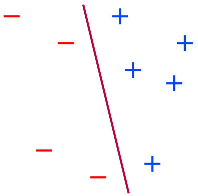
DDA5001 Machine Learning

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:

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

The linear classification model is represented as

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

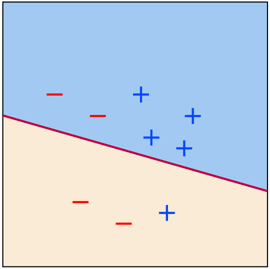
Recap: The Perceptron

- Pick a misclassified data $\boldsymbol{x}_i$ (any misclassified one)

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

- Update rule

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



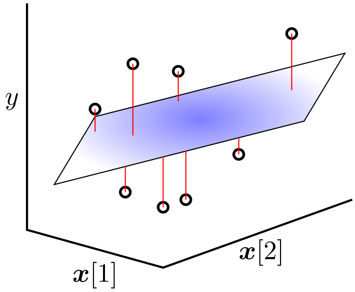
- The perceptron is an iterative algorithm.
- It may converge to any linear classifier that can classify the training data points.
- It is designed for binary classification and for linearly separable data.

Recap: Linear Regression

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

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

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



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

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

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)

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

Then

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

Facts:

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

Take the gradient

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

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

$$\boxed{\boldsymbol{X}^\top \boldsymbol{X}\widehat{\boldsymbol{\theta}} = \boldsymbol{X}^\top \boldsymbol{y}}$$

How to solve for $\widehat{\boldsymbol{\theta}}$?

Solution of Least Squares: Case I

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

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

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

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

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

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

Hence, we can also write

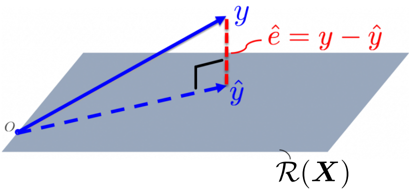
$$\widehat{\boldsymbol{\theta}} = \boldsymbol{X}^\dagger \boldsymbol{y}.$$

Solution of Least Squares: Case I

Geometric interpretation:

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

$$\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\theta}} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{X}^{\dagger}\boldsymbol{y} = \boldsymbol{y} - \mathcal{P}_{\boldsymbol{X}}(\boldsymbol{y}) := \boldsymbol{y} - \widehat{\boldsymbol{y}}.$$



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

Solution of Least Squares: Case II

Case II: $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ 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?
$\rightsquigarrow$ Regularization, which we will study in details later in the overfitting chapter (simple derivation will be discussed in Homework 1).

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,...}, 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:

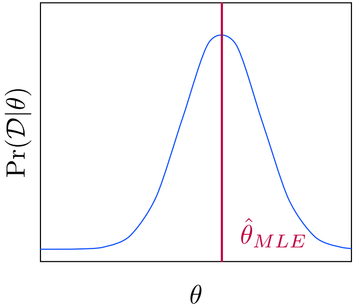
$$\Pr[\mathcal{D}|\theta] = \theta^k (1-\theta)^{n-k}$$

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

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

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

$$\boxed{\widehat{\theta}_{MLE} = \frac{k}{n}}$$

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 $\mathcal{D} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ following $p(\boldsymbol{x}|\boldsymbol{\theta}^\star)$.
- Build the likelihood function $p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^n p(\boldsymbol{x}_i|\boldsymbol{\theta})$ for some parameter $\boldsymbol{\theta}$.
- Log-likelihood $\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^n \log(p(\boldsymbol{x}_i|\boldsymbol{\theta}))$
- Maximum likelihood estimator: $\widehat{\boldsymbol{\theta}}_{MLE} = \mathrm{argmax}_{\boldsymbol{\theta}}\, \mathcal{L}(\boldsymbol{\theta})$.

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

Solution of Least Squares

Maximum Likelihood Estimation

LS Interpretation

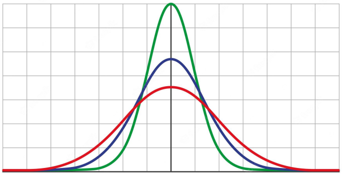
Recall: Multivariate Gaussian distribution

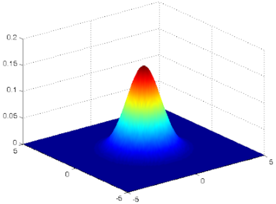
- Gaussian distribution. A random variable X is said to follow $\mathcal{N}(\mu, \sigma^2)$ (Gaussian distribution with mean μ and variance σ^2) if its probability density function (PDF) is given by

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

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

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





LS for Linear Regression: MLE Interpretation

- Recap LS:

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

- Recap our model:

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

- To be more explicit, consider

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

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

- In matrix notation

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

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

LS for Linear Regression: MLE Interpretation

Equivalently,

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

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

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

Log-likelihood:

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

LS for Linear Regression: MLE Interpretation

MLE principle:

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

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

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

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

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

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

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:

$$\mathbb{E}[\widehat{\boldsymbol{\theta}}_{MLE}] = \boldsymbol{\theta}^\star$$

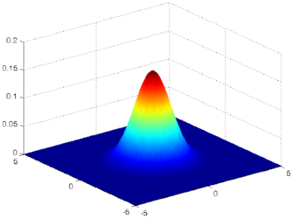
Appendix

MLE for Estimating Gaussian Parameters

MLE for Gaussian

Recall Multi-variate Gaussian distribution:

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



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

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

MLE for Gaussian

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

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

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

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

MLE for Gaussian

Optimization:

- We first optimization over $\boldsymbol{\mu}$, compute the **gradient**

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

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

$$\widehat{\boldsymbol{\mu}}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i.$$

MLE for Gaussian

Optimization:

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

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

setting the gradient to zero gives

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

Used facts:

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

and

$$\nabla_{\boldsymbol{A}} \operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{B}) = \boldsymbol{B}.$$

Unbiased and Biased Estimators

MLE for the mean of Gaussian is unbiased:

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

MLE for the variance of Gaussian is biased:

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