Week 2: Maximum Likelihood Estimation

Instructor: Sergey Levine

1 Introduction: estimating parameters of a binomial distribution

Let's say that we want to predict the probability that a loaded coin (or, say, a thumbtack) will land on a particular side – either pointy end up (which we call heads) or point end down (which we call tails). We can frame this as trying to estimate the probability of an event. Call this event x, where $x \in \{T, H\}$. How can we *learn* $p(x = H)$? Well, as with all machine learning problems, we need four things: we need the data, the hypothesis space, the objective, and the algorithm.

Question. What is the data?

Answer. The data consists of samples of x, which we can obtain, for example, by flipping the coin (or thumbtack...). Imagine we flip it five times, we get samples: $x_1 = H$, $x_2 = T$, $x_3 = H$, $x_4 = T$, $x_5 = H$. We will use $\mathcal{D} = \{x_i\}$ to denote our *dataset* of N samples.

Question. What is the hypothesis space?

Answer. We would like to estimate $p(x = H)$ (since $p(x = T) = 1 - p(x = H)$). This is a binomial distribution, and it can be *parameterized* by a single real-valued parameter, θ, where $p(x = H) = \theta$. So the hypothesis space is $\theta \in [0, 1]$.

Question. What is the objective?

Answer. We need to design an objective. Fortunately, probability theory can be our guide here: we can try to find the hypothesis θ that makes the observed dataset $\mathcal{D}$ the most probable. This is the *maximum likelihood* solution. Finding the maximum likelihood solution is referred to as maximum likelihood estimation. The likelihood of the data is simply the probability of observing the entire dataset given the hypothesis θ:

$$p(\mathcal{D}|\theta) = p(x_1, x_2, x_3, x_4, x_5|\theta)$$

We assume that the individual samples are independent from each other and all distribution according to the same distribution $p(x)$. This assumption is very common in machine learning, and we call it "independently and identically distributed" (i.i.d.). Recall that if two variables x and y are independent, then $p(x, y) = p(x)p(y)$. So we can rewrite our likelihood as:

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$$

Now, $p(x_i|\theta) \in [0, 1]$, so as N increases, $p(\mathcal{D}|\theta) \to 0$, since it's the product of many numbers, all of which are less than 1 (unless $p(x|\theta)$ is deterministic). This is not bad by itself, because computers struggle to represent extremely small numbers, it's very convenient for us to instead use the log of the likelihood as our objective:

$$\log p(\mathcal{D}|\theta) = \sum_{i=1}^{N} \log p(x_i|\theta)$$

This is why we typically say that maximum likelihood estimation maximizes the log-likelihood. So our learning problem can be written as:

$$\theta \leftarrow \arg\max_{\theta} \sum_{i=1}^{N} \log p(x_i|\theta)$$

Question. What is the algorithm?

Answer. We need to solve the above optimization problem. We know that $p(x = H|\theta) = \theta$ and $p(x = T|\theta) = (1 - \theta)$. Let α_H be the number of heads in our dataset, and let α_T be the number of tails. We can then rewrite our log-likelihood as:

$$\sum_{i=1}^{N} \log p(x_i|\theta) = \alpha_H \log \theta + \alpha_T \log(1 - \theta).$$

Now, if we want to find the value of θ that minimizes this, we can simply compute the derivative with respect to θ and set it to zero:

$$\begin{aligned} \frac{d}{d\theta} \left[\alpha_H \log \theta + \alpha_T \log(1 - \theta)\right] &= \alpha_H \frac{d}{d\theta} \log \theta + \alpha_T \frac{d}{d\theta} \log(1 - \theta) \\ &= \frac{\alpha_H}{\theta} - \frac{\alpha_T}{1 - \theta} = 0. \end{aligned}$$

We then solve for θ:

$$\begin{aligned}
\frac{\alpha_H}{\theta} - \frac{\alpha_T}{1-\theta} &= 0 \\
\frac{\alpha_H}{\theta} &= \frac{\alpha_T}{1-\theta} \\
\frac{1-\theta}{\theta} &= \frac{\alpha_T}{\alpha_H} \\
\frac{1}{\theta} - 1 &= \frac{\alpha_T}{\alpha_H} \\
\frac{1}{\theta} &= \frac{\alpha_T}{\alpha_H} + 1 = \frac{\alpha_T + \alpha_H}{\alpha_H} \\
\theta &= \frac{\alpha_H}{\alpha_T + \alpha_H}.
\end{aligned}$$

This is precisely the answer we expect: simply count up the number of heads, and divide by the size of the dataset.

2 How much data do we need?

Given the above method for estimating the most probable parameter θ, a natural next question to ask is: how certain are we that this estimate is correct? Intuitively, we feel more confident in our estimate if we've seen more data: that is, if N (the number of samples in $\mathcal{D}$) is larger. But how many samples N do we need to achieve a certain level of confidence?

In machine learning theory, these kinds of questions are often analyzed through the "probably approximately correct" (PAC) framework. First, Hoeffding's inequality tells us that the probability that the error in our parameter estimate θ is greater than some constant ϵ can be bounded as following:

$$p(|\theta - \theta^\star| \geq \epsilon) \leq 2e^{-2N\epsilon^2}.$$

Unpacking this inequality, we can see that, as N increases, the probability of the error being greater than ϵ decreases *exponentially.* We will not cover the derivation of Hoeffding's inequality in this class (this is something you might see in a graduate class), but we will do a little exercise to understand how it can be used to determine how much data we need to confidently estimate the parameter θ.

