

Maximum Likelihood Methods

Charles Elkan
elkan@cs.ucsd.edu

November 5, 2007

Consider a family of probability distributions defined by a set of parameters θ . The distributions may be either probability mass functions (pmfs) or probability density functions (pdfs). Suppose we have a random sample drawn from a fixed but unknown member of this family. The random sample is a training set of n examples x_1 to x_n . We assume that the examples are independent so the probability of the set is the product of the probabilities of the individual examples:

$$f(x_1, \dots, x_n; \theta) = \prod_j f_\theta(x_j; \theta).$$

Previously we have thought of the distribution θ as fixed and the examples x_j as unknown, or varying. However, we can think of the training data as fixed and consider alternative parameter values. This is the point of view behind the definition of the likelihood function:

$$L(\theta; x_1, \dots, x_n) = f(x_1, \dots, x_n; \theta).$$

Note that if $f(x; \theta)$ is a probability mass function, then the likelihood is always less than one, but if $f(x; \theta)$ is a probability density function, then the likelihood can be greater than one, since densities can be greater than one.

The principle of maximum likelihood says that we should use as our model the distribution $f(\cdot; \hat{\theta})$ that gives the greatest possible probability to the training data. Formally,

$$\hat{\theta} = \operatorname{argmax}_\theta L(\theta; x_1, \dots, x_n).$$

This value $\hat{\theta}$ is called the maximum likelihood estimator (MLE) of θ . Note that in general each x_j is a vector of values, and θ is a vector of real-valued parameters. For example, for a Gaussian distribution $\theta = \langle \mu, \sigma^2 \rangle$.

As a first example of finding a maximum likelihood estimator, consider the parameter of a Bernoulli distribution. A random variable with this distribution is a formalization of a coin toss. The value of the random variable is 1 with probability θ and 0 with probability $1 - \theta$. Let X be a Bernoulli random variable. We have

$$P(X = x) = \begin{cases} \theta & \text{if } x = 1 \\ 1 - \theta & \text{if } x = 0 \end{cases}.$$

For mathematical convenience write this as

$$P(X = x) = \theta^x (1 - \theta)^{1-x}.$$

Suppose the training data are x_1 through x_n where each $x_j \in \{0, 1\}$. We maximize the likelihood function

$$L(\theta; x_1, \dots, x_n) = f(x_1, \dots, x_n; \theta) = \theta^h (1 - \theta)^{n-h}$$

where $h = \sum_i x_i$. The maximization is over the possible values $0 \leq \theta \leq 1$.

We can do the maximization by setting the derivative with respect to θ equal to zero. The derivative is

$$\begin{aligned} \frac{\partial}{\partial \theta} \theta^h (1 - \theta)^{n-h} &= h\theta^{h-1} (1 - \theta)^{n-h} + \theta^h (n - h) (1 - \theta)^{n-h-1} (-1) \\ &= \theta^{h-1} (1 - \theta)^{n-h-1} [h(1 - \theta) - (n - h)\theta] \end{aligned}$$

which has solutions $\theta = 0$, $\theta = 1$, and $\theta = h/n$. The solution which is a maximum is clearly $\theta = h/n$ while $\theta = 0$ and $\theta = 1$ are minima. So we have the maximum likelihood estimate $\hat{\theta}_{\text{MLE}} = h/n$.

The log likelihood function is simply the logarithm of the likelihood function. Because logarithm is a monotonic strictly increasing function, maximizing the log likelihood is precisely equivalent to maximizing the likelihood, or to minimizing the negative log likelihood.

For an example of minimizing the negative log likelihood (NLL), consider the problem of estimating the parameters of a univariate Gaussian distribution. This distribution is

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

The NLL for one example x is

$$l(\mu, \sigma^2; x) = \log L(\mu, \sigma^2; x) = -\log \sigma - \log \sqrt{2\pi} - \frac{(x - \mu)^2}{2\sigma^2}.$$

Suppose we have training data $\{x_1, \dots, x_n\}$. The maximum likelihood estimates are

$$\langle \hat{\mu}, \hat{\sigma}^2 \rangle = \operatorname{argmin}_{\langle \mu, \sigma^2 \rangle} \left[-n \log \sigma - n \log \sqrt{2\pi} - \frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2 \right].$$

This expression is to be minimized simultaneously over two variables, but we can simplify it into two sequential univariate minimizations. The first is

$$\hat{\mu} = \operatorname{argmin}_{\mu} \sum_i (x_i - \mu)^2$$

while the second is

$$\hat{\sigma}^2 = \operatorname{argmin}_{\sigma^2} \left[-n \log \sigma \sqrt{2\pi} - \frac{1}{2\sigma^2} T \right]$$

where $T = \sum_i (x_i - \hat{\mu})^2$. In order to do the first minimization, write $(x_i - \mu)^2$ as $(x_i - \bar{x} + \bar{x} - \mu)^2$. Then

$$\sum_i (x_i - \mu)^2 = \sum_i (x_i - \bar{x})^2 - 2(\bar{x} - \mu) \sum_i (x_i - \bar{x}) + n(\bar{x} - \mu)^2.$$

The first term $\sum_i (x_i - \bar{x})^2$ does not depend on μ so it is irrelevant to the minimization. The second term equals zero, because $\sum_i (x_i - \bar{x}) = 0$. The third term is always positive, so it is clear that it is minimized when $\mu = \bar{x}$.

To perform the second minimization, take the derivative and set it equal to zero:

$$\begin{aligned} \frac{\partial}{\partial \sigma} \left[-n \log \sigma - n \log \sqrt{2\pi} - \frac{1}{2} \sigma^{-2} T \right] &= -n \sigma^{-1} - \frac{1}{2} (-2 \sigma^{-3}) T \\ &= \sigma^{-1} (-n + T \sigma^{-2}) \\ &= 0 \text{ if } \sigma^2 = T/n. \end{aligned}$$

Maximum likelihood estimators are typically reasonable, but they may have issues. Consider the Gaussian variance estimator $\hat{\sigma}_{\text{MLE}}^2 = \sum_i (x_i - \bar{x})^2 / n$ and the case where $n = 1$. In this case $\hat{\sigma}_{\text{MLE}}^2 = 0$. This estimate is guaranteed to be too small. Intuitively, the estimate is optimistically assuming that all future data points x_2 and so on will equal x_1 exactly.

It can be proved that in general the maximum likelihood estimate of the variance of a Gaussian is too small, on average:

$$E \left[\frac{1}{n} \sum_i (x_i - \bar{x})^2; \mu, \sigma^2 \right] = \frac{n-1}{n} \sigma^2 < \sigma^2.$$

This phenomenon can be considered an instance of overfitting: the observed spread around the observed mean \bar{x} is less than the unknown true spread σ^2 around the unknown true mean μ .