Chris Piech
CS109

Lecture #21
Nov 10th, 2017

Maximum Likelihood

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 θ to be a vector of all the parameters: In the real 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.

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

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: $X_1, X_2, \ldots X_n$ 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 (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

First we define the likelihood of our data give parameters θ:

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

This is the probability of all of our data. It evaluates to a product because all X_i are independent. Now we chose the value of θ that maximizes the likelihood function. Formally $\hat{\theta} = \underset{\theta}{\text{argmax}}\ L(\theta)$.

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. Instead of using likelihood, you should instead use log likelihood: $LL(\theta)$.

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

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. Most require computing the first derivative of the function.

Bernoulli MLE Estimation

Consider IID random variables $X_1, X_2, \dots X_n$ where $X_i \sim \text{Ber}(p)$. First we are going to write the PMF of a Bernoulli in a crazy way: The probability mass function $f(X_i|p) = p^{X_i}(1-p)^{1-X_i}$. Wow! Whats up with that? First convince yourself that when $X_i = 0$ and $X_i = 1$ this returns the right probabilities. We write the PMF this way because its derivable.

Now let's do some MLE estimation:

$$\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} \\
LL(\theta) &= \sum_{i=1}^{n} \log p^{X_i}(1-p)^{1-X_i} \\
&= \sum_{i=1}^{n} X_i(\log p) + (1-X_i)log(1-p) \\
&= Y \log p + (n-Y)log(1-p) && \text{where } Y = \sum_{i=1}^{n} X_i
\end{aligned}$$

Great Scott! Now we simply need to chose the value of p that maximizes our log-likelihood. One way to do that is to find the first derivative 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 we get the same thing as method of moments and sample mean...

Normal MLE Estimation

Consider IID random variables $X_1, X_2, \dots X_n$ where $X_i \sim N(\mu, \sigma^2)$.

$$\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} f(X_i|\mu, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}} \\
LL(\theta) &= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}} \\
&= \sum_{i=1}^{n} \left[-\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}(X_i-\mu)^2 \right]
\end{aligned}$$

If we chose the values of $\hat{\mu}$ and $\hat{\sigma^2}$ that maximize likelihood, we get:$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$.