Maximum Likelihood Estimation

Foundations of Data Analysis

February 4, 2020

The purpose of these notes is to review the definition of a *maximum likelihood estimate* (MLE), and show that the sample mean is the MLE of the μ parameter in a Gaussian. For more details about MLEs, see the Wikipedia article:

https://en.wikipedia.org/wiki/Maximum_likelihood (https://en.wikipedia.org/wiki/Maximum_likelihood)

Consider a random sample $X_1, X_2, \ldots, X_n$ coming from a distribution with parameter θ (for example, they could be from a Gaussian distribution with parameter μ). Remember the terminology "random sample" means that X_i random variables are independent and identically distributed (i.i.d.). Furthermore, let's assume that each X_i has a probability density function $p_{X_i}(x; \theta)$. Given a realization of our random sample, $x_1, x_2, \ldots, x_n$, (remember, these are the actual *numbers* that we have observed), we define the *likelihood function* $\mathcal{L}(\theta)$ as follows:

$$\begin{aligned} \mathcal{L}(\theta) &= p_{X_1,\ldots,X_n}(x_1, x_2, \ldots, x_n; \theta), \\ &= \prod_{i=1}^{n} p_{X_i}(x_i; \theta), && \text{using independence of the } X_i. \end{aligned}$$

Here, $p_{X_1,\ldots,X_n}$ is the joint pdf for all of the X_i variables. This pdf depends on the value of the parameter θ for the distribution, so that is in the notation after the semicolon. Notice an important point, we are treating the x_i as constants (they are the data that we've observed) and $\mathcal{L}$ is a function of θ. Maximum likelihood now says that we want to maximize this likelihood function as a function of θ.

MLE of Gaussian mean parameter, μ

Now, let's work this out for the Gaussian case, i.e., let $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$. We will focus only on the MLE of the μ parameter, essentially treating σ^2 as a known constant for simplicity of the example. The likelihood function looks like this:

$$\begin{aligned} \mathcal{L}(\mu) &= p_{X_1,\ldots,X_n}(x_1, x_2, \ldots, x_n; \mu), \\ &= \prod_{i=1}^{n} p_{X_i}(x_i; \theta), && \text{using independence of the } X_i, \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right), && \text{using Gaussian pdf for each } X_i, \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right), && \text{product turns into a sum inside exp.} \end{aligned}$$

To maximize this function, it is easier to think about maximizing it's natural log. We can do this because $\ln$ is a monotonically increasing function, so the value of μ that maximizes $\mathcal{L}$ also maximizes $\ln \mathcal{L}$. So, the *log likelihood function* is defined as

$$\ell(\mu) = \ln \mathcal{L}(\mu) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 + C,$$

where C is a constant in μ (we don't need it to maximize ℓ). Now, defining our estimate of μ to maximize the log likelihood, we get

$$\hat{\mu} = \arg\max_{\mu}\ \ell(\mu) = \arg\min_{\mu} \sum_{i=1}^{n}(x_i - \mu)^2.$$

Notice we changed the sign in the last equality, and this changes us from a max to a min problem. This is called *least squares*, as we are minimizing the sum-of-squared differences from the μ to our data x_i. We can solve this maximization problem exactly using the fact (from calculus) that the derivative of ℓ with respect to μ will be zero at a maxima. We get

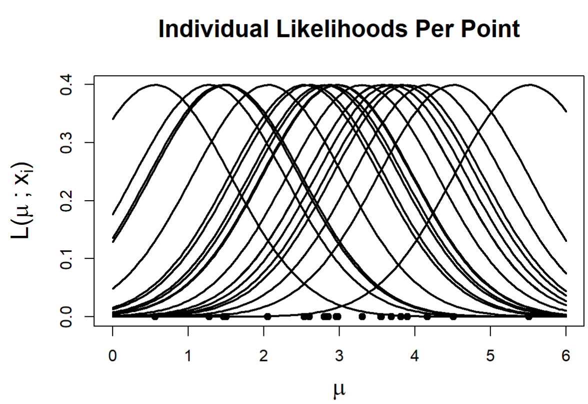
$$0 = \frac{d}{d\mu}\ell(\mu) = \frac{d}{d\mu}\sum_{i=1}^{n}(x_i - \mu)^2 = 2n\mu - 2\sum_{i=1}^{n} x_i.$$

Solving for μ, we get the sample mean as the MLE:

