$$

The expression we obtained is commonly know as the **ordinary least squares** (OLS).
We have found a general expression to obtain the unknown parameters of a linear regressor.

In some cases Eq. 3 can be ill-posed.

- If the features are not linearly independent
- If $N \ll D$

leading to errors in the estimation of $\hat{\mathbf{w}}$.

Even in such cases, it is possible to find a solution using of additional techniques (**not covered**)

Predictions

- Once $\hat{\mathbf{w}}$ has been estimated, the fitted model can be used to predict new values of $\hat{y}$:

$$\hat{\mathbf{y}}_{new} = \mathbf{X}_{new}\hat{\mathbf{w}}$$

where $\mathbf{X}_{new}$ is a set of "unseen" input data

Predictions

- Once $\hat{\mathbf{w}}$ has been estimated, the fitted model can be used to predict new values of $\hat{y}$:

$$\hat{\mathbf{y}}_{new} = \mathbf{X}_{new}\hat{\mathbf{w}}$$

  where $\mathbf{X}_{new}$ is a set of "unseen" input data

- The matrix $\mathbf{X}_{new}$ is constructed in the same way as it was done for the training set, but using $\mathbf{x}^*$.

Predictions

- Once $\hat{\mathbf{w}}$ has been estimated, the fitted model can be used to predict new values of $\hat{y}$:

$$\hat{\mathbf{y}}_{new} = \mathbf{X}_{new}\hat{\mathbf{w}}$$

  where $\mathbf{X}_{new}$ is a set of "unseen" input data

- The matrix $\mathbf{X}_{new}$ is constructed in the same way as it was done for the training set, but using $\mathbf{x}^*$.

- **Question:** What would be $\mathbf{X}_{new}$ in the 100m Olympics problem?

Solution to the 100m Olympic Games Problem

Implementing OLS

Implementing OLS solution in Python:

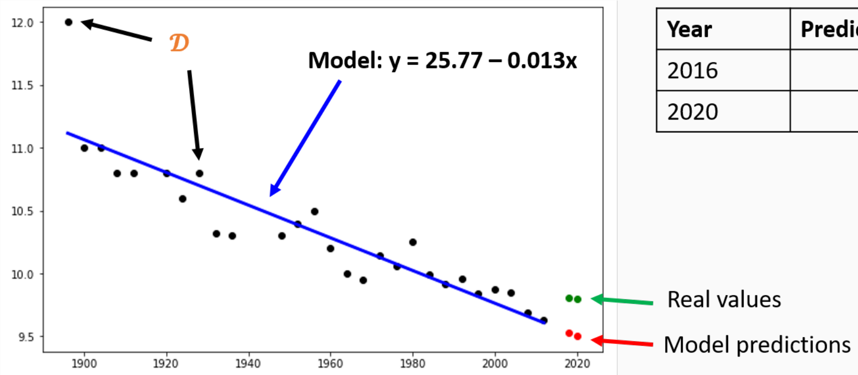
```
def least_squares(X,y):
    X_t = np.transpose(X) #X^T
    X_t_X = X_t.dot(X)    #X^TX
    X_inv = inv(X_t_X)    #(#X^TX)^-1
    X_T_y = X_t.dot(y)    #X^Ty
    w = X_inv.dot(X_T_y)

    return w
```

```
w=least_squares(X,y)
y_hat=np.sum(X*w,axis=1)
```

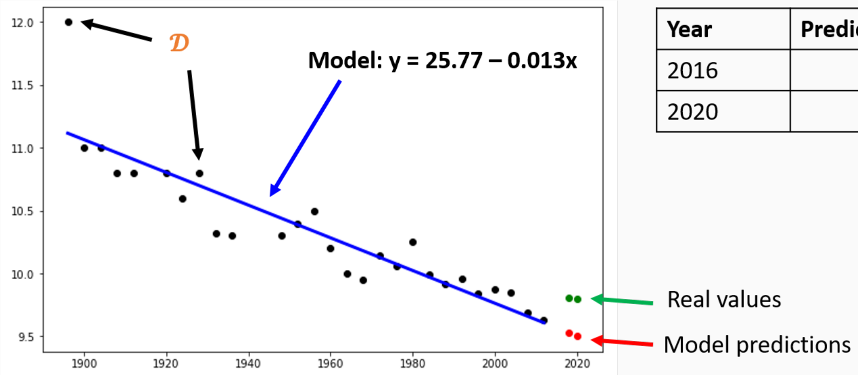
Notebook: See `01_linear_models.ipynb`

The fitted model



Year	Prediction	Real time
2016	9.52	9.81
2020	9.50	9.80

The fitted model



Year	Prediction	Real time
2016	9.52	9.81
2020	9.50	9.80

What can we say about this model?