Question. About how many samples (N) do we need in order to be certain with 95% probability that we've estimated θ with an error at most $\epsilon = 0.1$?

Answer. This is PAC learning: we want to know that we are probably (with 95% probability) approximately (within $\epsilon = 0.1$) correct. We typically use δ to denote the probability of being wrong, so that $\delta = 1 - 0.95 = 0.05$. So we would like $p(|\theta - \theta^\star| \geq \epsilon) \leq \delta$. One way to ensure that this is the case is to use the bound, and choose N such that

$$p(|\theta - \theta^\star| \geq \epsilon) \leq 2e^{-2N\epsilon^2} \leq \delta.$$

So now we simply solve for N:

$$\begin{aligned} 2e^{-2N\epsilon^2} &\leq \delta \\ \log 2 - 2N\epsilon^2 &\leq \log \delta \\ N &\geq \frac{\log 2 - \log \delta}{2\epsilon^2} = \frac{\log 2/\delta}{2\epsilon^2} \end{aligned}$$

Now we just plug in $\delta = 0.05$ and $\epsilon = 0.1$ to get:

$$N \geq \frac{\log 2/0.05}{2 \times 0.1^2} = \frac{\log 40}{0.02} \approx \frac{3.7}{0.02} = 185$$

So if we want to be certain with 95% probability that we are off by at most 0.1, it is sufficient to have 185 samples.

A variety of simple maximum likelihood estimation problems have bounds on sample complexity of this type. Some of the more complex estimation problems we will see later in the course do not have such nice closed form bounds, but the intuition is typically the same: the more samples we have, the more accurately we can estimate the model parameters. When we discuss learning theory, we will see some more general methods to estimate how the required amount of data changes as we vary our hypothesis class and dataset size.

3 Distributions over parameters

We saw how to compute the parameters θ that maximize the probability of the dataset. But here is an interesting thought: can we compute the parameters θ that are most probable *given* the dataset? Are they the same?

In order to answer this question, we must first construct the distribution over θ conditioned on the data: $p(\theta|\mathcal{D})$. We can obtain this from Bayes rule:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$$

We already know what $p(\mathcal{D}|\theta)$ is. The distribution $p(\theta)$ is called a prior – we'll come back to this in a second, but for now let's just say we don't know anything about θ, so $p(\theta) = 1$ (it's a constant). Since $\theta \in [0, 1]$, this integrates to 1. The denominator is

$$p(\mathcal{D}) = \int p(\mathcal{D}|\theta')p(\theta')d\theta'$$

That looks really complicated, but we can just not worry about it, since it's a constant, so we can just write:

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

and figure out the constant later. Let's plug in the equations for the binomial distribution:

$$p(\theta|\mathcal{D}) \propto \theta^{\alpha_H}(1-\theta)^{\alpha_T}$$

This kind of distribution is called the Beta distribution (with parameters α_H+1 and α_T+1), and it is normalized by the Beta function:

$$p(\theta|\mathcal{D}) = \frac{\theta^{\alpha_H}(1-\theta)^{\alpha_T}}{B(\alpha_H+1, \alpha_T+1)}$$

For reference $p(\theta|\mathcal{D})$ is called the posterior, because it is our distribution over θ after observing $\mathcal{D}$. $p(\theta)$ is the prior. The Beta function is just defined as the integral of $\theta^{\alpha_H}(1-\theta)^{\alpha_T}$ from 0 to 1. Since it does not depend on θ, we can find the most likely θ just by setting the derivative of $\log(p(\mathcal{D}|\theta)p(\theta))$ to zero. But that's exactly the same as what we had before, so the maximum likelihood estimate of θ is simply the most probable θ under a uniform prior.

Now imagine that if you believe that the coin is very likely to be a fair coin. Can we incorporate this *prior* information into the estimate of the most probable θ? We can do this by changing the prior $p(\theta)$ from a uniform prior to a more informative one. There are a number of different choices, but from looking at the form of the posterior, one thing that might jump out is that we should choose a form for $p(\theta)$ that makes it easy to multiply with the likelihood $p(\mathcal{D}|\theta)$. This is called a conjugate prior: a type of prior where the product with the distribution has an analytic answer. Since our distribution is binomial, the conjugate prior for this type of distribution is just the Beta distribution – the same type of distribution as our posterior! Note that in general, the conjugate prior does not have to be of the same type as the posterior.

The Beta distribution prior is given by

$$p(\theta) = \frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}{B(\beta_H, \beta_T)},$$

where β_H and β_T are the parameters of the prior. For example, if we believe that the coin is fair, we can set $\beta_H = \beta_T$. The larger the value, the more "confident" we are in this prior. If we look at the posterior, we get

$$p(\mathcal{D}|\theta)p(\theta) \propto \theta^{\alpha_H+\beta_H-1}(1-\theta)^{\alpha_T+\beta_T-1}$$

This is just a Beta distribution with parameters $\alpha_H+\beta_H$ and $\alpha_T+\beta_T$! We can compute the derivative and set it to zero to recover the most likely value of θ like we did before, but there is an even easier way: observe that the posterior with prior parameters β_H and β_T is exactly the same as the posterior with a uniform prior, if we instead observed $\alpha_H+\beta_H-1$ heads and $\alpha_T+\beta_T-1$ tails. That means that

$$\theta = \frac{\alpha_H+\beta_H-1}{\alpha_H+\beta_H+\alpha_T+\beta_T-2}$$

Note in particular as $N = \alpha_H+\alpha_T$ increases, the prior is gradually "forgotten," so the prior has the strongest effect on the estimate when N is small.