

DDA5001 Machine Learning

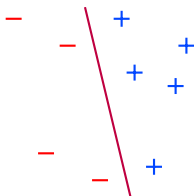
Solution of Least Squares
& Maximum Likelihood Estimation

Xiao Li

School of Data Science
The Chinese University of Hong Kong, Shenzhen



Recap: Linear Classification



Linear classification is to separate the training dataset:

$$\begin{cases} f_{\theta}(\mathbf{x}_i) > 0 & \text{if } y_i = +1, \\ f_{\theta}(\mathbf{x}_i) < 0 & \text{if } y_i = -1, \end{cases} \quad \forall i = 1, \dots, n.$$

The linear classification model is represented as

$$f_{\theta}(\mathbf{x}) = \theta^{\top} \mathbf{x}, \quad y = \text{sign}(f_{\theta}(\mathbf{x})).$$

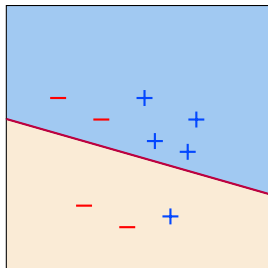
Recap: The Perceptron

- Pick a **misclassified** data \mathbf{x}_i (any misclassified one)

$$\underbrace{\text{sign}(f_{\boldsymbol{\theta}}(\mathbf{x}_i))}_{=\text{sign}(\boldsymbol{\theta}^{\top} \mathbf{x}_i)} \neq y_i.$$

- Update rule

$$\boldsymbol{\theta} = \boldsymbol{\theta} + y_i \mathbf{x}_i.$$



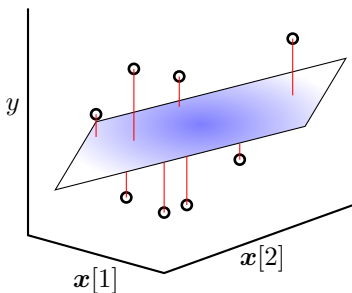
- The perceptron is an iterative algorithm.
- It may converge to **any** linear classifier that can classify the training data points.
- It is designed for **binary classification** and for **linearly separable** data.

Recap: Linear Regression

Linear regression is to use a linear model $f_{\theta}(\mathbf{x})$ to fit the **continuous real-valued** label y . Namely,

$$y_i \approx f_{\theta}(\mathbf{x}_i) = \boldsymbol{\theta}^{\top} \mathbf{x}_i,$$

where $\boldsymbol{\theta} \in \mathbb{R}^d$.



Least squares is a method for finding such a linear regression model:

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2$$

Solution of Least Squares

Maximum Likelihood Estimation

LS Interpretation

Solution of Least Squares

Let (we can ignore the $\frac{1}{n}$ without loss of generality)

$$\mathcal{L}(\theta) = \|X\theta - y\|_2^2$$

Then

$$\begin{aligned}\mathcal{L}(\theta) &= (X\theta - y)^\top (X\theta - y) \\ &= \theta^\top X^\top X\theta - 2\theta^\top X^\top y + y^\top y\end{aligned}$$

Facts:

- ▶ If $h(\theta) = c^\top \theta$, then $\nabla h(\theta) = c$.
- ▶ If $h(\theta) = \theta^\top M\theta$ (M is symmetric), then: $\nabla h(\theta) = 2M\theta$.

Take the gradient

$$\nabla_{\theta} \mathcal{L}(\theta) = 2X^\top (X\theta - y)$$

Setting the gradient to zero and due to **convexity** of \mathcal{L} , the optimal solution $\hat{\theta}$ satisfies

$$X^\top X\hat{\theta} = X^\top y$$

How to solve for $\hat{\theta}$?

Solution of Least Squares: Case I

Case I: $\mathbf{X} \in \mathbb{R}^{n \times d}$ has full column rank

- ▶ $\mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{d \times d}$ will be full rank / invertible (Homework 1).
- ▶ We have

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{y}.$$

Pseudo-inverse of matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ when it has full column rank is defined as

$$\mathbf{X}^\dagger = \left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{X}^\top \in \mathbb{R}^{d \times n}.$$

- ▶ We have $\mathbf{X}^\dagger \mathbf{X} = \mathbf{I}$, but $\mathbf{X} \mathbf{X}^\dagger \neq \mathbf{I}$.

