Note 20. MLE and MAP estimation

Alex Fu

Fall 2022

1 Maximum likelihood estimation

We now turn away from the problem of hypothesis testing, determining which of several given hypotheses, distributions, or classes is most likely, to the similar problem of **estimation**, finding the most likely value out of a usually larger space of possible values.

Definition 1 (Setup for maximum likelihood estimation)**.**

Let X be a random variable for the underlying state or true parameter, and let $X \sim \pi$ be the **prior** distribution. Y is a random variable for the given observation, which is drawn from the **model** $p_{Y|X}$. The **posterior** distribution of X is $p_{X|Y}$, where

$$p_{X|Y}(x \mid y) = \mathbb{P}(X = x \mid Y = y) = \frac{p_{Y|X}(y \mid x) \cdot \pi(x)}{\sum_{\tilde{x} \in S} p_{Y|X}(y \mid \tilde{x}) \cdot \pi(\tilde{x})}$$

by Bayes' rule and the law of total probability. Note that $p_{X|Y} \propto p_{Y|X} \cdot \pi$, as the *normalizing constant* $\mathbb{P}(Y = y)$ is fixed by y. In particular, if π is uniform or *uninformative* — we have no information about the distribution of X — then $p_{X|Y} \propto p_{Y|X}$.

Definition 2 (Maximum likelihood estimator)**.**

The **maximum likelihood estimator** (MLE) of X given Y is the function of Y

$$\hat{X}_{\mathsf{MLE}}(y) = \operatorname*{argmax}_x \mathbb{P}(Y = y \mid X = x).$$

Note that we do not have to find any probabilities to find the MLE, as the model $p_{Y|X}$ is given.

A common informal principle is that the MLE estimate is most often "the parameter that makes the most sense," such as the expected value or maximum, as we find in the following examples. The MLE can thus be thought of as the "most likely explanation" $X = x$ for $Y = y$.

Proposition 1 (MLE minimizes KL divergence asymptotically*).

We use a different conventional notation $(\theta, X) \leftarrow (X, Y)$. Let θ^* be an unknown parameter, and let $x_1, \dots, x_n$ be i.i.d. samples drawn from the distribution $\mathbb{P}_{\theta^*} \colon x \mapsto p(x \mid \theta^*)$. Then the MLE of θ is the parameter which minimizes $D_{\mathrm{KL}}(\mathbb{P}_{\hat{\theta}_{\mathsf{MLE}}} \parallel \mathbb{P}_{\theta^*})$ as $n \to \infty$, and $\hat{\theta}_{\mathsf{MLE}} = \theta^*$ if θ^* belongs to the space of estimation.

Proof. We maximize the log-likelihood function and use the law of large numbers to find

$$\begin{aligned}
\hat{\theta}_{\mathsf{MLE}}(x_1, \dots, x_n) = \operatorname*{argmax}_\theta \prod_{i=1}^n p(x_i \mid \theta) &= \operatorname*{argmax}_\theta \sum_{i=1}^n (\ln p(x_i \mid \theta) - \ln p(x_i \mid \theta^*)) \\
&= \operatorname*{argmin}_\theta \frac{1}{n} \sum_{i=1}^n \ln \frac{p(x_i \mid \theta^*)}{p(x_i \mid \theta)} \\
&\to \operatorname*{argmin}_\theta \mathbb{E}\left(\ln \frac{p(x \mid \theta^*)}{p(x \mid \theta)} \right) \\
&= \operatorname*{argmin}_\theta \int p(x \mid \theta^*) \ln \frac{p(x \mid \theta^*)}{p(x \mid \theta)} \, dx \\
&= \operatorname*{argmin}_\theta D_{\mathrm{KL}}(\mathbb{P}_{\theta^*} \parallel \mathbb{P}_\theta) \\
&= \operatorname*{argmin}_\theta H(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta).
\end{aligned}$$

Recall that the cross entropy is $H(\mathbb{P}_{\theta^*}, \mathbb{P}_\theta) = H(\mathbb{P}_{\theta^*}) + D_{\mathrm{KL}}(\mathbb{P}_{\theta^*} \parallel \mathbb{P}_\theta)$. In the interpretation of KL divergence as "distance between distributions" and cross entropy as "entropy when the wrong distribution is assumed," the "most likely" distribution $\mathbb{P}_{\hat{\theta}_{\mathsf{MLE}}}$ is also "closest" to $\mathbb{P}_{\theta^*}$. □

Example 1 (Parameter of uniform distribution).

Let $X \geq 0$ be unknown, and let $Y \sim \mathsf{Uniform}([0, X])$. Then the MLE of X given Y is

$$\hat{X}_{\mathsf{MLE}}(y) = \operatorname*{argmax}_x f_{Y|X}(y \mid x) = \operatorname*{argmax}_x \frac{1}{x} \cdot \mathbb{1}_{0 \leq y \leq x} = y.$$

That is, the MLE is simply the observation, as $Y = y$ is most likely given $X = Y$.

Example 2 (Mean of normal distribution I).

Let $X = \mu$, and let $Y \sim \mathcal{N}(\mu, \sigma^2)$. Then the MLE of X given Y is

$$\hat{X}_{\mathsf{MLE}}(y) = \operatorname*{argmax}_x \ln f_{Y|X}(y \mid x) = \operatorname*{argmax}_x \left(-\frac{(y - x)^2}{2\sigma^2} \right) = y.$$

Example 3 (Mean of normal distribution II).