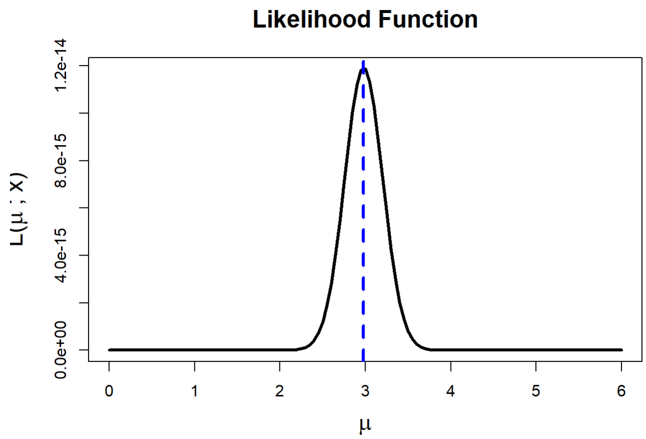
$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Here are some plots demonstrating the above MLE of the mean of a Gaussian. First, we generated a random sample, $x_1, \ldots, x_{20}$ from a normal distribution with $\mu = 3, \sigma = 1$.

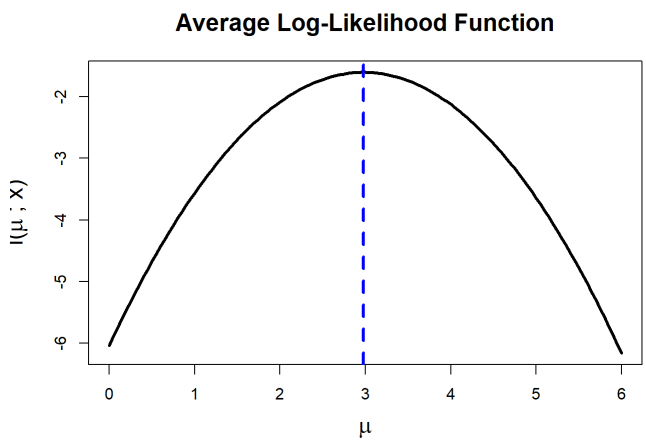
Next, we plot the likelihood functions, $p(x_i; \mu)$, for each of the points separately. Note that the x_i points are plotted on the bottom (x-axis) and each one has its own Gaussian pdf "hill" centered above it. These are the $p(x_i; \mu)$.



Next, we plot the likelihood function for all of the data, which is just the product of all of the $p(x_i; \mu)$. The vertical line is at the average of the x_i data. You can see that the maximum of the likelihood curve is indeed at the average.



Finally, we plot the log-likelihood function (the log of the previous plot, which is just a quadratic). The maximum is still at the same place.



MLE of a Bernoulli probability

The Bernoulli distribution is the binary variable distribution. If now our random variables X_i are binary variables, the notation is $X_i \sim \mathrm{Ber}(\theta)$. The parameter θ gives the probability that X_i is a one. In other words:

$$\begin{aligned} P(X_i = 1) &= \theta, \\ P(X_i = 0) &= 1 - \theta \end{aligned}$$

Now, what is the MLE for θ? The likelihood for a single x_i is:

$$p(x_i; \theta) = \theta^{x_i}(1-\theta)^{1-x_i}$$

Notice this is θ when $x_i = 1$ and $1 - \theta$ when $x_i = 0$. Now the joint likelihood of all x_i is just the product of these individual likelihoods:

$$\begin{aligned} L(\theta) &= p(x_1, \ldots, x_n; \theta) \\ &= p(x_1; \theta) \times p(x_2; \theta) \times \cdots \times p(x_n; \theta) \\ &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum_i x_i}(1-\theta)^{\sum_i (1-x_i)} \\ &= \theta^k (1-\theta)^{n-k}, \quad \text{where } k = \sum_{i=1}^{n} x_i \end{aligned}$$

To maximize $L(\theta)$, we can take the derivative (without first taking log this time):

$$\begin{aligned} \frac{dL}{d\theta} &= k\theta^{k-1}(1-\theta)^{n-k} - (n-k)\theta^k(1-\theta)^{n-k-1} \\ &= (k(1-\theta) - (n-k)\theta)\,\theta^{k-1}(1-\theta)^{n-k-1} \\ &= (k - n\theta)\,\theta^{k-1}(1-\theta)^{n-k-1} \end{aligned}$$

Setting this to zero ($dL/d\theta = 0$), and then solving for θ, gives the maximum likelihood estimate:

$$\hat{\theta} = \frac{k}{n}.$$

This is what we intuitively expect. The value k is the number of ones appearing in our data, so $\hat{\theta}$ is the proportion of ones in our data.