Hence, we can also write

$$\hat{\boldsymbol{\theta}} = \mathbf{X}^\dagger \mathbf{y}.$$

Solution of Least Squares: Case I

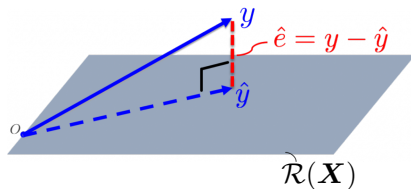
Geometric interpretation:

- Let

$$\mathcal{P}_X = \mathbf{X}\mathbf{X}^\dagger.$$

- \mathcal{P}_X is the **projection matrix** onto the range space $\mathcal{R}(\mathbf{X}) := \{\mathbf{X}\mathbf{a} : \mathbf{a} \in \mathbb{R}^{n \times d}\}$.
- What does the LS solution mean?

$$\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}} = \mathbf{y} - \mathbf{X}\mathbf{X}^\dagger \mathbf{y} = \mathbf{y} - \mathcal{P}_X(\mathbf{y}) := \mathbf{y} - \hat{\mathbf{y}}.$$



Conclusion: The LS solution is such that $\hat{\mathbf{y}} := \mathbf{X}\hat{\boldsymbol{\theta}} = \mathcal{P}_X(\mathbf{y})$ is the **orthogonal projection** of \mathbf{y} onto the range space $\mathcal{R}(\mathbf{X})$.

Solution of Least Squares: Case II

Case II: $\mathbf{X} \in \mathbb{R}^{n \times d}$ does not have full column rank.

- ▶ One typical case is, $n < d$. It means overfitting.
- ▶ We do not have a unique solution. Instead, we have infinitely many solutions.

We can represent all the solutions using singular value decomposition (SVD), and it will be discussed in Homework 1.

How to obtain a unique and meaningful solution in this case?

↪ Regularization, which we will study in details later in the overfitting chapter (simple derivation will be discussed in Homework 1).

Solution of Least Squares

Maximum Likelihood Estimation

LS Interpretation

Maximum Likelihood Estimation: A Simple Example

Flip Coin

- ▶ Setup: For a 'special' coin (with head and back), flip it.
- ▶ Outcome: Dataset $\mathcal{D} = \{H, H, B, H, B, \dots\}$, k heads out of n flips.
- ▶ Question:

What is the probability it will be head?

Assumption:

- ▶ $\Pr[\text{head}] = \theta$ and $\Pr[\text{back}] = 1 - \theta$.
- ▶ Flips are i.i.d.
 - The $i + 1$ -th flip is independent of i -th flip.
 - All flips follow the binomial distribution.

Task: Learning θ from data using Maximum Likelihood Estimation (MLE).

Conditional Probability and Likelihood

- ▶ The conditional probability

$$\Pr[y|x]$$

means **the probability of y given x** . An important quantity for learning (\rightsquigarrow later course on logistic regression and language modeling)

- ▶ How about

$$\Pr[\mathcal{D}|\theta]?$$

Likelihood of data

$\Pr[\mathcal{D}|\theta]$: The probability of observed data given parameter θ .

Write down the likelihood:

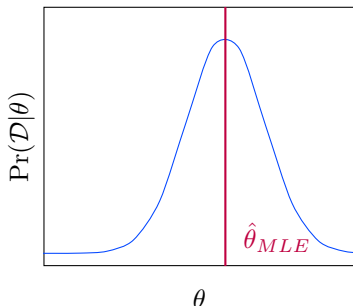
$$\Pr[\mathcal{D}|\theta] = \theta^k (1 - \theta)^{n-k}$$

since we have k heads out of n trials and i.i.d.

Maximum Likelihood Estimation

Maximum likelihood estimation: The principle

Maximize the probability of observed data over parameter θ .



Log-likelihood:

$$\log(\Pr[\mathcal{D}|\theta]) = k \log(\theta) + (n - k) \log(1 - \theta)$$

Maximum Likelihood Estimation

Set $\mathcal{L}(\theta) = \log(\Pr[\mathcal{D}|\theta])$, we have

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \frac{k}{\theta} - \frac{n-k}{1-\theta}$$

Optimization: Letting the derivative of $\mathcal{L}(\theta)$ to be zero gives

$$\hat{\theta}_{MLE} = \frac{k}{n}$$

Why log-likelihood?

