Maximum Likelihood Estimation

Albert Dorador

UW-Madison

March 4, 2025

Introduction

- Formally introduced over 100 years ago (Fisher, 1922) but still relevant to this day
- Useful to estimate the parameters of a probabilistic model (density estimation): find the parameter values that best fit the data observed
- More precisely: find the (log) likelihood of θ, termed $L(\theta)$ (i.e. the joint pdf when $x_1, \ldots, x_n$ are viewed as fixed) and find the $\hat{\theta}$ that maximizes it. That is: $\hat{\theta} \in \arg\max_{\theta \in \Theta} p(\theta | x_1, \ldots, x_n)$ i.e. under i.i.d., $\hat{\theta} \in \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} p(\theta | x_i)$
- Vast majority of times, this is an unconstrained optimization problem: set 1st θ-derivative equal to 0. BUT sometimes that doesn't work (e.g. no stationary points or there are boundary constraints).
- Always check that the $\hat{\theta}$ you found is indeed a maximizer! Usual way: check $d^2 L(\theta)/d\theta^2 < 0$ when evaluated at $\theta = \hat{\theta}$

Example 1: Poisson

Let $X_1, \ldots, X_n$ i.i.d. Poisson(λ) having pmf $p_\lambda(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} I(k \in \mathbb{N})$ with $\lambda > 0$. Then,

$$\arg\max_{\lambda > 0} \prod_{i=1}^{n} p_\lambda(k_i) = \arg\max_{\lambda > 0} \log \left(\prod_{i=1}^{n} p_\lambda(k_i) \right) = \arg\max_{\lambda > 0} \sum_{i=1}^{n} \log \left(p_\lambda(k_i) \right)$$

Plugging in the assumed pmf and simplifying,

$$\arg\max_{\lambda > 0} \sum_{i=1}^{n} \log \left(\frac{\lambda^{k_i}}{k_i!} e^{-\lambda} \right) = \arg\max_{\lambda > 0} \sum_{i=1}^{n} \left[k_i \log(\lambda) - \log(k_i!) - \lambda \right]$$

Differentiating with respect to λ and setting to zero (to find the stationary point) we find that $\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} k_i$. Check $\hat{\lambda} > 0$ and negative second derivative.

Example 2: Uniform

Let $X_1, \ldots, X_n$ i.i.d. $U(0, \theta)$ having pdf $p_\theta(X = x) = \frac{1}{\theta} I(0 \leq x \leq \theta)$ with $\theta > 0$. Then,

$$\max_{\theta > 0} \prod_{i=1}^{n} p_\theta(x_i) = \max_{\theta > 0} \prod_{i=1}^{n} \frac{1}{\theta} I(0 \leq x_i \leq \theta)$$

Observe that the target function is 0 if $0 < \theta < x_i$ for some i, and hence we must impose $\theta \geq x_i$ for all i i.e. $\theta \geq \max_i x_i$. Under this constraint, the above maximization problem becomes

$$\max_{\theta > 0} \prod_{i=1}^{n} \frac{1}{\theta} = \max_{\theta > 0} \frac{1}{\theta^n}$$

which yields $\hat{\theta} = \max_i x_i$, as θ^{-n} is a decreasing function of θ. Check $\hat{\theta} > 0$.

Example 3: Linear regression

We observe the dataset $\{x_i, y_i\}_{i=1}^n$ and suppose $y_i = \beta^T x_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$ for $i = 1, \ldots, n$, and $x_i \in \mathbb{R}^p, y_i \in \mathbb{R}$. Let's find the MLE of β. First, as ϵ_i is assumed Gaussian, y_i is Gaussian given X, so

$$p_\beta(y|X) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta^T x_i)^2\right) = \exp\left(-\frac{1}{2\sigma^2}||y - X\beta||_2^2\right)$$

Hence, the MLE $\hat{\beta} \in \mathbb{R}^p$ is the (arg) solution to

$$\max_\beta p_\beta(y|X) = \min_\beta -\log p_\theta(y|X) = \min_\beta ||y - X\beta||_2^2 = \left(X^T X\right)^{-1} X^T y$$

where the last step assumes X has full rank, and uses matrix calculus. Here's a quick refresher for $\beta, a \in \mathbb{R}^p$ and $A \in \mathbb{R}^{p\times p}$:

$$\frac{d}{d\beta}\beta^T a = a \ , \ \frac{d}{d\beta}\beta^T A\beta = 2A\beta \ , \ \frac{d}{d\beta}A\beta = A$$

Connection with KL divergence?

To be continued next class…