Maximum Likelihood Estimate (MLE) and the Gaussian model

S. Marchand-Maillet

Computer Science – University of Geneva

Given $\mathcal{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ where $\boldsymbol{x}_i \subset \mathbb{X} \subseteq \mathbb{R}^D$, we wish to model the generative process of $\mathcal{X}$. The Maximum Likelihood criterion:

1. assumes a model $p_{\boldsymbol{\theta}}$ with parameters $\boldsymbol{\theta}$ for the generative density (of which the data is supposed to be a sample)
2. selects its optimal parameters $\boldsymbol{\theta}^*$ such that the likelihood of $\mathcal{X}$ as a sample of $p_{\boldsymbol{\theta}^*}$ is maximized over all possible parameters.

Definition 1 (Maximum Likelihood Estimate of parameters)**.** Given a dataset $\mathcal{X} \subset \mathbb{X}$, a density model $p_{\boldsymbol{\theta}}$ with parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the Maximum Likelihood Estimate (MLE) parameters $\boldsymbol{\theta}^*$ maximize the likelihood of data $\mathcal{X}$ under sampling of density $p_{\boldsymbol{\theta}}$ and are given by:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}}\, p_{\boldsymbol{\theta}}(\mathcal{X})$$

Now,

$$p_{\boldsymbol{\theta}}(\mathcal{X}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N) = p_{\boldsymbol{\theta}}(\boldsymbol{x}_1) \Pi_{i=2}^{N} p_{\boldsymbol{\theta}}(\boldsymbol{x}_i | \boldsymbol{x}_{j<i})$$

We therefore assume that the data is identically and independently distributed (i.i.d) in order to alleviate the conditional dependence between data points:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_i | \boldsymbol{x}_{j<i}) \overset{\text{i.i.d}}{=} p_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \quad \Rightarrow \quad p_{\boldsymbol{\theta}}(\mathcal{X}) \overset{\text{i.i.d}}{=} \Pi_{i=1}^{N} p_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$$

Hence:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}}\, p_{\boldsymbol{\theta}}(\mathcal{X}) \overset{\text{i.i.d}}{=} \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}}\, \Pi_{i=1}^{N} p_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \tag{1}$$

Assuming a Gaussian generative model:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})} \qquad \boldsymbol{\theta} = [\boldsymbol{\mu}, \boldsymbol{\Sigma}]$$

which, in the case where $D = 1$ reduces to

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_i) = \mathcal{N}(\boldsymbol{x}_i | \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x_i - \mu)^2}{\sigma^2}} \qquad \boldsymbol{\theta} = [\mu, \sigma]$$

Inserting into (1):

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}}\, \Pi_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x_i - \mu)^2}{\sigma^2}}$$

We further write

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \log \left(\Pi_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x_i - \mu)^2}{\sigma^2}} \right) \qquad \text{since } \log(.) \text{ is a monotonically increasing function}$$

Denote:

$$\begin{aligned}
\mathcal{L}_{\boldsymbol{\theta}}^{\log}(\mathcal{X}) &= \log \left(\Pi_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x_i - \mu)^2}{\sigma^2}} \right) \\
&= \sum_{i=1}^{N} \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x_i - \mu)^2}{\sigma^2}} \right) \\
&= \sum_{i=1}^{N} -\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2}\frac{(x_i - \mu)^2}{\sigma^2} \\
&= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{N}\frac{(x_i - \mu)^2}{\sigma^2} \\
&= -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{1}{2}\sum_{i=1}^{N}\frac{(x_i - \mu)^2}{\sigma^2}
\end{aligned}$$

so that

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \mathcal{L}_\theta^{\log}(\mathcal{X}) = \underset{\theta \in \Theta}{\operatorname{argmax}} \left[-\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{1}{2}\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{\sigma^2} \right]$$

and

$$\left.\frac{\partial \mathcal{L}_\theta^{\log}(\mathcal{X})}{\partial \theta}\right|_{\theta=\theta^*} = 0$$

Now,

$$\frac{\partial \mathcal{L}_\theta^{\log}(\mathcal{X})}{\partial \mu} = -\frac{1}{2}\sum_{i=1}^{N} 2(x_i-\mu) = -\sum_{i=1}^{N}(x_i-\mu)$$

so that

$$\frac{\partial \mathcal{L}_\theta^{\log}(\mathcal{X})}{\partial \mu} = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{N}(x_i-\mu) = 0 \quad \Leftrightarrow \quad \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Similarly,

$$\frac{\partial \mathcal{L}_\theta^{\log}(\mathcal{X})}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{N}(x_i-\mu)^2$$

so that

$$\frac{\partial \mathcal{L}_\theta^{\log}(\mathcal{X})}{\partial \sigma} = 0 \quad \Leftrightarrow \quad -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{N}(x_i-\mu)^2 = 0 \quad \Leftrightarrow \quad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

We see that the classical empirical mean (μ) and variance (σ^2) computation correspond to the Maximum Likelihood Estimate of the data against a Gaussian model.