Lisa Yan
CS109

Lecture Notes #22
May 27, 2020

Maximum A Posteriori

Based on a chapter by Chris Piech

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 choose 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.

One of the disadvantages of MLE is that it best explains data we have seen and makes no attempt to generalize to unseen data. In MAP, we incorporate *prior* belief about our parameters, and then we update our posterior belief of the parameters based on the data we have seen.

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.

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

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).

Beta

We've covered that Beta is a conjugate distribution for Bernoulli. The MAP of a Bernoulli distribution with a Beta prior is the *mode* of the Beta posterior. The *mode* of a distribution is the value that maximizes the probability mass function (if discrete) or probability density function (if continuous).

If $X \sim \text{Beta}(a, b)$, where a, b are integers where $a + b > 2$, the mode is $\underset{x}{\text{argmax}} f(x) = \frac{a-1}{a+b-2}$, where $f(x)$ is the PDF of X.

Example 1

Flip $n + m$ coins and observe n heads. If we assume a prior on p of $\text{Beta}(n_{imag} + 1, m_{imag} + 1)$, the posterior on the parameter p is $\text{Beta}(n + n_{imag} + 1, m + m_{imag} + 1)$. The MAP estimator is therefore the mode of this distribution: $\frac{n+n_{imag}}{n+n_{imag}+m+m_{imag}}$.

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 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 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. The Laplace estimate for a Multinomial RV is $p_i = \frac{X_i + 1}{n + m}$ for $i = 1, \ldots, m$, where n is the number of actual trials in your observed experiment.

Gamma

The Gamma(α, β) 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 mode of the Gamma distribution is $\frac{\alpha - 1}{\beta}$.

The hyperparameters can be interpreted as: you saw $\alpha - 1$ total imaginary events during β imaginary time periods. After observing n events during the next k time periods the posterior distribution is Gamma$(\alpha + n, \beta + k)$.

For example, Gamma(11, 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(22, 7), which is equivalent to an updated rate of 3.