Maximum-Likelihood Estimation

Jessica Fridrich

EECE 566

Introduction

Problem

- Observation $\mathbf{x} \in \mathbb{R}^n \sim p(\mathbf{x}; \boldsymbol{\theta}), \boldsymbol{\theta} \in \mathbb{R}^k$, the parameter vector. Estimate $\boldsymbol{\theta}$ from observations
  - e. g. $\mathcal{N}(\mathbf{x}; \mu, \sigma^2)$, $\boldsymbol{\theta} = (\mu, \sigma^2)$

MLE (Maximum-Likelihood Estimation)

- Estimate as $\widehat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}} p(\mathbf{x}; \boldsymbol{\theta})$
- $p(\mathbf{x}; \boldsymbol{\theta})$ is the likelihood of observing $\mathbf{x}$ given the parameter vector $\boldsymbol{\theta}$

Link to MAP

- Estimate as $\widehat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}} p(\boldsymbol{\theta}; \mathbf{x}) = \text{argmax}_{\boldsymbol{\theta}} p(\mathbf{x}; \boldsymbol{\theta}) p(\boldsymbol{\theta}) / p(\mathbf{x}) = \text{argmax}_{\boldsymbol{\theta}} p(\mathbf{x}; \boldsymbol{\theta}) p(\boldsymbol{\theta})$
- Weighted max. likelihood (need a meaningful prior $p(\boldsymbol{\theta})$)

Maximum-Likelihood Estimation

MLE (Maximum-Likelihood Estimation)

- Estimate as $\widehat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}} p(\mathbf{x}; \boldsymbol{\theta}) = \text{argmax}_{\boldsymbol{\theta}} \ln p(\mathbf{x}; \boldsymbol{\theta})$
- Necessary conditions for maximum:

$$\frac{\partial \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_i} = 0, \; i = 1, \ldots, k$$

Example: Estimating Gaussian parameters

Gaussian distribution parameters

$$
\begin{aligned}
x_i &\sim \mathcal{N}(\mu, \sigma^2), i = 1, \ldots, n, \text{ iid observations} \\
p(\mathbf{x}; \mu, \sigma^2) &= \prod_{i=1}^{n} p(x_i; \mu, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}\right) \\
\ln p(\mathbf{x}; \mu, \sigma^2) &= \frac{-n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2
\end{aligned}
$$

Example: Estimating Gaussian parameters

MLE estimator of μ

$$
\begin{aligned}
\ln p(\mathbf{x};\mu,\sigma^2) &= \frac{-n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \\
\frac{\partial \ln p(\mathbf{x};\mu,\sigma^2)}{\partial\mu} &= \frac{-1}{2\sigma^2}\cdot(-2)\cdot\sum_{i=1}^{n}(x_i-\mu) \\
&= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}x_i - n\mu\right) = 0 \\
\Rightarrow \widehat{\mu} &= \frac{1}{n}\sum_{i=1}^{n}x_i = \overline{\mathbf{x}} \text{ sample mean}
\end{aligned}
$$

Example: Estimating Gaussian parameters

MLE estimator of σ^2

$$\begin{aligned}
\ln p(\mathbf{x};\mu,\sigma^2) &= \frac{-n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \\
\frac{\partial \ln p(\mathbf{x};\mu,\sigma^2)}{\partial\sigma^2} &= \frac{-n}{2}\cdot\frac{2\pi}{2\pi\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 \\
&= \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \\
\Rightarrow \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\widehat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{\mathbf{x}})^2 \text{ sample variance}
\end{aligned}$$

Example: Estimating linear model

Linear model with AWGN

$$y_i = \theta_1 x_{i1} + \ldots + \theta_k x_{ik} + \xi_i$$

where, $\xi_i \sim \mathcal{N}(0, \sigma^2)$ iid. We observe $\mathbf{x}_i, y_i$. Note that $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})$ is a row vector.

$$\begin{aligned}
y_i &= \mathbf{x}_i \cdot \boldsymbol{\theta} + \xi_i \\
y_i | \mathbf{x}_i, \boldsymbol{\theta} &\sim \mathcal{N}(\mathbf{x}_i \cdot \boldsymbol{\theta}, \sigma^2), \text{ and independent} \\
p(y_i; \mathbf{x}_i, \boldsymbol{\theta}, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i \cdot \boldsymbol{\theta})^2}{2\sigma^2}\right) \\
p(\mathbf{y}; \overbrace{\mathbf{x}_1, \ldots, \mathbf{x}_n}^{X}, \boldsymbol{\theta}, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i \cdot \boldsymbol{\theta})^2}{2\sigma^2}\right) \\
\ln p(\mathbf{y}; X, \boldsymbol{\theta}, \sigma^2) &= \frac{-n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{x}_i \cdot \boldsymbol{\theta})^2
\end{aligned}$$

