AIA Exercise

Bayesian Estimation

Simon Matern

Computer Vision and Remote Sensing
Technische Universität Berlin

May 19, 2025

Overview

1. Maximum Likelihood Estimation
2. Maximum A Posteriori Estimation
3. Bayesian Estimation
4. Bayesian Decision Theory

Maximum Likelihood Estimation (MLE)

Assuming our data comes from a parametrized distribution, how can one estimate its parameters given the observations?

MLE Definition

Given a set of observations $D = \{x_1, \cdots, x_n\}$ with i.i.d $x_i \sim p(x|\theta)$ The MLE is defined as

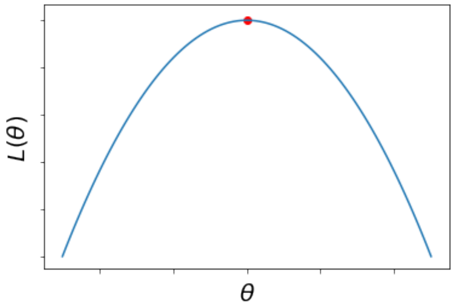
$$\hat{\theta}_{ML} := \arg\max_{\theta} \underbrace{p(D|\theta)}_{\text{likelihood}} = \arg\max_{\theta} \prod_{i=1}^{n} p(x_i|\theta)$$

$$= \arg\max_{\theta} \underbrace{\log p(D|\theta)}_{\text{log-likelihood}} = \arg\max_{\theta} \sum_{i=1}^{n} \log p(x_i|\theta)$$

MLE

MLE procedure

1. formulate likelihood analytically: $p(D|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$
2. formulate log-likelihood analytically: $L(\theta) := \sum_{i=1}^{n} \log p(x_i|\theta)$
3. compute gradient: $\nabla L(\theta)$
4. find extrema: $\nabla L(\hat{\theta}_{ML}) \stackrel{!}{=} 0$



MLE Example

Example

We observe a coin-toss experiment $D = \{x_1, \cdots, x_n\}$ with i.i.d. $x_i \sim p(x_i|\theta)$

$$p(x_i|\theta) = \begin{cases} \theta & \text{if } x_i = 1 \ (head), \\ 1 - \theta & \text{if } x_i = 0 \ (tail) \end{cases}$$

Example

1. Likelihood: $p(D|\theta) = \theta^k \cdot (1-\theta)^{n-k}$ where k is the number of heads
2. Log-likelihood: $L(\theta) = k \log(\theta) + (n-k)\log(1-\theta)$
3. Gradient: $\nabla L(\theta) = \dfrac{k}{\theta} - \dfrac{n-k}{1-\theta}$
4. Extremum: $\nabla L(\theta) \stackrel{!}{=} 0$

MLE Example

Example

4. Extremum:

$$
\begin{aligned}
\nabla L(\theta_{ML}) &\stackrel{!}{=} 0 \\
\Leftrightarrow \frac{k}{\theta} - \frac{n-k}{1-\theta} &= 0 \\
\Leftrightarrow \frac{k(1-\theta) - \theta(n-k)}{\theta(1-\theta)} &= 0 \\
\Leftrightarrow k(1-\theta) - \theta(n-k) &= 0 \\
\Leftrightarrow k - \theta n &= 0 \\
\Rightarrow \hat{\theta}_{ML} = \frac{k}{n}
\end{aligned}
$$

MLE Visualization

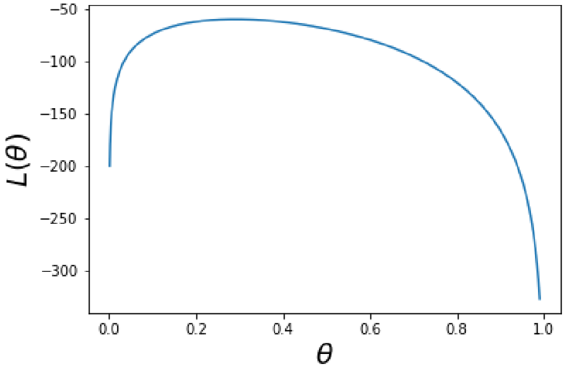


Figure: The graph shows the log-likelihood of a Bernoulli distribution with $\theta = 0.3$

MLE Visualization II

Summary: As shown before, the

Log-likelihood of Bernoulli distribution

is defined as

$$L(\theta \mid \mathbf{x}) = k \log \theta + (n-k)\log(1-\theta), \qquad 0 < \theta < 1.$$

We fixed the data and are scanning over all $\theta \in [0,1]$
to see which parameter value makes the data most plausible.

MLE: Exercise

Task 1

- Why apply a logarithm on the likelihood?
- What are analytical reasons?
- What are numerical reasons?
- Does it affect the estimator?

Task 2:

We observe an experiment $D = \{x_1, \cdots, x_n\}$ with i.i.d. $x_i \sim p(x_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

What is the MLE for μ and σ^2 ?

Task 3: Regression

We observe am experiment $D = \{(x_1, y_1), \cdots, (x_n, y_n)\}$. We assume a linear model with Gaussian noise: $y_i = x_i \cdot a + b + \epsilon_i$ with i.i.d. $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. What is the MLE for a, b and σ^2 ?

Maximum a posteriori estimation (MAP)

Problem

- MLE is purely data-driven. This leads to some unstable behavior for estimations with low amount of data.
- How can one incorporate additional knowledge into the estimation?

