Maximum Likelihood Estimation

Our first algorithm for estimating parameters is called Maximum Likelihood Estimation (MLE). The central idea behind MLE is to select that parameters (θ) that make the observed data the most likely.

The data that we are going to use to estimate the parameters are going to be n independent and identically distributed (IID) samples: $X_1, X_2, \ldots X_n$.

Likelihood

We made the assumption that our data are identically distributed. This means that they must have either the same probability mass function (if the data are discrete) or the same probability density function (if the data are continuous). To simplify our conversation about parameter estimation we are going to use the notation $f(X|\theta)$ to refer to this shared PMF or PDF. Our new notation is interesting in two ways. First, we have now included a conditional on θ which is our way of indicating that the likelihood of different values of X depends on the values of our parameters. Second, we are going to use the same symbol f for both discrete and continuous distributions.

What does likelihood mean and how is "likelihood" different than "probability"? In the case of discrete distributions, likelihood is a synonym for the probability mass, or joint probability mass, of your data. In the case of continuous distribution, likelihood refers to the probability density of your data.

Since we assumed that each data point is independent, the likelihood of all of our data is the product of the likelihood of each data point. Mathematically, the likelihood of our data give parameters θ is:

$$L(\theta) = \prod_{i=1}^{n} f(X_i|\theta)$$

For different values of parameters, the likelihood of our data will be different. If we have correct parameters our data will be much more probable than if we have incorrect parameters. For that reason we write likelihood as a function of our parameters (θ).

Maximization

In maximum likelihood estimation (MLE) our goal is to chose values of our parameters (θ) that maximizes the likelihood function from the previous section. We are going to use the notation $\hat{\theta}$ to represent the best choice of values for our parameters. Formally, MLE assumes that:

$$\hat{\theta} = \underset{\theta}{\text{argmax}}\ L(\theta)$$

Argmax is short for Arguments of the Maxima. The argmax of a function is the value of the domain at which the function is maximized. It applies for domains of any dimension.

A cool property of argmax is that since log is a monotone function, the argmax of a function is the same as the argmax of the log of the function! That's nice because logs make the math simpler. If we find the argmax of the log of likelihood it will be equal to the armax of the likelihood. Thus for MLE we first write the Log Likelihood function (LL)

$$LL(\theta) = \log L(\theta) = \log \prod_{i=1}^{n} f(X_i|\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

To use a maximum likelihood estimator, first write the log likelihood of the data given your parameters. Then chose the value of parameters that maximize the log likelihood function. Argmax can be computed in many ways. All of the methods that we cover in this class require computing the first derivative of the function.

Bernoulli MLE Estimation

For our first example, we are going to use MLE to estimate the p parameter of a Bernoulli distribution. We are going to make our estimate based on n data points which we will refer to as IID random variables $X_1, X_2, \dots X_n$. Every one of these random variables is assumed to be a sample from the same Bernoulli, with the same p, $X_i \sim \text{Ber}(p)$. We want to find out what that p is.

Step one of MLE is to write the likelihood of a Bernoulli as a function that we can maximize. Since a Bernoulli is a discrete distribution, the likelihood is the probability mass function.

The probability mass function of a Bernoulli X can be written as $f(x) = p^x(1-p)^{1-x}$. Wow! What's up with that? It's an equation that allows us to say that the probability that $X = 1$ is p and the probability that $X = 0$ is $1 - p$. Convince yourself that when $X_i = 0$ and $X_i = 1$ the PMF returns the right probabilities. We write the PMF this way because its derivable.

Now let's do some MLE estimation:

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} && \text{First write the likelihood function} \\
LL(\theta) &= \sum_{i=1}^{n} \log p^{x_i}(1-p)^{1-x_i} && \text{Then write the log likelihood function} \\
&= \sum_{i=1}^{n} x_i(\log p) + (1-x_i) log(1-p) \\
&= Y \log p + (n - Y) \log(1-p) && \text{where } Y = \sum_{i=1}^{n} x_i
\end{aligned}
$$

Great Scott! We have the log likelihood equation. Now we simply need to chose the value of p that maximizes our log-likelihood. As your calculus teacher probably taught you, one way to find the value which maximizes a function that is to find the first derivative of the function and set it equal to 0.

$$
\frac{\delta LL(p)}{\delta p} = Y\frac{1}{p} + (n - Y)\frac{-1}{1-p} = 0
$$

$$
\hat{p} = \frac{Y}{n} = \frac{\sum_{i=1}^{n} x_i}{n}
$$

All that work and find out that the MLE estimate is simply the sample mean...