Let $X = \mu$, and now let $Y_1, \ldots, Y_n$ be i.i.d. samples drawn from $\mathcal{N}(\mu, \sigma^2)$. Then

$$\hat{X}_{\mathsf{MLE}}(y_1, \ldots, y_n) = \operatorname*{argmax}_x \prod_{i=1}^{n} \exp\left(-\frac{(y_i - x)^2}{2\sigma^2}\right) = \operatorname*{argmin}_x \sum_{i=1}^{n} (y_i - x)^2$$

is the minimizer of the *ordinary least squares error*, which we can find by differentiation:

$$\frac{d}{dx} \sum_{i=1}^{n} (y_i - x)^2 = -\sum_{i=1}^{n} 2(y_i - x) = 0$$

when x is the average of the observations. That is,

$$\hat{\mu}_{\mathsf{MLE}}(y_1, \ldots, y_n) = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

Example 4 (Parameters of random graph I).

Let $p \in [0, 1]$, let $X = n$, and let $Y = m$ be the number of edges in $G \sim \mathcal{G}(n, p)$. Then

$$\hat{n}_{\mathsf{MLE}}(m) = \operatorname*{argmax}_n \ln\left(p^m \cdot (1-p)^{\binom{n}{2} - m}\right) = \operatorname*{argmax}_n \left(\binom{n}{2} - m\right) \ln(1-p)$$

is the largest value of $n \in \mathbb{N}$ so that $\binom{n}{2} \leq m$.

Example 5 (Parameters of random graph II).

Let $X = p \in [0, 1]$ be unknown, and suppose that $Y = G \sim \mathcal{G}(n, p)$ has m edges. Then

$$\hat{p}_{\mathsf{MLE}}(G) = \operatorname*{argmax}_p \left[m \cdot \ln(p) + \left(\binom{n}{2} - m\right) \cdot \ln(1-p)\right]$$

can be found by differentiation with respect to p:

$$\begin{aligned} \frac{d}{dp}\left[m \cdot \ln(p) + \left(\binom{n}{2} - m\right) \cdot \ln(1-p)\right] &= \frac{m}{p} + \frac{m}{1-p} - \frac{\binom{n}{2}}{(1-p)} \\ &= \frac{1}{p(1-p)}\left(m - \binom{n}{2} p\right) \end{aligned}$$

is 0 when $m = \binom{n}{2} p$, which is precisely the expected number of edges. That is,

$$\hat{p}_{\mathsf{MLE}}(G) = \frac{m}{\binom{n}{2}}.$$

2 Maximum a posteriori estimation

Definition 3 (Maximum a posteriori estimator).

The **maximum a posteriori** (MAP) estimate of X given by Y is the function of Y

$$\hat{X}_{\mathsf{MAP}}(y) = \operatorname*{argmax}_x \mathbb{P}(X = x \mid Y = y) = \operatorname*{argmax}_x \mathbb{P}(X = x, Y = y).$$

The MLE is the special case of the MAP when the prior distribution is uniform. In general, the MAP estimate may not be the same as the MLE, as the posterior $p_{X|Y}$ or joint distribution $p_{X,Y}$ is weighted by the prior π, which we find in the following examples.

Example 6 (Parameter of uniform distribution, again).

The MAP estimate of $X \sim \mathsf{Uniform}([0,1])$ given $Y \sim \mathsf{Uniform}([0,X])$ is

$$\hat{X}_{\mathsf{MAP}}(y) = \operatorname*{argmax}_x f_{Y|X}(y \mid x) \cdot \pi(x) = \operatorname*{argmax}_x f_{Y|X}(y \mid x) = \hat{X}_{\mathsf{MLE}}(y) = y.$$

If the prior is instead $X \sim \mathsf{Bernoulli}(\frac{1}{4}) + 1$, then

$$\hat{X}_{\mathsf{MAP}}(y) = \operatorname*{argmax}_x \frac{1}{4} \cdot \mathbb{1}_{x=1, 0 \le y \le 1} + \frac{3}{8} \cdot \mathbb{1}_{x=2, 0 \le y \le 2} = 2.$$

Example 7 (Mean of normal distribution, again).

Now suppose that $X = \mu \sim \mathcal{N}(\mu_*, \sigma_*^2)$, and let $Y_1, \ldots, Y_n \sim \mathcal{N}(X, \sigma_i^2)$ i.i.d. Then

$$\begin{aligned}
\hat{X}_{\mathsf{MAP}}(y_1, \ldots, y_n) &= \operatorname*{argmax}_x \ln \pi(x) + \sum_{i=1}^n \ln f_{Y_i|X}(y_i \mid x) \\
&= \operatorname*{argmin}_x \frac{(x - \mu_*)^2}{\sigma_*^2} + \sum_{i=1}^n \frac{(y_i - x)^2}{\sigma_i^2} \\
\frac{d}{dx}\left[\frac{(x - \mu_*)^2}{\sigma_*^2} + \sum_{i=1}^n \frac{(y_i - x)^2}{\sigma_i^2}\right] &= \frac{2(x - \mu_*)}{\sigma_*^2} + \sum_{i=1}^n \frac{2(x - y_i)}{\sigma_i^2} \\
&= 2\left(\frac{1}{\sigma_*^2} + \sum_{i=1}^n \frac{1}{\sigma_i^2}\right) x - 2\left(\frac{\mu_*}{\sigma_*^2} + \sum_{i=1}^n \frac{y_i}{\sigma_i^2}\right) \\
\hat{\mu}_{\mathsf{MAP}}(y_1, \ldots, y_n) &= \frac{1}{\frac{1}{\sigma_*^2} + \sum_{i=1}^n \frac{1}{\sigma_i^2}} \left(\frac{\mu_*}{\sigma_*^2} + \sum_{i=1}^n \frac{y_i}{\sigma_i^2}\right).
\end{aligned}$$