UPPSALA
UNIVERSITET

Statistical Machine Learning

Lecture 2 – Maximum likelihood refresher

UPPSALA
UNIVERSITET

Sebastian Mair
https://smair.github.io/
Department of Information Technology
Uppsala University

Course webpage

Maximum likelihood

Let $x_1, x_2, \ldots, x_n$ be a sample of n iid random variables $X_1, X_2, \ldots, X_n$ with pmf $\mathbf{P}_\theta(x)$ (or pdf $f_\theta(x)$) parametrized by $\theta \in \Theta$.

Goal: Estimate θ based on the sample.

Idea: Choose $\theta^\star$ that maximizes the joint probability of the observed data.

$$\theta^\star = \underset{\theta \in \Theta}{\arg\max}\ \mathbf{P}_\theta[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n].$$

Maximum likelihood

$$
\begin{aligned}
&\mathbf{P}_\theta[X_1 = x_1, X_2 = x_2, \dots, X_n = x_n] \\
\overset{\text{iid}}{=}\, &\mathbf{P}_\theta[X_1 = x_1] \cdot \mathbf{P}_\theta[X_2 = x_2] \cdot \ldots \cdot \mathbf{P}_\theta[X_n = x_n] \\
= &\prod_{i=1}^{n} \mathbf{P}_\theta[X_i = x_i] =: \mathcal{L}(\theta).
\end{aligned}
$$

We call $\mathcal{L}(\theta)$ the **likelihood function**. Our estimate is now given by

$$\theta^\star = \underset{\theta \in \Theta}{\arg\max}\; \mathcal{L}(\theta)$$

UPPSALA UNIVERSITET

Maximum likelihood

Instead of the product, we often maximize the the logarithmized version $\ell(\theta) = \log \mathcal{L}(\theta)$, called the **log-likelihood function**. Thus,

$$\theta^\star = \underset{\theta \in \Theta}{\arg\max}\ \ell(\theta)$$

Note: Using the probability $\mathbf{P}_\theta[X_i = x_i]$ does not makes sense in the continuous case as it is zero for all x_i. Instead, we use the pdf $f_\theta(x_i)$.

Maximum likelihood - Example: Poisson distribution

Assumption: The sample $x_1, x_2, \ldots, x_n$ is iid and follows a Poisson distribution.

Reminder: The pmf of a Poisson distribution with parameter $\lambda > 0$ is given by

$$\mathbf{P}(X = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda} \quad (k \in \mathbb{N}_0).$$

Maximum likelihood - Example: Poisson distribution

The likelihood function is given by

$$\mathcal{L}(\lambda) = \mathbf{P}_\lambda[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] \overset{\text{iid}}{=} \prod_{i=1}^{n} \mathbf{P}_\lambda[X_i = x_i] = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} \cdot e^{-\lambda}$$

and the corresponding log-likelihood function is

$$\begin{aligned}
\ell(\lambda) = \log \mathcal{L}(\lambda) &= \sum_{i=1}^{n} \log\left(\frac{\lambda^{x_i}}{x_i!} \cdot e^{-\lambda}\right) = \sum_{i=1}^{n} \left[\log\left(\frac{\lambda^{x_i}}{x_i!}\right) + \log\left(e^{-\lambda}\right)\right] \\
&= \sum_{i=1}^{n} \left[\log\left(\lambda^{x_i}\right) - \log(x_i!) - \lambda\right] = \sum_{i=1}^{n} x_i \log(\lambda) - \sum_{i=1}^{n} \log(x_i!) - n\lambda \\
&= \log(\lambda) \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \log(x_i!) - n\lambda
\end{aligned}$$

Maximum likelihood - Example: Poisson distribution

$$\ell(\lambda) = \log(\lambda)\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\log(x_i!) - n\lambda$$

We can maximize this log-likelihood function by setting the derivative

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\ell(\lambda) = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n$$

to zero and re-arrange the terms

$$\frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0 \implies \frac{1}{\lambda}\sum_{i=1}^{n} x_i = n \implies \sum_{i=1}^{n} x_i = n\lambda \implies \lambda = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}.$$

Thus, our estimate for λ is given by $\frac{1}{n}\sum_{i=1}^{n} x_i$.

UPPSALA UNIVERSITET

Maximum likelihood - Example: Normal distribution

Assumption: The sample $x_1, x_2, \ldots, x_n$ is iid and follows a normal distribution.

Reminder: The pdf of a normal distribution with param. $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ is given by

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

Maximum likelihood - Example: Normal distribution

$$
\begin{aligned}
\ell(\mu, \sigma^2) = \log \mathcal{L}(\mu, \sigma^2) &= \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2} \right) \\
&= \sum_{i=1}^{n} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2} \right) \right] \\
&= \sum_{i=1}^{n} \underbrace{\log 1}_{=0} - \log \sqrt{2\pi\sigma^2} - \frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2} \\
&= \sum_{i=1}^{n} -\frac{1}{2} \log 2\pi - \log \sigma + -\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2} \\
&= -\frac{n}{2} \log 2\pi - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 .
\end{aligned}
$$

Maximum likelihood - Example: Normal distribution

For μ, we derive

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \mu} = \frac{\partial}{\partial \mu}\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \right] = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{\partial}{\partial \mu}\left[(x_i - \mu)^2\right]$$

$$= -\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)(-1) \overset{!}{=} 0.$$

Hence,

$$0 = \sum_{i=1}^{n}(x_i - \mu) \implies 0 = \sum_{i=1}^{n} x_i - n\mu \implies \mu = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Maximum likelihood - Example: Normal distribution

For σ^2, we derive

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma} = \frac{\partial}{\partial \sigma}\left[- n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right]$$
$$= -n \frac{\partial}{\partial \sigma} \log \sigma - \frac{1}{2} \frac{\partial}{\partial \sigma} \sigma^{-2} \sum_{i=1}^{n} (x_i - \mu)^2$$
$$= -\frac{n}{\sigma} - \frac{-2}{2} \sigma^{-3} \sum_{i=1}^{n} (x_i - \mu)^2 = -\frac{n}{\sigma} + \sigma^{-3} \sum_{i=1}^{n} (x_i - \mu)^2 \overset{!}{=} 0.$$

Hence,

$$\frac{n}{\sigma} = \sigma^{-3} \sum_{i=1}^{n} (x_i - \mu)^2 \implies \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2.$$

UPPSALA UNIVERSITET