Chris Piech
CS109

Lecture #23
Nov 14th, 2018

Maximum A Posteriori

Today we are going to cover our third Parameter Estimator, Maximum A Posteriori (MAP). The other two were Unbiased estimation and Maximum Likelihood (MLE). 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 choses the value of parameters that makes the *data* most likely. Formally, we observe data: $x^{(1)},\dots,x^{(n)}$ which we think of as assignments to IID random variables $X^{(1)},\dots,X^{(n)}$. We also think of a random variable for all of our parameters Θ:

$$\theta_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}}\ f(\Theta=\theta|X^{(1)}=x^{(1)},X^{(2)}=x^{(2)},\dots,X^{(n)}=x^{(n)})$$

That is pretty complex notation. And it seems quite redundant (each datum has a corresponding random variable). To make it easier to read I am going to use the datum value (eg $x^{(2)}$) as short hand for the **event** that the corresponding random variable takes on said value: (eg $X^{(2)}=x^{(2)}$). Using my new notation:

$$\theta_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}}\ f(\theta|x^{(1)},x^{(2)},\dots,x^{(n)})$$

In the equation above we are trying to calculate the conditional probability of an unobserved parameter given observed data, and we only know the probability the other way around. Think Bayes Theorem! Lets expand the function f using the continuous version of Bayes Theorem:

$$\begin{aligned}\theta_{\text{MAP}} &= \underset{\theta}{\operatorname{argmax}}\ f(\theta|x^{(1)},x^{(2)},\dots,x^{(n)}) && \text{Now apply Bayes Theorem}\\ &= \underset{\theta}{\operatorname{argmax}}\ \frac{g(\theta)f(x^{(1)},x^{(2)},\dots,x^{(n)}|\theta)}{h(x^{(1)},x^{(2)},\dots x^{(n)})} && \text{Ahh much better}\end{aligned}$$

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:

$$\begin{aligned}\theta_{\text{MAP}} &= \underset{\theta}{\operatorname{argmax}}\ \frac{g(\theta)\prod_{i=1}^{n} f(x^{(i)}|\theta)}{h(x^{(1)},x^{(2)},\dots,x^{(n)})} && \text{Since the samples are IID}\\ &= \underset{\theta}{\operatorname{argmax}}\ g(\theta)\prod_{i=1}^{n} f(x^{(i)}|\theta) && \text{Since } h \text{ is a positive constant with respect to } \theta\end{aligned}$$

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.

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

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? Here is a list of different parameters and the distribution used as their priors:

Parameter	**Prior 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 don't cover the Inverse Gamma in CS109. 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 estimation. Before you run MAP you decide on the values of (a, b).

Dirichlet

The Dirichlet distribution generalizes Beta in same way Multinomial generalizes Bernoulli. If estimating a Multinomial using the MAP paradigm, we are going to need a prior on our belief of each of the multinomial parameters: welcome the Dirichlet. A random variable $\mathbf{X}$ that is Dirichlet is parametrized as $\mathbf{X} \sim \text{Dirichlet}(a_1, a_2, \ldots, a_m)$. The PDF of the distribution is:

$$f(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m) = K \prod_{i=1}^{m} x_i^{a_i - 1} \qquad \text{Where } K \text{ is a normalizing constant.}$$

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 has a different probability). We will estimate the probabilities of rolling each side of this dice by rolling the dice n times. This will produce n IID samples. 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 $X \sim \text{Gamma}(k, \theta)$ distribution is the conjugate prior for the λ parameter of the Poisson distribution (It is also the conjugate for Exponential). The parameters of Gamma 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.