Mathematics for Machine Learning

— When Models Meet Data

Parameter Estimation

Joseph Chuang-Chieh Lin

Department of Computer Science & Engineering,
National Taiwan Ocean University

Fall 2025

Credits for the resource

- The slides are based on the textbooks:
  - *Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong: Mathematics for Machine Learning. Cambridge University Press. 2020.*
  - *Arnold J. Insel, Lawrence E. Spence, Stephen H. Friedberg: Linear Algebra, 4th Edition. Prentice Hall. 2013.*
  - *Howard Anton, Chris Rorres, Anton Kaul: Elementary Linear Algebra, 12th Edition. Wiley. 2019.*
- We could partially refer to the monograph:
  Francesco Orabona: A Modern Introduction to Online Learning. https://arxiv.org/abs/1912.13213

Outline

1. Maximum Likelihood Estimation

2. Maximum A Posteriori Estimation

Goal

- Use probabilistic distributions to model our uncertainty due to:
  - the observation process.
  - the uncertainty in the parameters of the predictor.

Outline

Maximum Likelihood Estimation (MLE)

For data represented by a random variable $\mathbf{x}$ and for a family of probability densities $p(\mathbf{x} \mid \boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$, we aim at the negative log-likelihood:

$$\mathcal{L}_{\mathbf{x}}(\boldsymbol{\theta}) = -\log p(\mathbf{x} \mid \boldsymbol{\theta}).$$

- **Note:** The parameter $\boldsymbol{\theta}$ is varying and the data $\mathbf{x}$ is fixed.
- $\mathcal{L}_{\mathbf{x}}(\boldsymbol{\theta})$: a function of $\boldsymbol{\theta}$.

Maximum Likelihood Estimation (MLE)

For data represented by a random variable $\mathbf{x}$ and for a family of probability densities $p(\mathbf{x} \mid \boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$, we aim at the negative log-likelihood:

$$\mathcal{L}_{\mathbf{x}}(\boldsymbol{\theta}) = -\log p(\mathbf{x} \mid \boldsymbol{\theta}).$$

- **Note:** The parameter $\boldsymbol{\theta}$ is varying and the data $\mathbf{x}$ is fixed.
- $\mathcal{L}_{\mathbf{x}}(\boldsymbol{\theta})$: a function of $\boldsymbol{\theta}$.

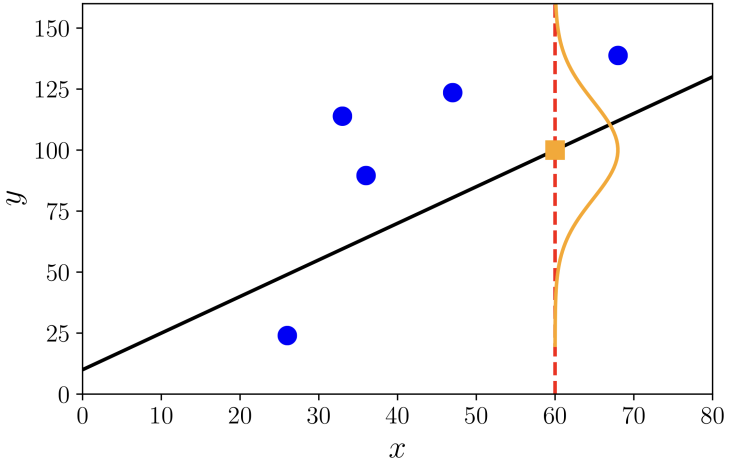
For a given dataset $\mathbf{x}$, the likelihood allows us choose the settings of $\boldsymbol{\theta}$ that more "likely" has generated the data or how "likely" $\boldsymbol{\theta}$ is for the observations $\mathbf{x}$.

Example

- Specify that the conditional probability of the labels given the examples is a Gaussian distribution.
- Assume that we can explain our observation uncertainty by independent Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.
- We assume the linear model $\mathbf{x}_i^\top \boldsymbol{\theta}$ is used for prediction.