Example: Estimating linear model

$$
\begin{aligned}
y_i &= \theta_1 x_{i1} + \ldots + \theta_k x_{ik} + \xi_i \text{ or } \mathbf{y} = X\boldsymbol{\theta} + \boldsymbol{\xi} \\
\ln p(\mathbf{y}; X, \boldsymbol{\theta}, \sigma^2) &= \frac{-n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{x}_i \cdot \boldsymbol{\theta})^2
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \ln p(\mathbf{y}; X, \boldsymbol{\theta}, \sigma^2)}{\partial \theta_j} &= \frac{2}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{x}_i \cdot \boldsymbol{\theta}) x_{ij} = 0 \\
\Rightarrow \sum_{i=1}^{n} (y_i - \mathbf{x}_i \cdot \boldsymbol{\theta}) x_{ij} &= 0, j = 1, \ldots, k \\
\underbrace{\sum_{i=1}^{n} y_i x_{ij}}_{\mathbf{y}^T X} - \underbrace{\sum_{i=1}^{n} (\mathbf{x}_i \cdot \boldsymbol{\theta}) x_{ij}}_{(X\boldsymbol{\theta})^T X} &= 0, j = 1, \ldots, k
\end{aligned}
$$

Example: Estimating linear model

$$
\begin{aligned}
(X\widehat{\boldsymbol{\theta}})^T X &= \mathbf{y}^T X \\
X^T(X\widehat{\boldsymbol{\theta}}) = X^T X\widehat{\boldsymbol{\theta}} &= X^T\mathbf{y} \\
\widehat{\boldsymbol{\theta}} &= (X^T X)^{-1} X^T \mathbf{y}
\end{aligned}
$$

- $(X^T X)^{-1} X^T = X^+$: Moore–Penrose pseudo-inverse of X. Need to invert $X^T X$.
- $\widehat{\boldsymbol{\theta}}$: does not depend on σ^2

$$
\begin{aligned}
\frac{\partial \ln p(\mathbf{y}; X, \widehat{\boldsymbol{\theta}}, \sigma^2)}{\partial \sigma^2} &= \frac{-n}{2} \cdot \frac{2\pi}{2\pi\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (y_i - \mathbf{x}_i \cdot \widehat{\boldsymbol{\theta}})^2 = 0 \\
\hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i \cdot \widehat{\boldsymbol{\theta}})^2
\end{aligned}
$$

Example: Estimating linear model

MLE estimate of linear model with WGN - properties

$$\text{Recall } \mathbf{y} = X\boldsymbol{\theta} + \boldsymbol{\xi} \text{ and } X^{+} = (X^T X)^{-1} X^T$$

$$\begin{aligned} E(\widehat{\boldsymbol{\theta}}) &= E(X^{+}\mathbf{y}) \\ &= E((X^T X)^{-1} X^T X \boldsymbol{\theta} + X^{+}\boldsymbol{\xi}) \\ &= E(\boldsymbol{\theta} + X^{+}\boldsymbol{\xi}) = \boldsymbol{\theta} \end{aligned}$$

since $E(X^{+}\boldsymbol{\xi}) = 0$, $X^{+}\boldsymbol{\xi}$ is a linear combination of iid Gaussians with zero mean

$\Rightarrow$ MLE estimator $\widehat{\boldsymbol{\theta}}$ is unbiased

MLE estimate is also MVU

MLE estimate of linear model with WGN - properties

$$\frac{\partial \ln p(\mathbf{y}; X, \boldsymbol{\theta}, \sigma^2)}{\partial \theta_j} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{x}_i \cdot \boldsymbol{\theta}) x_{ij}$$

$$\frac{\partial \ln p(\mathbf{y}; X, \boldsymbol{\theta}, \sigma^2)}{\partial \boldsymbol{\theta}} = \frac{1}{\sigma^2} \left(X^T \mathbf{y} - X^T X \boldsymbol{\theta} \right)$$

$$= \underbrace{\frac{1}{\sigma^2}(X^T X)}_{I(\boldsymbol{\theta})} \left(\underbrace{(X^T X)^{-1} X^T \mathbf{y}}_{\widehat{\boldsymbol{\theta}}} - \boldsymbol{\theta} \right)$$

- Cramer-Rao necessary and sufficient condition (from lecture on estimation)
- $\Rightarrow$ MLE estimator $\widehat{\boldsymbol{\theta}}$ is MVU