Models & Assumptions

"All models are wrong but some are useful" - G. Box

- Is the straight line too simple? Should we try to fit a more complex model?
- Is it really always decreasing?
- Our assumptions: It decreases $\not\Leftrightarrow$ it cannot be negative
- Are we being too precise?

Models & Assumptions

“All models are wrong but some are useful” - G. Box

- Is the straight line too simple? Should we try to fit a more complex model?
- Is it really always decreasing?
- Our assumptions: It decreases $\not\Leftrightarrow$ it cannot be negative
- Are we being too precise?

How useful is our model depends on what we are trying to answer

The role of σ^2

The role of σ^2

As the distribution of $\mathbf{y}$ has two parameters, $\hat{\mathbf{w}}$ and σ:

$$p(y_i | \mathbf{x}_i; \mathbf{w}, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{\left(y_i - \mathbf{w}^T \mathbf{x}_i\right)^2}{2\sigma^2} \right\}$$

we can also use the MLE to find an estimation of σ.

The role of σ^2

As the distribution of $\mathbf{y}$ has two parameters, $\hat{\mathbf{w}}$ and σ:

$$p(y_i|\mathbf{x}_i; \mathbf{w}, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{\left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2}{2\sigma^2}\right\}$$

we can also use the MLE to find an estimation of σ.
For simplicity, we will use the matrix notation for this derivation:

$$\prod_{i=1}^{N} p(y_i|\mathbf{x}_i; \mathbf{w}, \sigma^2) = p(\mathbf{y}|\mathbf{X}; \mathbf{w}, \boldsymbol{\Sigma})$$

Link between $\boldsymbol{\Sigma}$ and σ^2

We need to find the link between $\boldsymbol{\Sigma}$ and $\boldsymbol{\sigma}$. For this, lets have a look at the distribution of $\mathbf{y}$ before plugin it into the MLE:

$$p(\mathbf{y}|\mathbf{X}; \mathbf{w}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{N/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{y} - \mathbf{X}\mathbf{w}\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{y} - \mathbf{X}\mathbf{w}\right)\right\}$$

Link between $\boldsymbol{\Sigma}$ and σ^2

We need to find the link between $\boldsymbol{\Sigma}$ and $\boldsymbol{\sigma}$. For this, lets have a look at the distribution of $\mathbf{y}$ before plugin it into the MLE:

$$p(\mathbf{y}|\mathbf{X};\mathbf{w},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{N/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\mathbf{w})^T \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\mathbf{w})\right\}$$

As the noise is independent for every $\mathbf{x}_i$:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}$$

Link between $\boldsymbol{\Sigma}$ and σ^2

We need to find the link between $\boldsymbol{\Sigma}$ and $\boldsymbol{\sigma}$. For this, lets have a look at the distribution of $\mathbf{y}$ before plugin it into the MLE:

$$p(\mathbf{y}|\mathbf{X};\mathbf{w},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{N/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\mathbf{X}\mathbf{w})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\mathbf{w})\right\}$$

As the noise is independent for every $\mathbf{x}_i$:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}$$

Replacing the term for $\boldsymbol{\Sigma}$:

$$p(\mathbf{y}|\mathbf{X};\mathbf{w},\sigma^2\mathbf{I}) = p(\mathbf{y}|\mathbf{X};\mathbf{w},\sigma^2)\frac{1}{(2\pi\sigma)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\mathbf{w})^T(\mathbf{y}-\mathbf{X}\mathbf{w})\right\}$$

Derivation: MLE HOWTO Steps 1 & 2

Plugin this into the MLE and applying the log:

$$\hat{\sigma}^2 = \arg\max_{\sigma^2} \log p(\mathbf{y} | \mathbf{X}; \mathbf{w}, \sigma^2)$$

Derivation: MLE HOWTO Steps 1 & 2

Plugin this into the MLE and applying the log:

$$\begin{aligned}\hat{\sigma}^2 &= \arg\max_{\sigma^2} \log p(\mathbf{y}|\mathbf{X}; \mathbf{w}, \sigma^2) \\ &= \arg\max_{\sigma^2} \log \frac{1}{(2\pi\sigma^2)^{N/2}} + \log\left(\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w})\right\}\right)\end{aligned}$$