► $\log(\cdot)$ is increasing, thus

$$\max_{\theta} \Pr[\mathcal{D}|\theta] \iff \max_{\theta} \log(\Pr[\mathcal{D}|\theta])$$

► $\log(\Pr[\mathcal{D}|\theta])$ is often easier to maximize than $\Pr[\mathcal{D}|\theta]$.

Maximum Likelihood Estimation: Survey

- ▶ Observe (i.i.d.) data $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ following $p(\mathbf{x}|\boldsymbol{\theta}^*)$.
- ▶ Build the likelihood function $p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^n p(\mathbf{x}_i|\boldsymbol{\theta})$ for some parameter $\boldsymbol{\theta}$.
- ▶ Log-likelihood $\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^n \log(p(\mathbf{x}_i|\boldsymbol{\theta}))$
- ▶ Maximum likelihood estimator: $\hat{\boldsymbol{\theta}}_{MLE} = \operatorname{argmax}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$.

Key point: Know $p(\mathbf{x}|\boldsymbol{\theta})$.

Solution of Least Squares

Maximum Likelihood Estimation

LS Interpretation

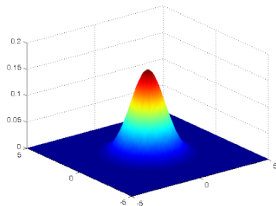
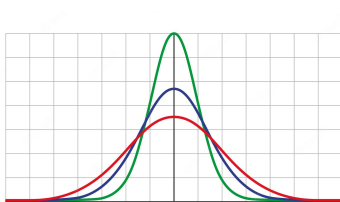
Recall: Multivariate Gaussian distribution

- **Gaussian distribution.** A random variable X is said to follow $\mathcal{N}(\mu, \sigma^2)$ (Gaussian distribution with mean μ and variance σ^2) if its probability density function (PDF) is given by

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

- **Multivariate Gaussian distribution.** We say the random vector $\mathbf{X} \in \mathbb{R}^d$ follows Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ (assumed to be PD), if its PDF is given by

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$



LS for Linear Regression: MLE Interpretation

- ▶ Recap LS:

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2$$

- ▶ Recap our model:

$$y_i \approx f_{\boldsymbol{\theta}}(\mathbf{x}_i) = \boldsymbol{\theta}^\top \mathbf{x}_i.$$

- ▶ To be more explicit, consider

$$y_i = \boldsymbol{\theta}^\top \mathbf{x}_i + \epsilon_i \quad \text{assume} \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2),$$

where ϵ_i are i.i.d. Gaussian for $i = 1, \dots, n$.

- ▶ In matrix notation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \quad \text{with} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = \text{diag}(\sigma^2, \dots, \sigma^2)$.

LS for Linear Regression: MLE Interpretation

Equivalently,

$$\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}).$$

The likelihood: The probability of observed data given parameters $\boldsymbol{\theta}$

$$\begin{aligned} p(\mathcal{D}|\boldsymbol{\theta}) &= \frac{1}{(2\pi)^{n/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{\epsilon}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\epsilon}\right) \\ &= \frac{1}{(2\pi)^{n/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})\right). \end{aligned}$$

Log-likelihood:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \log(p(\mathcal{D}|\boldsymbol{\theta})) = \text{constant} - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \\ &= \text{constant} - \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2. \end{aligned}$$

LS for Linear Regression: MLE Interpretation

MLE principle:

$$\hat{\boldsymbol{\theta}}_{MLE} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2,$$

which is exactly LS.

Take-home message

Least square is a maximum likelihood estimator under Gaussian noise assumption

How Good Is This Estimator? Unbiased or Biased.

Suppose the target hypothesis g is

$$\mathbf{y} = g(\mathbf{X}) = \mathbf{X}\boldsymbol{\theta}^* + \boldsymbol{\epsilon} \quad \text{with} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

Goal: To learn the target $\boldsymbol{\theta}^*$ (then we know g to the most extent).