For each example-label pair $(\mathbf{x}_i, y_i)$,

$$p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}) = \mathcal{N}(y_n \mid \mathbf{x}_i^\top \boldsymbol{\theta}, \sigma^2).$$


150
125
100
75
50
25
0
y
0
10
20
30
40
50
60
70
80
x

MLE for i.i.d. examples

- Assume that $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)$ are i.i.d.
- The likelihood factorizes into a product of likelihoods of each individual example

$$p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) = \prod_{i=1}^{N} p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}).$$

MLE for i.i.d. examples

- Assume that $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)$ are i.i.d.
- The likelihood factorizes into a product of likelihoods of each individual example

$$p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) = \prod_{i=1}^{N} p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}).$$

Then,

$$\mathcal{L}(\boldsymbol{\theta}) = -\log p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) = -\sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}).$$

MLE for i.i.d. examples

- Assume that $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)$ are i.i.d.
- The likelihood factorizes into a product of likelihoods of each individual example

$$p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) = \prod_{i=1}^{N} p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}).$$

Then,

$$\mathcal{L}(\boldsymbol{\theta}) = -\log p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) = -\sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}).$$

Note: Do not forget that $\mathcal{L}(\boldsymbol{\theta})$ is a function of $\boldsymbol{\theta}$.

Example (contd.)

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= -\sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}) = -\sum_{i=1}^{N} \log \mathcal{N}(y_i \mid \mathbf{x}_i^\top \boldsymbol{\theta}, \sigma^2) \\
&= -\sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\theta})^2}{2\sigma^2}\right) \\
&= -\sum_{i=1}^{N} \log \exp\left(-\frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\theta})^2}{2\sigma^2}\right) - \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} \\
&= \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mathbf{x}_i^\top \boldsymbol{\theta})^2 - \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}}.
\end{aligned}
$$

$\Longrightarrow$ minimizing $\mathcal{L}(\boldsymbol{\theta})$

Example (contd.)

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= -\sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}) = -\sum_{i=1}^{N} \log \mathcal{N}(y_i \mid \mathbf{x}_i^\top \boldsymbol{\theta}, \sigma^2) \\
&= -\sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\theta})^2}{2\sigma^2}\right) \\
&= -\sum_{i=1}^{N} \log \exp\left(-\frac{(y_i - \mathbf{x}_i^\top \boldsymbol{\theta})^2}{2\sigma^2}\right) - \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} \\
&= \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mathbf{x}_i^\top \boldsymbol{\theta})^2 - \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}}.
\end{aligned}$$

The second term is constant.

$\Longrightarrow$ minimizing $\mathcal{L}(\boldsymbol{\theta}) \Rightarrow$ solving the least-squares problem.

Issues

- This example has a closed-form solution.

Issues

- This example has a closed-form solution.
    - No closed-form $\Rightarrow$ resort to numerical optimization.
- MLE may suffer from overfitting

Issues

- This example has a closed-form solution.
  - No closed-form $\Rightarrow$ resort to numerical optimization.
- MLE may suffer from overfitting (analogous to unregularized empirical risk minimization).

Outline

Motivation (1/2)

What if we have prior knowledge about the distribution of the parameters θ?

Motivation (1/2)

What if we have prior knowledge about the distribution of the parameters $\boldsymbol{\theta}$?

We can multiply an additional term (i.e., $p(\boldsymbol{\theta})$) to the likelihood.

