MLE and MAP: Statistical Justifications

Somani, Neel - Bao, Jason

February 20, 2018

1 Introduction

We started today's lecture with an example of a situation where you cannot calculate the exact value of a statistic, because you don't have access to all of the data. Specifically, we used the example of a website that wants to estimate the average time that a user stays on the website.

We reviewed the concept of a *confidence interval*, that is, a range that we think the true value of a statistic falls in. We used the concept of *bootstrapping* to demonstrate this: we resampled from our training data (the amount of time that each user was on the website) with replacement some number of times, and calculated our statistic (the mean) each time. This range of values allowed us to determine a confidence interval.

Our best estimate for the true mean of the population was the mean of our training data, because this population distribution would have maximized the probability that we observed this training data.

As we'll see, similar reasoning applies to machine learning models. After hypothesizing what the true model looks like, we use our sample data to determine what the parameters of this model should be to maximize the probability that we observed this data.

2 Probability Review

Linearity of expectation: $\mathbf{E}[aX + Y] = a\mathbf{E}[X] + \mathbf{E}[Y]$

If X, Y are independent*: $Var(X + Y) = Var(X) + Var(Y)$ and in general for a constant a: $Var(aX) = a^2Var(X)$

* Try deriving this from $Var(X) = E[(X - E[X])^2]$, and from
$X, Y independent \Rightarrow E[XY] = E[X] * E[Y]$

2.1 Chebyshev's Inequality

Way to bound probability: $Pr(|x-\mu| \geq k\sigma) \leq \frac{1}{k^2}$ where σ = standard deviation, k is an arbitrary number and μ = mean.

The reason we care about this is that it leads to the weak law of large numbers since the standard deviation goes to 0 as the size of the sample approaches infinity (why is this the case with the equations given above? hint: if we set $x = (\sum_{i=0}^{n} x_n)/n$ what do we set σ to?). In other words, the mean of the sample converges in probability to μ.

We covered the exponential distribution, which has the unique property of being a *memoryless* continuous probability distribution.

2.2 The Envelope Problem: An Application of Bayes' Rule

We reviewed the puzzle from last lecture:

There are two envelopes, one with M dollars, and one with N dollars. Without loss of generality, $M < N$.

You're given one envelope, and you see the amount. You don't know the amount of the other envelope, but you have the option to switch envelopes if you want to. How can you guarantee greater than 50% odds of picking the larger amount?

One answer: Flip a coin until you see one heads, and record the number of flips until you see heads. If the number is greater than or equal to the amount that you see, then switch. Otherwise, don't switch. Why does this strategy work?

First, let's consider the probability that you win if the number of heads does not exceed M. In this case, you don't switch, so your chance of winning is the same as the chance that you picked the correct envelope: 50%.

Next, let's consider the probability of winning if the number of heads falls between M and N. If you picked M to begin with, then you'll switch. If you picked N, then you'll stay. So the probability of winning is 1.

Finally, we'll consider the probability of winning if the number of heads exceeds N. You'll switch if you picked N, and you'll switch if you picked M, so the probability of winning is .5.

So our total probability of winning is:
$.5 * Pr[H < M] + 1 * Pr[M <= H < N] + .5 * Pr[N <= H]$
$= .5 * (Pr[H < M] + Pr[N <= H]) + 1 * Pr[M <= H < N] > .5$

3 Maximum Likelihood Estimation

Suppose we have $X = x_1, ..., x_n$. We're trying to estimate what the true model looks like, so we make a hypothesis. For example, we might guess that the true distribution is normally distributed. (Note that this is an assumption that we're making with MLE. Bootstrapping didn't make any kind of assumption like this.)

$$pdf(X) = \frac{1}{\sqrt{2\pi\sigma^2}} * e^{\frac{-(X-\mu)^2}{2\sigma^2}} \tag{1}$$

So we think that the truth is a normal distribution, but how should we pick σ and μ? A good idea might to be to maximize the probability that we observed this set of data.

$$arg\, max_{\sigma,\mu} Pr[X|\sigma,\mu] = \prod_{i=1}^{n} Pr[x_i|\sigma,\mu] \tag{2}$$

which we can substitute with the *pdf* of X:

$$arg\, max_{\sigma,\mu} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} * e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \tag{3}$$

And that's the idea behind MLE.

3.1 Simplifying the Expression

Given that $x_1, x_2, ...x_i$ are i.i.d (identically and independently distributed) we can express the equation as

$$arg\, max_{\sigma,\mu} \prod_{i=1}^{n} Pr[x_i|\sigma,\mu] \tag{4}$$

Now taking the derivative of this thing would be messy, as we have a bunch of exponentials that don't simplify easily. A trick we can use is to *log* the entire expression to get a summation that we can easily take the derivative of (remember that $log(ab) = log(a) + log(b)$) The reason we can do this is because the log function is monotonically increasing, so the *argmax* of the original expression is the same as the *argmax* of the logged expression, revealing:

$$arg\, max_{\sigma,\mu} log(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} * e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}) \tag{5}$$

which we can continue to simplify:

$$= arg\, max_{\sigma,\mu} log(\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} * e^{\frac{-(x_i-\mu)^2}{2\sigma^2}})$$

$$
\begin{aligned}
&= arg\,max_{\sigma,\mu} \sum_{i=1}^{n} log(\frac{1}{\sqrt{2\pi\sigma^2}} * e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}) \\
&= arg\,max_{\sigma,\mu} \sum_{i=1}^{n} log(\frac{1}{\sqrt{2\pi\sigma^2}}) + log(e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}) \\
&= arg\,max_{\sigma,\mu} \sum_{i=1}^{n} -log(\sqrt{2\pi\sigma^2}) - \frac{(x_i-\mu)^2}{2\sigma^2} \\
&= arg\,max_{\sigma,\mu} \sum_{i=1}^{n} -log(\sigma) - \frac{(x_i-\mu)^2}{2\sigma^2} \qquad (6)
\end{aligned}
$$

At this point, we can differentiate with respect to σ or μ and solve for the values that maximize the probability. If you do so, you'll find that the optimal value for μ is our classic sample mean, and the optimal value for σ is our classic sample standard deviation.

So what's the point of MLE? In short, it allows us to statistically model a situation, and solve for the parameters of this model that maximize the probability that we observed the data that we did.

At the very end of lecture, we touched on the result of modeling a situation as having unit normal noise and solving for the optimum: the result is ordinary least squares. If we impose a prior, then we get MAP, which we'll dive more deeply into next lecture.