Solution

- Treat parameter θ as a random variable.
- Find mostly likely θ given the data

$$
\begin{aligned}
\hat{\theta}_{MAP} &= \arg\max_{\theta} p(\theta|D) \\
&= \arg\max_{\theta} \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta} \\
&= \arg\max_{\theta} p(D|\theta)p(\theta)
\end{aligned}
$$

MAP

MAP procedure

- A prior distribution $p(\theta)$ can model a certainty over the parameter space
- $\hat{\theta}_{MAP}$ can be found the same way as MLE. The only difference is that the likelihood has an additional constraint.

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \underbrace{p(D|\theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}}$$

Bayesian Estimation

Problem

MLE and MAP are **point estimators**. They provide no certainty over the found solution. What is the distribution for a new measurement x given our data D?

Bayesian Estimation

$$p(x|D) = \int \underbrace{p(x|\theta)}_{\text{pdf}} \underbrace{p(\theta|D)}_{\text{Posterior probability}} d\theta$$

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta}$$

Bayesian Estimation: Exercise

Task

Let $D = (x_1, x_2, ..., x_7) = (0, 0, 1, 1, 0, 0, 1)$. Assume $p(x_i|\theta) = \begin{cases} \theta & \text{if } x_i = 1 \ (head), \\ 1-\theta & \text{if } x_i = 0 \ (tail) \end{cases}$

- Let $p(\theta) = \mathcal{N}(0.5, 0.1)$. What is the MAP estimator θ_{MAP}? What is the probability of tossing tails two times $P(x_8 = 0, x_9 = 0|\theta_{MAP})$
- Let $p(\theta) = \mathcal{U}(0, 1)$. What is the probability of the next toss to be head $P(x_8 = 1|D)$

Bayesian Decision Theory

Discriminant Functions

- Select class i with highest probability given measurement x:

$$\arg\max_i P(\omega_i|x) = \frac{p(x|\omega_i)P(\omega_i)}{p(x)}$$

- Alternatively use any functions $g_i(x)$ with

$$k = \arg\max_i \; g_i(x) \Leftrightarrow k = \arg\max_i P(\omega_i|x)$$

Examples

- $g_i(x) = P(\omega_i|x)$
- $g_i(x) = p(x|\omega_i)P(\omega_i)$
- $g_i(x) = \log p(x|\omega_i) + \log P(\omega_i)$
- $g_i(x) = f(\hat{g}_i(x))$ for any monotonic function f and some discriminant $\hat{g}_i(x)$

Bayesian Decision Theory: Error

Using our discriminant functions for decision making what is the expected error ?

Error Metric

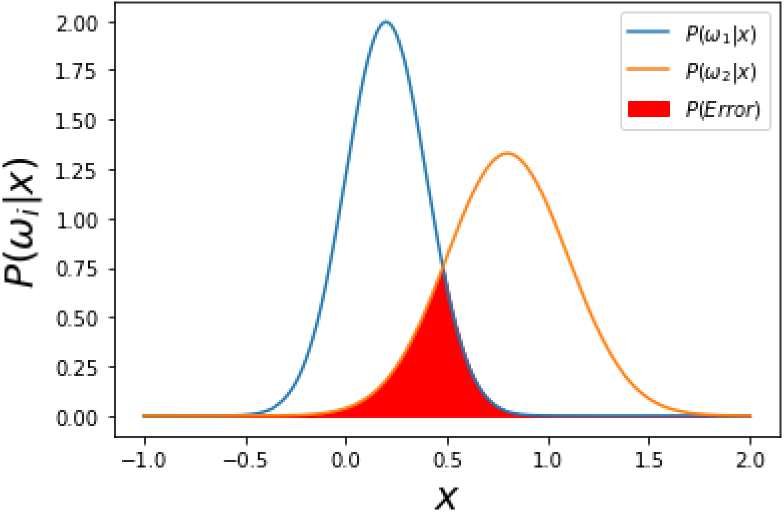
- Conditional error:

$$
\begin{aligned}
P(\text{error}|x) &= 1 - \max_i (P(\omega_i|x)) \\
&= \min(P(\omega_1|x), P(\omega_2|x)) \quad \text{for binary classification}
\end{aligned}
$$

- Expected error:

$$
P(error) = \int P(\text{error}|x)p(x)dx
$$

Bayesian Decision Theory: Error



Bayesian Decision Theory: Exercise

Task 1

- If $p(x|\omega_i)$ is assumed to be Gaussian $p(x|\omega_i) = \mathcal{N}(\mu_i, \Sigma_i)$
  - compute the discriminant function: $g_i(x) = \log[p(x|\omega_i)P(\omega_i)]$
  - When is the decision boundary linear? $w^T(x - x_0) = 0 \quad \forall x$ with $g_i(x) = g_j(x)$
  - In which case is the optimal decision rule to always choose class ω_1? Explain the parameters of this scenario.
- How does the distribution of the features $p(x)$ affect the classification error?
- Are the following statements correct of wrong?
  - If $P(\omega_1) > P(\omega_2)$ it is always better to select class ω_1
  - If $\forall i, j : P(\omega_i) = P(\omega_j)$ then $g_i(x) = p(x|\omega_i)$ are valid discriminator functions?
- In which case are $g_i(x) = P(\omega_i)$ valid discriminator functions ?