Maximum Likelihood Estimation

Wilker Aziz

February 9, 2018

Notation We use capital Roman letters (e.g. X) for random variables (rvs) and lowercase letters for assignments (e.g. x). We use X_1^n as a shorthand for $X_1, \ldots, X_n$ and similarly with x_1^n. We write P_X for probability distributions, and $P_X(X = x)$ for probability values—where we sometimes omit one or both occurrences of X, e.g. $P_X(x)$, $P(X = x)$, or $P(x)$, if no ambiguity is possible. We denote a probability mass function (pmf) by $p(x; \alpha)$, where α is a deterministic set of parameters. Throughout, we also assume that argmax returns a single point.

Assume we have a dataset of n iid observations $\mathcal{D} = \{x_1, \ldots, x_n\}$ of an rv $X \sim P_X$, i.e. $(X_i \sim P_X)_{i=1}^n$. First of all, from independence, we know that

$$P_{X_1^n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} P_{X_i}(x_i) = \prod_{i=1}^{n} P_X(x_i) \tag{1}$$

and we then model the probability $P_X(x)$ with a member $p(x; \alpha)$ of a parametric family and proceed to derive a maximum likelihood estimate of α. In the following sections we use $\mathcal{L}(\alpha|\mathcal{D})$ for the log-likelihood function

$$\mathcal{L}(\alpha|\mathcal{D}) = \sum_{i=1}^{n} \log p(x_i; \alpha) \tag{2}$$

and we often omit the dependency on data writing simply $\mathcal{L}(\alpha)$. Our objective is then

$$\theta^\star = \operatorname*{argmax}_{\theta \in \Theta} \; \mathcal{L}(\theta) \tag{3}$$

where the space of valid parameters Θ is possibly subject to constraints.

1 Bernoulli

Suppose X takes on values in the set $\{0, 1\}$, then we say X is Bernoulli-distributed

$$X \sim \text{Bern}(\theta) \tag{4}$$

where $0 < \theta < 1$ the Bernoulli parameter. The Bernoulli pmf is

$$p(x; \theta) = \text{Bern}(X = x|\theta) = \theta^x (1 - \theta)^{(1-x)} \tag{5}$$

and therefore the Bernoulli parameter corresponds to the probability of the positive class—i.e. $P_X(X = 1) = \theta$.

We now derive the maximum likelihood estimate of the parameter θ. We start by rewriting the objective (3) in terms for the Bernoulli pmf (4)

$$\begin{aligned}
\theta^\star &= \underset{\theta \in [0,1]}{\text{argmax}} \; \mathcal{L}(\theta) && \text{(6a)} \\
&= \underset{\theta \in (0,1)}{\text{argmax}} \; \sum_{i=1}^{n} \log p(x_i; \theta) && \text{(6b)} \\
&= \underset{\theta \in (0,1)}{\text{argmax}} \; \sum_{i=1}^{n} x_i \log \theta + (1 - x_i) \log(1 - \theta) && \text{(6c)} \\
&= \underset{\theta \in (0,1)}{\text{argmax}} \; \log \theta \underbrace{\left(\sum_{i=1}^{n} x_i \right)}_{n_1} + \log(1 - \theta) \underbrace{\left(\sum_{i=1}^{n} 1 - x_i \right)}_{n_0} && \text{(6d)} \\
&= \underset{\theta \in (0,1)}{\text{argmax}} \; n_1 \log \theta + n_0 \log(1 - \theta) && \text{(6e)}
\end{aligned}$$

where we use n_1 for the number of positive observations and n_0 for the number of negative observations—and note that $n = n_1 + n_0$ is the total number of observations.

Now we find the first derivative of $\mathcal{L}(\theta)$ with respect to θ:

$$\begin{aligned}
\frac{\mathrm{d}\mathcal{L}(\theta)}{\mathrm{d}\theta} &= \frac{\mathrm{d}}{\mathrm{d}\theta} \left[n_1 \log \theta + n_0 \log(1 - \theta) \right] && \text{(7a)} \\
&= n_1 \frac{\mathrm{d}}{\mathrm{d}\theta} \log \theta + n_0 \frac{\mathrm{d}}{\mathrm{d}\theta} \log(1 - \theta) && \text{(7b)} \\
&= n_1 \frac{\mathrm{d}}{\mathrm{d}\theta} \log \theta + n_0 \underbrace{\frac{\mathrm{d}}{\mathrm{d}(1 - \theta)} \log(1 - \theta) \frac{\mathrm{d}(1 - \theta)}{\mathrm{d}\theta}}_{\text{chain rule of derivatives}} && \text{(7c)} \\
&= \frac{n_1}{\theta} + \frac{n_0}{1 - \theta}(-1) && \text{(7d)}
\end{aligned}$$

Setting the derivative to 0 and solving for θ gives us the MLE solution.

$$0 = \frac{n_1}{\theta} + \frac{n_0}{1-\theta}(-1) \tag{8a}$$
$$= \frac{n_1(1-\theta) - n_0\theta}{\theta(1-\theta)} \tag{8b}$$
$$= n_1(1-\theta) - n_0\theta \tag{8c}$$
$$= n_1 - n_1\theta - n_0\theta \tag{8d}$$
$$n_1 = \theta(n_1 + n_0) \tag{8e}$$
$$\theta = \frac{n_1}{n_1 + n_0} \tag{8f}$$
$$= \frac{n_1}{n} \tag{8g}$$

Note that in (8b) the denominator is never zero because the Bernoulli parameter can never be 0 or 1.

