11. Parameter Estimation

Chris Piech and Mehran Sahami

May 2017

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 θ to be a vector of all the 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	$\theta = (m, b)$

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.

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

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 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 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 $X_1, X_2, \ldots X_n$. 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 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:

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

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

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.

Maximum A Posterior Estimation

MLE is great, but it is not the only way to estimate parameters! This section introduces an alternate algorithm, Maximum A Posteriori (MAP).The paradigm of MAP is that we should chose the value for our parameters that is the most likely given the data. At first blush this might seem the same as MLE, however notice that MLE chooses the value of parameters that makes the *data* most likely. Formally, for IID random variables $X_1, \ldots, X_n$:

$$\theta_{\text{MAP}} = \underset{\theta}{\text{argmax}}\ f(\theta | X_1, X_2, \ldots X_n)$$

In the equation above we trying to calculate the conditional probability of unobserved random variables given observed random variables. When that is the case, think Bayes Theorem! Expand the function f using the continuous version of Bayes Theorem.

$$\begin{aligned}
\theta_{\text{MAP}} &= \underset{\theta}{\text{argmax}}\ f(\theta | X_1, X_2, \ldots X_n) && \text{Now apply Bayes Theorem} \\
&= \underset{\theta}{\text{argmax}}\ \frac{f(X_1, X_2, \ldots, X_n | \theta) g(\theta)}{h(X_1, X_2, \ldots X_n)} && \text{Ahh much better}
\end{aligned}$$

Note that f, g and h are all probability densities. I used different symbols to make it explicit that they may have different functions. Now we are going to leverage two observations. First, the data is assumed to be IID so we can decompose the density of the data given θ. Second, the denominator is a constant with respect to θ. As such its value does not affect the argmax and we can drop that term. Mathematically:

$$\begin{aligned}
\theta_{\text{MAP}} &= \underset{\theta}{\text{argmax}}\ \frac{\prod_{i=1}^{n} f(X_i | \theta) g(\theta)}{h(X_1, X_2, \ldots X_n)} && \text{Since the samples are IID} \\
&= \underset{\theta}{\text{argmax}}\ \prod_{i=1}^{n} f(X_i | \theta) g(\theta) && \text{Since } h \text{ is constant with respect to } \theta
\end{aligned}$$

As before, it will be more convenient to find the argmax of the log of the MAP function, which gives us the final form for MAP estimation of parameters.

$$\theta_{\text{MAP}} = \underset{\theta}{\text{argmax}} \left(\log(g(\theta)) + \sum_{i=1}^{n} \log(f(X_i | \theta)) \right)$$

Using Bayesian terminology, the MAP estimate is the mode of the "posterior" distribution for θ. If you look at this equation side by side with the MLE equation you will notice that MAP is the argmax of the exact same function *plus* a term for the log of the prior.

Parameter Priors

In order to get ready for the world of MAP estimation, we are going to need to brush up on our distributions. We will need reasonable distributions for each of our different parameters. For example, if you are predicting a Poisson distribution, what is the right random variable type for the prior of λ?

A desiderata for prior distributions is that the resulting posterior distribution has the same functional form. We call these "conjugate" priors. In the case where you are updating your belief many times, conjugate priors makes programming in the math equations much easier.

Here is a list of different parameters and the distributions most often used for their priors:

Parameter	Distribution
Bernoulli p	Beta
Binomial p	Beta
Poisson λ	Gamma
Exponential λ	Gamma
Multinomial p_i	Dirichlet
Normal μ	Normal
Normal σ^2	Inverse Gamma

You are only expected to know the new distributions on a high level. You do not need to know Inverse Gamma. I included it for completeness.

The distributions used to represent your "prior" belief about a random variable will often have their own parameters. For example, a Beta distribution is defined using two parameters (a,b). Do we have to use parameter estimation to evaluate a and b too? No. Those parameters are called "hyperparameters". That is a term we reserve for parameters in our model that we fix before running parameter estimate. Before you run MAP you decide on the values of (a,b).

Dirichlet

The Dirichlet distribution generalizes Beta in same way Multinomial generalizes Bernoulli. A random variable X that is Dirichlet is parametrized as $X \sim \text{Dirichlet}(a_1, a_2, \ldots, a_m)$. The PDF of the distribution is:

$$f(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m) = K \prod_{i=1}^{m} x_i^{a_i - 1}$$

Where K is a normalizing constant.

You can intuitively understand the hyperparameters of a Dirichlet distribution: imagine you have seen $\sum_{i=1}^{m} a_i - m$ imaginary trials. In those trials you had $(a_i - 1)$ outcomes of value i. As an example consider estimating the probability of getting different numbers on a six-sided Skewed Dice (where each side is a different shape). We will estimate the probabilities of rolling each side of this dice by repeatedly rolling the dice n times. This will produce n IID samples. For the MAP paradigm, we are going to need a prior on our belief of each of the parameters $p_1 \ldots p_6$. We want to express that we lightly believe that each roll is equally likely.

Before you roll, let's imagine you had rolled the dice six times and had gotten one of each possible values. Thus the "prior" distribution would be Dirichlet$(2,2,2,2,2,2)$. After observing $n_1 + n_2 + \cdots + n_6$ new trials with n_i results of outcome i, the "posterior" distribution is Dirichlet$(2+n_1, \ldots 2+n_6)$. Using a prior which represents one imagined observation of each outcome is called "Laplace smoothing" and it guarantees that none of your probabilities are 0 or 1.

Gamma

The Gamma(k,θ) distribution is the conjugate prior for the λ parameter of the Poisson distribution (It is also the conjugate for Exponential, but we won't delve into that).

The hyperparameters can be interpreted as: you saw k total imaginary events during θ imaginary time periods. After observing n events during the next t time periods the posterior distribution is Gamma$(k+n, \theta+t)$.

For example Gamma(10, 5) would represent having seen 10 imaginary events in 5 time periods. It is like imagining a rate of 2 with some degree of confidence. If we start with that Gamma as a prior and then see 11 events in the next 2 time periods our posterior is Gamma(21,7) which is equivalent to an updated rate of 3.