Will Monroe
CS 109

Lecture Notes #20
August 9, 2017

Parameter Learning

Based on a chapter by Chris Piech

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 approaches: Maximum Likelihood Estimation (MLE) and Maximum A Posteriori (MAP). Both of these approaches assume that your data are IID samples: $X_1, X_2, \ldots X_n$ where all X_i are independent and have the same distribution.

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 $f(X \mid \theta)$ 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 each data point is independent, the likelihood of all our data is the product of the likelihood of each data point. Mathematically, the likelihood of our data given 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:

$$\hat{\theta} = \arg\max_{\theta} L(\theta)$$

"Arg max" is short for *argument of the maximum*. The arg max 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 arg max is that since log is a monotone function, the arg max of a function is the same as the arg max of the log of the function! That's nice because logs make the math simpler.

If we find the arg max of the log of likelihood, it will be equal to the arg max of the likelihood. Therefore, 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.

You may not have realized before that the probability mass function of a Bernoulli X can be written as $f(X) = p^X(1-p)^{1-X}$. Interesting! Where did that come from? 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 it is differentiable.

Let's do some maximum likelihood estimation:

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

We have a formula for the log likelihood. 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 to find out that the maximum likelihood estimate is simply the sample mean...

Normal MLE Estimation

Practice is key. Next, we will estimate the best parameter values for a normal distribution. All we have access to are n samples from our normal, which we represent 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 because 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, and 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 choose 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 then 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 $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$.

Maximum A Posteriori 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, remember that MLE chooses the value of parameters that makes the *data* most likely. Formally, for IID random variables $X_1, \ldots, X_n$:

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

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:

$$
\begin{aligned}
\theta_{\text{MAP}} &= \arg\max_{\theta} f(\theta|X_1, X_2, \ldots X_n) \\
&= \arg\max_{\theta} \frac{f(X_1, X_2, \ldots, X_n|\theta)g(\theta)}{h(X_1, X_2, \ldots X_n)} && \text{by Bayes' Theorem}
\end{aligned}
$$

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 arg max, and we can drop that term. Mathematically:

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

As before, it will be more convenient to find the arg max 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 arg max 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

We won't cover the inverse gamma distribution in this class. The remaining two, Dirichlet and gamma, you will not be required to know, but details for them are included below 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 $X \sim \text{Dir}(a_1, a_2, \ldots, a_m)$. 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 die" (where each side is a different shape). We will estimate the probabilities of rolling each side of this die by repeatedly rolling the die 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 diee six times and had gotten one of each possible value. Thus, the "prior" distribution would be $\text{Dir}(2, 2, 2, 2, 2, 2)$. After observing $n_1 + n_2 + \cdots + n_6$ new trials with n_i results of outcome i, the "posterior" distribution is $\text{Dir}(2 + n_1, \ldots 2 + n_6)$. 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 $\text{Gamma}(k, \theta)$ distribution is the conjugate prior for the λ parameter of the Poisson distribution. (It is also the conjugate for the λ in the exponential, but we won't cover that here.)

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 $\text{Gamma}(k + n, \theta + t)$.

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.