Derivation: MLE HOWTO Steps 1 & 2

Plugin this into the MLE and applying the log:

$$
\begin{aligned}
\hat{\sigma}^2 &= \underset{\sigma^2}{\arg\max}\ \log p(\mathbf{y}|\mathbf{X};\mathbf{w},\sigma^2) \\
&= \underset{\sigma^2}{\arg\max}\ \log \frac{1}{(2\pi\sigma^2)^{N/2}} + \log\left(\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\mathbf{w})^T(\mathbf{y}-\mathbf{X}\mathbf{w})\right\}\right) \\
&= \underset{\sigma^2}{\arg\max}\ -\frac{N}{2}\log 2\pi\sigma^2 + \log\left(\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\mathbf{w})^T(\mathbf{y}-\mathbf{X}\mathbf{w})\right\}\right)
\end{aligned}
$$

Derivation: MLE HOWTO Steps 1 & 2

Plugin this into the MLE and applying the log:

$$
\begin{aligned}
\hat{\sigma}^2 &= \arg\max_{\sigma^2} \log p(\mathbf{y}|\mathbf{X}; \mathbf{w}, \sigma^2) \\
&= \arg\max_{\sigma^2} \log \frac{1}{(2\pi\sigma^2)^{N/2}} + \log\left(\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w})\right\}\right) \\
&= \arg\max_{\sigma^2} -\frac{N}{2}\log 2\pi\sigma^2 + \log\left(\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w})\right\}\right) \\
&= \arg\max_{\sigma^2} -\cancel{\frac{N}{2}\log(2\pi)} - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w})
\end{aligned}
$$

Derivation: MLE HOWTO Steps 1 & 2

Plugin this into the MLE and applying the log:

$$
\begin{aligned}
\hat{\sigma}^2 &= \arg\max_{\sigma^2} \log p(\mathbf{y}|\mathbf{X}; \mathbf{w}, \sigma^2) \\
&= \arg\max_{\sigma^2} \log \frac{1}{(2\pi\sigma^2)^{N/2}} + \log \left(\exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) \right\} \right) \\
&= \arg\max_{\sigma^2} -\frac{N}{2} \log 2\pi\sigma^2 + \log \left(\exp \left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) \right\} \right) \\
&= \arg\max_{\sigma^2} -\frac{N}{2} \cancel{\log(2\pi)} - \frac{N}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) \\
&= \arg\min_{\sigma^2} \frac{N}{2} \log \sigma^2 + \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w})
\end{aligned}
$$

Derivation: MLE HOWTO Step 3

According to the MLE HOWTO, to find the minimum, we now derive the obtained expression and equal it to zero:

$$\frac{\partial}{\partial \sigma^2}\left(\frac{N}{2}\log \sigma^2 + \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w})\right) = 0$$

Derivation: MLE HOWTO Step 3

According to the MLE HOWTO, to find the minimum, we now derive the obtained expression and equal it to zero:

$$\frac{\partial}{\partial \sigma^2}\left(\frac{N}{2}\log \sigma^2 + \frac{1}{2\sigma^2}\left(\mathbf{y}-\mathbf{X}\mathbf{w}\right)^T\left(\mathbf{y}-\mathbf{X}\mathbf{w}\right)\right) = 0$$

$$\Rightarrow \quad \hat{\sigma}^2 = \frac{1}{N}\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{w}}\right)^T\left(\mathbf{y}-\mathbf{X}\hat{\mathbf{w}}\right)$$

Estimating $\hat{\sigma}^2$ using MLE

The obtained expression is nothing else than the standard estimate of the variance:

$$\hat{\sigma}^2 = \frac{1}{N} \left(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}} \right)^T \left(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}} \right)$$

Estimating $\hat{\sigma}^2$ using MLE

The obtained expression is nothing else than the standard estimate of the variance:

$$\hat{\sigma}^2 = \frac{1}{N} \left(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}} \right)^T \left(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}} \right)$$

- Note that $\hat{\mathbf{w}}$ intervenes in the estimation of $\hat{\sigma}^2$.