Assume full column rank of \mathbf{X} (non-overfitting case), then MLE / LS solution satisfies

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{LS} &= \hat{\boldsymbol{\theta}}_{MLE} = \mathbf{X}^\dagger \mathbf{y} \\ &= \mathbf{X}^\dagger (\mathbf{X}\boldsymbol{\theta}^* + \boldsymbol{\epsilon}) \\ &= \underbrace{\mathbf{X}^\dagger \mathbf{X}}_{\mathbf{I}} \boldsymbol{\theta}^* + \mathbf{X}^\dagger \boldsymbol{\epsilon} \\ &= \boldsymbol{\theta}^* + \mathbf{X}^\dagger \boldsymbol{\epsilon}\end{aligned}$$

Definition: An estimator of a given parameter is said to be **unbiased** if its expected value is equal to the true/underlying value of the parameter.

Unbiased Estimator:

$$\mathbb{E}[\hat{\boldsymbol{\theta}}_{MLE}] = \boldsymbol{\theta}^*$$

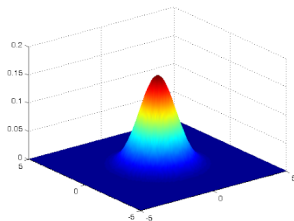
Appendix

MLE for Estimating Gaussian Parameters

MLE for Gaussian

Recall Multi-variate Gaussian distribution:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$



Data: Draw a set of i.i.d. samples $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ follow $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$.

Qes: What is the $\hat{\boldsymbol{\theta}}_{MLE}$ for $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ given data?

MLE for Gaussian

Likelihood: Set parameters $\theta = (\mu, \Sigma)$

$$\begin{aligned} p(\mathcal{D}|\theta) &= \prod_{i=1}^n p(\mathbf{x}_i|\theta) \\ &= \frac{1}{(2\pi)^{nd/2}} \frac{1}{|\Sigma|^{n/2}} \prod_{i=1}^n \exp\left(-\frac{1}{2}(\mathbf{x}_i - \mu)^\top \Sigma^{-1}(\mathbf{x}_i - \mu)\right) \end{aligned}$$

Log-likelihood: Set parameters $\theta = (\mu, \Sigma)$

$$\begin{aligned} \mathcal{L}(\theta) &= \sum_{i=1}^n \log(p(\mathbf{x}_i|\theta)) \\ &= -\frac{nd}{2} \log(2\pi) - \frac{n}{2} \log(|\Sigma|) - \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \mu)^\top \Sigma^{-1}(\mathbf{x}_i - \mu) \end{aligned}$$

MLE for Gaussian

Optimization:

- We first optimization over μ , compute the **gradient**

$$\nabla_{\mu} \mathcal{L}(\theta) = \Sigma^{-1} \sum_{i=1}^n (x_i - \mu).$$

Setting the gradient to be 0, together with the fact that Σ^{-1} is positive definite, provides

$$\hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^n x_i.$$

MLE for Gaussian

Optimization:

- We then optimize over Σ . Note that $|\Sigma|^{-1} = |\Sigma^{-1}|$. Taking gradient over Σ^{-1} yields

$$\nabla_{\Sigma^{-1}} \mathcal{L}(\theta) = \frac{n}{2} \Sigma - \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{MLE})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{MLE})^\top$$

setting the gradient to zero gives

$$\hat{\Sigma}_{MLE} = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{MLE})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_{MLE})^\top$$

Used facts:

$$\nabla_{\mathbf{A}} \log(|\mathbf{A}|) = \mathbf{A}^{-\top}$$

and

$$\nabla_{\mathbf{A}} \text{tr}(\mathbf{A}^\top \mathbf{B}) = \mathbf{B}.$$

Unbiased and Biased Estimators

MLE for the mean of Gaussian is **unbiased**:

$$\mathbb{E} [\hat{\boldsymbol{\mu}}_{MLE}] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\mathbf{x}_i] = \boldsymbol{\mu}^{\star}$$

MLE for the variance of Gaussian is **biased**:

$$\mathbb{E} \left[\hat{\boldsymbol{\Sigma}}_{MLE} \right] \neq \boldsymbol{\Sigma}^{\star}$$