2 Categorical

Suppose X takes on values in the discrete interval $[1, k]$, then we say X is Categorically-distributed

$$X \sim \mathrm{Cat}(\theta_1, \dots, \theta_k) \tag{9}$$

where $\theta_1, \dots, \theta_k$ are the Categorical parameters subject to $\theta_x > 0$ and $\sum_{x=1}^{k} \theta_x = 1$. The Categorical pmf is

$$p(a; \theta_1^k) = \mathrm{Cat}(X = a|\theta_1, \dots, \theta_k) = \prod_{x=1}^{k} \theta_x^{\delta_{xa}} \tag{10}$$

where δij is the Kronecker delta and therefore each Categorical parameter corresponds to the probability of the respective category—i.e. $P_X(X = x) = \theta_x$.

We now derive the maximum likelihood estimate of the parameters θ_1^k. We start by rewriting the objective (3) in terms for the Categorical pmf (4)

$$
\begin{aligned}
\theta^\star &= \underset{\theta_1^k \in \Delta}{\operatorname{argmax}} \;\; \mathcal{L}(\theta_1^k) && \text{(11a)}\\
&= \underset{\theta_1^k \in \Delta}{\operatorname{argmax}} \;\; \sum_{i=1}^{n} \log p(x_i; \theta_1^k) && \text{(11b)}\\
&= \underset{\theta_1^k \in \Delta}{\operatorname{argmax}} \;\; \sum_{i=1}^{n} \log \prod_{x=1}^{k} \theta_x^{\delta_{x x_i}} && \text{(11c)}\\
&= \underset{\theta_1^k \in \Delta}{\operatorname{argmax}} \;\; \sum_{i=1}^{n} \sum_{x=1}^{k} \delta_{x x_i} \log \theta_x && \text{(11d)}\\
&= \underset{\theta_1^k \in \Delta}{\operatorname{argmax}} \;\; \sum_{x=1}^{k} \log \theta_x \underbrace{\sum_{i=1}^{n} \delta_{x x_i}}_{n_x} && \text{(11e)}\\
&= \underset{\theta_1^k \in \Delta}{\operatorname{argmax}} \;\; \sum_{x=1}^{k} n_x \log \theta_x && \text{(11f)}
\end{aligned}
$$

where we denote the number of observations of class $x \in [1, k]$ by n_x.

To avoid optimising the constrained objective, where $\sum_{x=1}^{k} \theta_x = 1$, we employ a Lagrange multiplier λ such that the new objective is

$$
\begin{aligned}
\theta^\star = \underset{\theta_1^k \in \mathbb{R}^k}{\operatorname{argmax}} \;\; & \underbrace{\mathcal{L}(\theta_1^k) - \lambda \left[\left(\sum_{x=1}^{k} \theta_x \right) - 1 \right]}_{\mathcal{L}(\theta_1^k, \lambda)} && \text{(12a)}\\
\text{s.t. } \; & \theta_x > 0 \text{ for } x \in [1, k]
\end{aligned}
$$

Now we find the first partial derivative of $L(\theta_1^k, \lambda)$ with respect to λ

$$
\begin{aligned}
\frac{\partial}{\partial \lambda} \mathcal{L}(\theta_1^k, \lambda) &= \frac{\partial}{\partial \lambda} \left\{ \sum_{x=1}^{k} n_x \log \theta_x - \lambda \left[\left(\sum_{x=1}^{k} \theta_x \right) - 1 \right] \right\} && \text{(13a)}\\
&= \cancelto{0}{\frac{\partial}{\partial \lambda} \left\{ \sum_{x=1}^{k} n_x \log \theta_x \right\}} - \left[\left(\sum_{x=1}^{k} \theta_x \right) - 1 \right] && \text{(13b)}\\
&= 1 - \sum_{x=1}^{k} \theta_x && \text{(13c)}
\end{aligned}
$$

where setting the derivative to zero yields

$$
\sum_{x=1}^{k} \theta_x = 1 \quad . \tag{14}
$$

We now turn to the first partial derivative of $L(\theta_1^k, \lambda)$ with respect to θ_j for $j \in [1, k]$

$$
\begin{aligned}
\frac{\partial}{\partial \theta_j} \mathcal{L}(\theta_1^k, \lambda) &= \frac{\partial}{\partial \theta_j} \left\{ \sum_{x=1}^{k} n_x \log \theta_x - \lambda \left[\left(\sum_{x=1}^{k} \theta_x \right) - 1 \right] \right\} && \text{(15a)} \\
&= \sum_{x=1}^{k} n_x \frac{\partial}{\partial \theta_j} \log \theta_x - \lambda \sum_{x=1}^{k} \frac{\partial}{\partial \theta_j} \theta_x && \text{(15b)} \\
&= \sum_{x=1}^{k} n_x \frac{\delta j x}{\theta_x} - \lambda \sum_{x=1}^{k} \delta_{jx} && \text{(15c)} \\
&= \frac{n_j}{\theta_j} - \lambda && \text{(15d)}
\end{aligned}
$$

where setting the derivative to zero yields

$$
\begin{aligned}
0 &= \frac{n_j}{\theta_j} - \lambda && \text{(16a)} \\
\lambda &= \frac{n_j}{\theta_j} && \text{(16b)} \\
\theta_j &= \frac{n_j}{\lambda} && \text{(16c)}
\end{aligned}
$$

Now substituting (16c) into (14) we have

$$
\sum_{x=1}^{k} \theta_x = \sum_{x=1}^{k} \frac{n_x}{\lambda} = \frac{1}{\lambda} \sum_{x=1}^{k} n_x = \frac{1}{\lambda} n = 1 \tag{17a}
$$

and therefore $\lambda = \frac{1}{n}$. And finally, substituting λ in (16c)

$$
\theta_j = \frac{n_j}{n} \tag{18}
$$

yields the maximum likelihood estimate. Note that $\theta_j > 0$ is satisfied as long as $n_j > 0$.