Chris Piech
CS109

Lecture #20
Nov 9th, 2018

# Maximum Likelihood Estimation

We have learned many different distributions for random variables and all of those distributions had parameters: the numbers that you provide as input when you define a random variable. So far when we were working with random variables, we either were explicitly told the values of the parameters, or, we could divine the values by understanding the process that was generating the random variables.

What if we don't know the values of the parameters and we can't estimate them from our own expert knowledge? What if instead of knowing the random variables, we have a lot of examples of data generated with the same underlying distribution? In this chapter we are going to learn formal ways of estimating parameters from data.

These ideas are critical for artificial intelligence. Almost all modern machine learning algorithms work like this: (1) specify a probabilistic model that has parameters. (2) Learn the value of those parameters from data.

## Parameters

Before we dive into parameter estimation, first let's revisit the concept of parameters. Given a model, the parameters are the numbers that yield the actual distribution. In the case of a Bernoulli random variable, the single parameter was the value $p$. In the case of a Uniform random variable, the parameters are the $a$ and $b$ values that define the min and max value. Here is a list of random variables and the corresponding parameters. From now on, we are going to use the notation $\theta$ to be a vector of all the parameters: In the real

| Distribution | Parameters |
|---|---|
| Bernoulli(p) | $\theta = p$ |
| Poisson($\lambda$) | $\theta = \lambda$ |
| Uniform(a,b) | $\theta = (a, b)$ |
| Normal($\mu, \sigma^2$) | $\theta = (\mu, \sigma^2)$ |
| Y = mX + b | $\theta = (m, b)$ |

world often you don't know the "true" parameters, but you get to observe data. Next up, we will explore how we can use data to estimate the model parameters.

It turns out there isn't just one way to estimate the value of parameters. There are two main schools of thought: Maximum Likelihood Estimation (MLE) and Maximum A Posteriori (MAP). Both of these schools of thought assume that your data are independent and identically distributed (IID) samples: $X_1, X_2, \dots X_n$ where $X_i$.

## Maximum Likelihood

Our first algorithm for estimating parameters is called Maximum Likelihood Estimation (MLE). The central idea behind MLE is to select that parameters ($\theta$) 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: $X_1, X_2, \dots X_n$.

## 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 $\theta$ 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 joint probability of your data. In the case of continuous distribution, likelihood refers to the joint 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 $\theta$ 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 ($\theta$).

## Maximization

In maximum likelihood estimation (MLE) our goal is to chose values of our parameters ($\theta$) 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:

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

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 $X_1, X_2, \ldots X_n$. 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! Whats up with that? Its 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.

$$
\begin{aligned}
\frac{\delta LL(p)}{\delta p} &= Y\frac{1}{p} + (n-Y)\frac{-1}{1-p} = 0 \\
\hat{p} &= \frac{Y}{n} = \frac{\sum_{i=1}^{n} X_i}{n}
\end{aligned}
$$

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 $X_1, X_2, \ldots X_n$. 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 $\theta$ is a vector with two values, the first is the mean ($\mu$) parameter. The second is the variance($\sigma^2$) 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 $\theta$ that maximize the log likelihood function. In this case we can calculate the partial derivative of the $LL$ function with respect to both $\theta_0$ and $\theta_1$, set both equations to equal 0 and than solve for the values of $\theta$. 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 $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$.