Normal MLE Estimation

Practice is key. Next up we are going to try and estimate the best parameter values for a normal distribution. All we have access to are n samples from our normal which we refer to as IID random variables $X_1, X_2, \dots X_n$. We assume that for all i, $X_i \sim N(\mu = \theta_0, \sigma^2 = \theta_1)$. This example seems trickier since a normal has **two** parameters that we have to estimate. In this case θ is a vector with two values, the first is the mean (μ) parameter. The second is the variance(σ^2) parameter.

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} f(X_i|\theta) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta_1}} e^{-\frac{(x_i-\theta_0)^2}{2\theta_1}} && \text{Likelihood for a continuous variable is the PDF} \\
LL(\theta) &= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\theta_1}} e^{-\frac{(x_i-\theta_0)^2}{2\theta_1}} && \text{We want to calculate log likelihood} \\
&= \sum_{i=1}^{n} \left[-\log(\sqrt{2\pi\theta_1}) - \frac{1}{2\theta_1}(x_i - \theta_0)^2 \right]
\end{aligned}
$$

Again, the last step of MLE is to chose values of θ that maximize the log likelihood function. In this case we can calculate the partial derivative of the LL function with respect to both θ_0 and θ_1, set both equations to equal 0 and than solve for the values of θ. Doing so results in the equations for the values $\hat{\mu} = \hat{\theta_0}$ and $\hat{\sigma^2} = \hat{\theta_1}$ that maximize likelihood. The result is: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$.

Linear Transform Plus Noise

MLE is an algorithm that can be used for any probability model with a derivable likelihood function. As an example lets estimate the parameter θ in a model where there is a random variable Y such that $Y = \theta X + Z$, $Z \sim N(0, \sigma^2)$ and X is an unknown distribution.

In the case where you are told the value of X, θX is a number and $\theta X + Z$ is the sum of a gaussian and a number. This implies that $Y|X \sim N(\theta X, \sigma^2)$. Our goal is to chose a value of θ that maximizes the probability IID: $(X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n)$.

We approach this problem by first finding a function for the log likelihood of the data given θ. Then we find the value of θ that maximizes the log likelihood function. To start, use the PDF of a Normal to express the probability of $Y|X, \theta$:

$$f(Y_i|X_i, \theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(Y_i - \theta X_i)^2}{2\sigma^2}}$$

Now we are ready to write the likelihood function, then take its log to get the log likelihood function:

$$\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} f(Y_i, X_i|\theta) && \text{Let's break up this joint} \\
&= \prod_{i=1}^{n} f(Y_i|X_i, \theta) f(X_i) && f(X_i) \text{ is independent of } \theta \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(Y_i - \theta X_i)^2}{2\sigma^2}} f(X_i) && \text{Substitute in the definition of } f(Y_i|X_i)
\end{aligned}$$

$$\begin{aligned}
LL(\theta) &= \log L(\theta) \\
&= \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(Y_i - \theta X_i)^2}{2\sigma^2}} f(X_i) && \text{Substitute in } L(\theta) \\
&= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(Y_i - \theta X_i)^2}{2\sigma^2}} + \sum_{i=1}^{n} \log f(X_i) && \text{Log of a product is the sum of logs} \\
&= n \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \theta X_i)^2 + \sum_{i=1}^{n} \log f(X_i)
\end{aligned}$$

Remove constant multipliers and terms that don't include θ. We are left with trying to find a value of θ that maximizes:

$$\begin{aligned}
\hat{\theta} &= \underset{\theta}{\text{argmax}} - \sum_{i=1}^{m} (Y_i - \theta X_i)^2 \\
&= \underset{\theta}{\text{argmin}} \sum_{i=1}^{m} (Y_i - \theta X_i)^2
\end{aligned}$$

This result says that the value of θ that makes the data most likely is one that minimizes the squared error of predictions of Y. We will see in a few days that this is the basis for linear regression.