Notes on MLE

Jamie Morgenstern

January 2023

1 Maximum Likelihood estimation for a bernoulli distribution

Suppose we have a coin whose probability of heads anytime we flip it is some unknown value θ, and that each flip of the coin has this same probability of heads independently of any other flip of the coin. These are assumptions about the coin as a *data-generating process*, namely that it produces heads and tails independently each time it is flipped, and each flip comes up heads with probability θ.

We now ask the question of how would we estimate the value of θ if we had access to this coin? If we flipped the coin n times, and we saw a sequence $D = (THHHHTT...)$, what would we do with this information? If k of the n flips were heads, the *empirical* estimate of the probability of heads is k/n. A natural question is whether this empirical estimate is any good, and whether there are other justifications of whether θ is roughly k/n.

If we knew the value θ, the probability of a sequence D of n coin flips containing exactly k heads is

$$P(D|\theta) = \theta^k(1-\theta)^{n-k} \qquad \text{(Since each coin flip is an independent bernoulli, CSE 312/STAT390.)}$$

This expression is known as the *likelihood* function of our data. The *maximum likelihood estimate* of the parameter θ, denoted $\hat{\theta}_{MLE}$, is defined as the value of θ that maximizes the likelihood of seeing a given observation sequence D. So, rather than thinking of θ as fixed, we think of θ as a variable, and D as fixed, and find the θ which maximizes the above probability of observing D:

$$\begin{aligned}
\hat{\theta}_{MLE} &= \arg\max_\theta P(D|\theta) \\
&= \arg\max_\theta \theta^k(1-\theta)^{n-k} \\
&= \arg\max_\theta \log(\theta^k(1-\theta)^{n-k}) && \text{(Because log is monotone, } x > y \text{ means } \log x > \log y) \\
&= \arg\max_\theta k\log(\theta) + (n-k)\log(1-\theta) && \text{(Using logarithmic identities } \log(xy) = \log x + \log y \text{ and } \log x^k = k\log x)
\end{aligned}$$

Then, how do we find θ which maximizes the above expression? Using our MATH 126 skills, we know that the above expression is maximized in θ only at points where the derivative of the expression with respect to θ is zero. So, we take the derivative of the above expression with respect to θ, set to zero and solve:

$$\begin{aligned}
0 = \frac{\partial}{\partial\theta} k\log(\theta) + (n-k)\log(1-\theta) = \frac{k}{\theta} - \frac{(n-k)}{1-\theta} && \text{Since } \frac{\partial}{\partial\theta}\log(\theta) = \frac{1}{\theta} \\
\Leftrightarrow \frac{n-k}{k} = \frac{1-\theta}{\theta} && \text{Rearranging terms} \\
\Leftrightarrow n/k - 1 = 1/\theta - 1 \Rightarrow k/n = \theta.
\end{aligned}$$

So, the derivative of the log likelihood function is 0 when $\theta = k/n$. Is this a value of θ which maximizes the log likelihood? If we graph this function for fixed values of n and k, we will notice the function increases for awhile and then decreases in θ, and therefore has a unique maximum which is achieved where the derivative is zero.

How "good" is this estimate $\hat{\theta}_{MLE}$? In short, we can describe several properties of an estimator, all of which give us some understanding of how much we should trust an estimate. Let us call the *true* parameter generating our data θ^*. First, we will discuss whether $\mathbb{E}_D[\hat{\theta}_{MLE}] = \theta^*$, i.e. whether the expected value of our estimator is equal to the true value (over the randomness of the data draw). Notice that

$$\mathbb{E}_D[\hat{\theta}_{MLE}(D)] = \mathbb{E}_D[\text{fraction of } D \text{ which are heads}] = \theta^*$$

where the first equality is by the definition of the MLE and the second comes from our assumption that each coin flip comes up heads with probability θ^* independently. Because $\mathbb{E}[\hat{\theta}_{MLE}] = \theta^*$, we call this estimate *unbiased*. The *bias* of an estimate measures how far the average value of the estimator is from the quantity it estimates. This is one way to measure how well an estimator works. We might also care about how *concentrated* an estimator is, or how much its value deviates from its expected value. There are several ways to measure spread of an estimator (or a random variable); one is to measure its *variance*, and another is to calculate the probability it deviates from its mean by more than some ϵ:

$$\mathbb{P}[|\hat{\theta} - \theta^*| \geq \epsilon] \leq \delta.$$

One can generally upper-bound δ by using concentration inequalities like Chernoff-Hoeffding (CSE 312). In the case of our MLE estimator, because our coin flips are independent, we have that

$$\mathbb{P}[|\hat{\theta}_{MLE} - \theta^*| \geq \epsilon] \leq 2e^{-2n\epsilon^2},$$

so the probability our estimate is off by some amount shrinks exponentially in the number of coin flips we observe. This means that the spread of our estimator shrinks as our dataset size grows, so our confidence in our estimator should grow as our dataset size grows.