Estimating $\hat{\sigma}^2$ using MLE

The obtained expression is nothing else than the standard estimate of the variance:

$$\hat{\sigma}^2 = \frac{1}{N} (\mathbf{y} - \mathbf{X}\hat{\mathbf{w}})^T (\mathbf{y} - \mathbf{X}\hat{\mathbf{w}})$$

- Note that $\hat{\mathbf{w}}$ intervenes in the estimation of $\hat{\sigma}^2$.
- What information do we gain by having $\hat{\sigma}^2$?

Predictions

New values $\hat{y}_{new}$ are obtained through:

$$\hat{y}_{new} = \mathbb{E}[y_{new} | \mathbf{X}; \mathbf{w}, \sigma^2] = \mathbf{X}_{new} \hat{\mathbf{w}}$$

Predictions

New values $\hat{y}_{new}$ are obtained through:

$$\hat{y}_{new} = \mathbb{E}[y_{new}|\mathbf{X};\mathbf{w},\sigma^2] = \mathbf{X}_{new}\hat{\mathbf{w}}$$

- What information do we gain by having $\hat{\sigma}^2$?

$$\hat{y}_{new} = \mathcal{N}(\hat{\mathbf{w}}^T\mathbf{x}_{new}, \hat{\sigma}^2)$$

Predictions

New values $\hat{y}_{new}$ are obtained through:

$$\hat{y}_{new} = \mathbb{E}[y_{new} | \mathbf{X}; \mathbf{w}, \sigma^2] = \mathbf{X}_{new}\hat{\mathbf{w}}$$

- What information do we gain by having $\hat{\sigma}^2$?

$$\hat{y}_{new} = \mathcal{N}(\hat{\mathbf{w}}^T \mathbf{x}_{new}, \hat{\sigma}^2)$$

- It measures the uncertainty of the prediction

Predictions

New values $\hat{y}_{new}$ are obtained through:

$$\hat{y}_{new} = \mathbb{E}[y_{new}|\mathbf{X}; \mathbf{w}, \sigma^2] = \mathbf{X}_{new}\hat{\mathbf{w}}$$

- What information do we gain by having $\hat{\sigma}^2$?

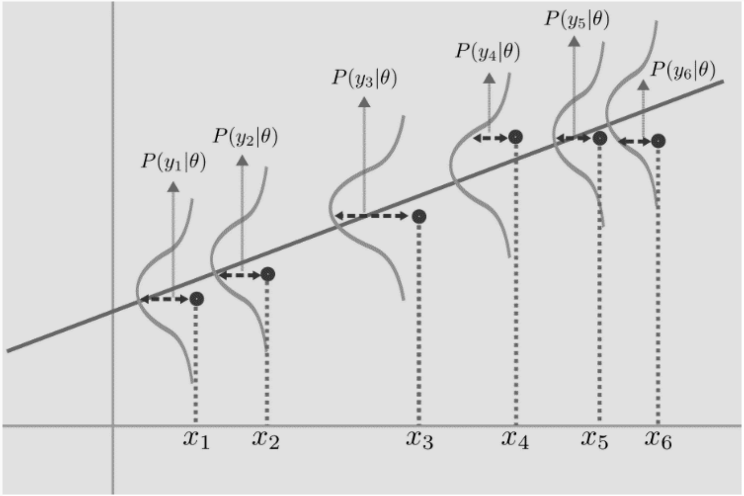
$$\hat{y}_{new} = \mathcal{N}(\hat{\mathbf{w}}^T\mathbf{x}_{new}, \hat{\sigma}^2)$$

- It measures the uncertainty of the prediction

Important

This is not the best measurement for uncertainty (not covered)

Recap: The Role of $\hat{\sigma}^2$



Source: `http://complx.me/2017-01-22-mle-linear-regression/`

Recap

Recap

- We saw linear regression models: our second family of methods
- We introduced the concept of likelihood
- We used Maximum Likelihood Estimation to learn the parameters in linear regression
- We saw that OLS is a solution to the MLE
- MLE allows to have an estimate on the uncertainty of the predictions

Key Concepts

- Linear Regression
- Likelihood
- Ordinary Least Squares (OLS)
- Maximum Likelihood Estimation (MLE)
- Model Parameters

References

Further Reading and Useful Material

Source	Notes
Pattern Recognition and Machine Learning	Ch. 2 and 3
Wikipedia	Multinormal Gaussian distribution (link)
Standford's ML Course	Review Notes on Probability (link)
The Matrix Cook Book	
Introduction to Linear Applied Linear Algebra	Part III Least Squares