Motivation (2/2)

- For a given prior, after observing some data $\mathbf{x}$, how should we update $p(\boldsymbol{\theta})$?
  - $\Rightarrow$ Bayes's theorem.
  - $\star$ Compute a posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{x})$.

$$p(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})}. \quad \text{(the more specific knowledge on } \boldsymbol{\theta}\text{)}$$

Motivation (2/2)

- For a given prior, after observing some data $\mathbf{x}$, how should we update $p(\boldsymbol{\theta})$?
  - $\Rightarrow$ Bayes's theorem.
  - $\star$ Compute a posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{x})$.

$$p(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})}. \quad \text{(the more specific knowledge on } \boldsymbol{\theta}\text{)}$$

So,

$$p(\boldsymbol{\theta} \mid \mathbf{x}) \propto p(\mathbf{x} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$$

since $p(\mathbf{x})$ does not depend on $\boldsymbol{\theta}$.

Motivation (2/2)

- For a given prior, after observing some data $\mathbf{x}$, how should we update $p(\boldsymbol{\theta})$?
  - $\Rightarrow$ Bayes's theorem.
  - $\star$ Compute a posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{x})$.

$$p(\boldsymbol{\theta} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})}. \quad \text{(the more specific knowledge on } \boldsymbol{\theta}\text{)}$$

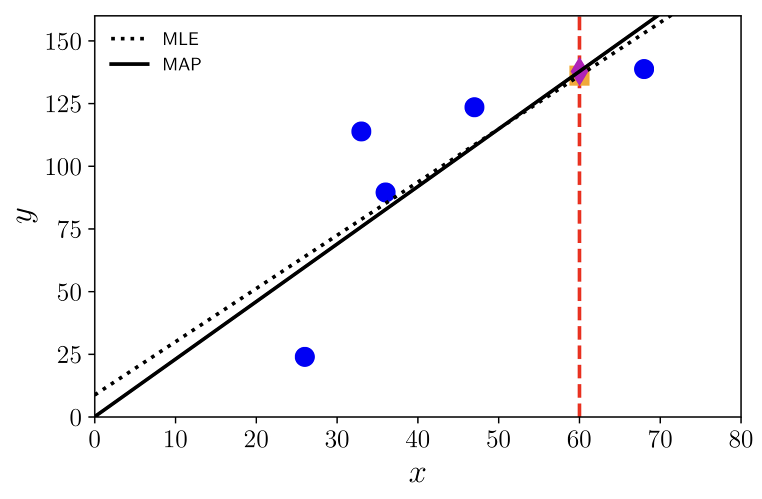
  So,

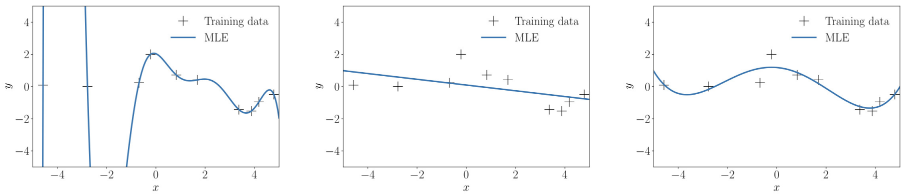
$$p(\boldsymbol{\theta} \mid \mathbf{x}) \propto p(\mathbf{x} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$$

  since $p(\mathbf{x})$ does not depend on $\boldsymbol{\theta}$.

- We now estimate the minimum of the negative log-posterior, which is referred to as maximum a posteriori estimation (MAP estimation).

MLE vs. MAP

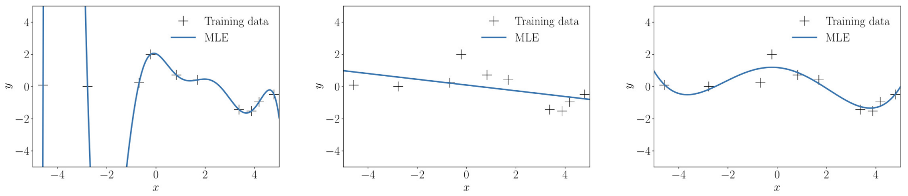




(a) Overfitting

(b) Underfitting.

(c) Fitting well.



(a) Overfitting (b) Underfitting. (c) Fitting well.

- We will see the details of MAP in linear regression later.

Discussions