2 Gaussian RV's ad MLE

Suppose we now wish to understand a distribution of continuous random variables, for example, we wish to understand the distribution of heights across UW's campus. Again, we can formulate a hypothesis (or collection of assumptions) about these heights and follow the same recipe we did for the bernoulli distribution above. A natural assumption would be that the heights of people on campus are distributed according to a gaussian $\mathcal{N}(\mu, \sigma^2)$, with mean μ and variance σ^2, and that we can sample n heights of campus people i.i.d. from this distribution, $D = (x_1, \ldots, x_n)$. Under these assumptions and a set of heights gathered on campus, we can again ask for a maximum likelihood estimate of the mean and variance of heights that generated our dataset. Our *empirical* mean of our heights is simply $\sum_i x_i/n$, which we will now show also turns out to be our MLE under these assumptions. Recall that $P(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, so

$$\begin{aligned}
\hat{\mu}_{MLE} &= \arg\max_{\mu} P(D|\mu, \sigma^2) \\
&= \arg\max_{\mu} \prod_i \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\
&= \arg\max_{\mu} \sum_i \ln(\frac{1}{\sigma\sqrt{2\pi}}) - \frac{(x_i-\mu)^2}{2\sigma^2} \qquad \text{(Because ln is monotone, } x > y \text{ means } \ln x > \ln y \text{, using logarithmic iden} \\
&= \arg\max_{\mu} n\ln(\frac{1}{\sigma\sqrt{2\pi}}) - \sum_i \frac{(x_i-\mu)^2}{2\sigma^2}
\end{aligned}$$

Now, the above expression is maximized when its derivative with respect to μ is zero, so we can take the derivative of the expression and set it to zero and solve:

$$\frac{\partial}{\partial \mu} n \ln(\frac{1}{\sigma\sqrt{2\pi}}) - \sum_i \frac{(x_i-\mu)^2}{2\sigma^2} = -2\sum_i \frac{(x_i-\mu)}{2\sigma^2} \qquad \text{Since derivative of first term is zero}$$

$$0 \Leftrightarrow \sum_i (x_i - \mu) = 0 \Rightarrow (\sum_i x_i)/n = \mu.$$

Again, one can plot the likelihood function and observe it has a unique local and global maximum as a function of μ, so the point where its derivative is zero is that maximum. Therefore, $\hat{\mu}_{MLE} = (\sum_i x_i)/n$. One can also argue that this estimate is unbiased, and could describe the spread of this random variable using a similar Chernoff-Hoeffding argument.

If we want to compute the MLE of the *variance*, $\hat{\sigma^2}_{MLE}$, we can also follow the same formula, and find

$$\begin{aligned}\hat{\sigma}^2_{MLE} &= \arg\max_{\sigma^2} P(D|\hat{\mu}_{MLE}, \sigma^2) \\ &= \arg\max_{\sigma^2} \sum_i \ln(\frac{1}{\sigma\sqrt{2\pi}}) - \frac{(x_i - \hat{\mu}_{MLE})^2}{2\sigma^2}\end{aligned}$$

We then take the derivative with respect to σ and set it to 0 and solve:

$$\begin{aligned}\frac{\partial}{\partial \sigma} \sum_i \ln(\frac{1}{\sigma\sqrt{2\pi}}) - \frac{(x_i - \hat{\mu}_{MLE})^2}{2\sigma^2} &= -n\frac{1}{\sigma} + \sum_i \frac{2(x_i - \hat{\mu}_{MLE})^2}{2\sigma^3} \\ &= 0 \Rightarrow n\sigma^2 = \sum_i (x_i - \hat{\mu}_{MLE})^2 \\ &\Rightarrow \sigma^2 = \frac{\sum_i (x_i - \hat{\mu}_{MLE})^2}{n}\end{aligned}$$

So, once you've verified the likelihood function has a unique local (and global) maximum, this implies that $\hat{\sigma^2}_{MLE} = \frac{\sum_i (x_i - \hat{\mu}_{MLE})^2}{n}$.

This estimator is our first example of one which is biased, since

$$
\begin{aligned}
\mathbb{E}[\hat{\sigma^2}_{MLE}] &= \mathbb{E}[\frac{\sum_i (x_i - \hat{\mu}_{MLE})^2}{n}] \\
&= \frac{1}{n}\sum_i \mathbb{E}[(x_i - \hat{\mu}_{MLE})^2] \\
&= \frac{1}{n}\sum_i \mathbb{E}[(x_i^2)] - 2\mathbb{E}[x_i\hat{\mu}_{MLE})] + \mathbb{E}[\hat{\mu}^2_{MLE}] \\
&= \frac{1}{n}\sum_i \mathbb{E}[(x_i^2)] - \mathbb{E}[\hat{\mu}^2_{MLE}] && \text{Since } \hat{\mu}_{MLE} = \sum_i x_i/n. \\
&= \frac{1}{n}\sum_i \mathbb{E}[(x_i^2)] - \frac{1}{n^2}\mathbb{E}[(\sum_i x_i)^2] && \text{expanding the definition again} \\
&= \frac{1}{n}\sum_i \mathbb{E}[(x_i^2)] - \frac{1}{n^2}\mathbb{E}[\sum_i x_i^2 + 2\sum_{i\neq j} x_i x_j] \\
&= \frac{1}{n}\sum_i \mathbb{E}[(x_i^2)] - \frac{1}{n}\mathbb{E}[x_i^2] - \mathbb{E}[\frac{2}{n^2}\sum_{i\neq j} x_i x_j] \\
&= \frac{n-1}{n}\sum_i \mathbb{E}[(x_i^2)] - \mathbb{E}[\frac{2}{n^2}\sum_{i\neq j} x_i x_j] \\
&= \frac{n-1}{n}\sum_i \mathbb{E}[(x_i^2)] - \mathbb{E}_{i\neq j}[\frac{(n)(n-1)}{n^2} x_i x_j] && \text{Counting the number of pairs of unequal heights} \\
&= \frac{n-1}{n}\sum_i \mathbb{E}[(x_i^2)] - \frac{n(n-1)}{n^2}\mathbb{E}[x_i]\mathbb{E}[x_j] && \text{Since the unequal pairs are independent} \\
&= \frac{n-1}{n}\sum_i \mathbb{E}[(x_i^2)] - \frac{n-1}{n}\mu^2 \\
&= \frac{n-1}{n}\sum_i (\sigma^2 + \mu^2) - \frac{n-1}{n}\mu^2 \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}
$$

which is smaller than σ^2.