Week 2: Maximum Likelihood Estimation

Instructor: Sergey Levine

1 Recap: MLE for the binomial distribution

In the previous lecture, we covered maximum likelihood estimation for the binomial distribution. Let's recap the key ideas:

Question. What is the data?

Answer. The data consists of a set of samples from the binomial distribution $p(x)$. Let's assume that $x \in \{T, H\}$, then our dataset looks like $x_1 = H$, $x_2 = T$, etc. The entire dataset is denoted $\mathcal{D} = \{x_1, \ldots, x_N\}$.

Question. What is the hypothesis space?

Answer. The binomial distribution is defined by a single parameter, given as $\theta = p(x = H)$.

Question. What is the objective?

Answer. The objective in MLE is to maximize the probability of the data, given by $p(\mathcal{D}|\theta)$. Typically, we use the log-likelihood:

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

Note that this is equivalent to the objective $p(\theta|\mathcal{D})$ when the prior $p(\theta)$ is uniform. We can also use a non-uniform prior, such as a Beta distribution, to encode our *prior knowledge* about θ (e.g., a prior belief that θ encodes the probability of heads for a fair coin).

Question. What is the algorithm?

Answer. The algorithm must solve the following problem

$$\hat{\theta} \leftarrow \arg\max_{\theta} \log p(\mathcal{D}|\theta),$$

(or $p(\theta|\mathcal{D})$ in the Bayesian case). We can solve this problem in the case of the binomial distribution by computing the derivative and setting it to zero. For more complex MLE problems, we might require a more sophisticated optimization algorithm. We'll see some examples of this later in the class, but for now, let's go through MLE for a different class of distributions.

2 Continuous data: Gaussian distributions

What if instead of predicting whether the coin (or thumbtack...) will land heads or tails, we instead want to predict the probability that it will land at a particular point on the table (imagine for now that we only care about horizontal position – 1 dimension)? Now the variable x that we would like to model is real-valued. When dealing with real-valued random variables, one very popular choice of distribution is the Gaussian or Normal distribution, given by

$$p(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Question. What is the hypothesis space if we want to model x using a Gaussian distribution?

Answer. The Gaussian is defined by two parameters: the mean μ and the standard deviation σ. Intuitively, μ corresponds to the "center" of the Gaussian, and σ corresponds to its width. The hypothesis space is fully defined by $\theta = \{\mu, \sigma\}$, where $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$.

A normally distributed random variable is typically written as

$$x \sim \mathcal{N}(\mu, \sigma^2).$$

Gaussians have a few really useful properties that make them a popular choice for modeling continuous random variables. First, affine transformations of Gaussians are themselves Gaussian: if $x \sim \mathcal{N}(\mu, \sigma^2)$, and $y = ax + b$, then $y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$. Second, the sum of two Gaussian random variables is also normally distributed: if $x \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $y \sim \mathcal{N}(\mu_y, \sigma_y^2)$, and $z = x + y$, then $z \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$. There are also natural generalizations of the univariate normal distribution to the multivariate case, where $\vec{x}$ is a multidimensional vector: in that case, μ is also a vector, and instead of the standard deviation σ, we use the *covariance matrix* Σ, which is a $d \times d$ matrix (where d is the dimensionality of $\vec{x}$). But for now, let's work with univariate Gaussians.

Say that we record a dataset of samples from our (unknown) Gaussian, e.g. $x_1 = 0.2$, $x_2 = 0.35$, $x_4 = 0.5$, etc. Our goal is to learn the parameters μ and σ.

Like before, we can write the learning problem as

$$\hat{\mu}, \hat{\sigma} = \arg\max_{\mu,\sigma} \sum_{i=1}^{N} \log p(x_i)$$

Let's derive the log likelihood:

$$\begin{aligned}
\sum_{i=1}^{N} \log p(x_i) &= \sum_{i=1}^{N} \log \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right) \\
&= \sum_{i=1}^{N} -\log \sigma - \frac{1}{2}\log 2\pi - \frac{(x_i-\mu)^2}{2\sigma^2} \\
&= -N \log \sigma - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2} + \text{const.}
\end{aligned}$$

Now, let's compute the optimal mean:

$$\begin{aligned}
\frac{d}{d\mu}\left[-N \log \sigma - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2} + \text{const} \right] &= -\sum_{i=1}^{N} \frac{d}{d\mu} \frac{(x_i-\mu)^2}{2\sigma^2} \\
&= \sum_{i=1}^{N} \frac{x_i - \mu}{\sigma^2} = 0
\end{aligned}$$

Rearranging the terms, we get:

$$\begin{aligned}
\sum_{i=1}^{N} \frac{x_i}{\sigma^2} &= N \frac{\mu}{\sigma^2} \\
\frac{1}{N} \sum_{i=1}^{N} x_i &= \mu
\end{aligned}$$

This is the answer we expect: the optimal mean μ is the average value of all of our data points. Now let's repeat the process for the standard deviation:

$$\begin{aligned}
\frac{d}{d\sigma}\left[-N \log \sigma - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2} + \text{const} \right] &= \frac{d}{d\sigma}\left[-N \log \sigma \right] - \sum_{i=1}^{N} \frac{d}{d\sigma}\left[\frac{(x_i-\mu)^2}{2\sigma^2} \right] \\
&= -\frac{N}{\sigma} + \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{\sigma^3} = 0
\end{aligned}$$

Rearranging, we get:

$$-\frac{N}{\sigma} + \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^3} = 0$$
$$\sum_{i=1}^{N} (x_i - \mu)^2 = N\sigma^2$$
$$\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 = \sigma^2$$

Again, this is the equation we would expect for the variance, and the standard deviation is $\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}$.

3 Bayesian learning with Gaussians

Just like we did with the binomial distribution, we can also use Bayesian learning with Gaussian distributions.

Question. What is the objective in Bayesian learning?

Answer. The objective is the (log) probability of the parameters $\theta = \{\mu, \sigma\}$ given the data:

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$
$$\log p(\theta|\mathcal{D}) = \log p(\mathcal{D}|\theta) + \log p(\theta) + \text{const}$$

For this exercise, let's assume that we know the standard deviation σ, and we're just trying to learn μ (we'll see how to build a prior on σ later). The conjugate prior for the mean of a Gaussian distribution is simply another Gaussian, with parameters μ_0 and σ_0:

$$p(\mu) = \frac{1}{\sigma_0 \sqrt{2\pi}} e^{-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}}$$

If we evaluate the posterior, we get:

$$\begin{aligned} \log p(\mu|\mathcal{D}) &= \log p(\mathcal{D}|\mu) + \log p(\mu) + \text{const} \\ &= -N \log \sigma - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2} - \log \sigma_0 - \frac{(\mu - \mu_0)^2}{2\sigma_0^2} + \text{const} \end{aligned}$$

Since all we want is a distribution over μ, we can fold any terms that don't depend on μ into the constant (which we'll figure out later), giving us

$$\log p(\mu|\mathcal{D}) = -\sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2} - \frac{(\mu - \mu_0)^2}{2\sigma_0^2} + \text{const}$$

We can expand the quadratics in the numerators to get:

$$\begin{aligned}
\log p(\mu|\mathcal{D}) &= -\sum_{i=1}^{N} \frac{x_i^2 + \mu^2 - 2\mu x_i}{2\sigma^2} - \frac{\mu^2 + \mu_0^2 - 2\mu_0\mu}{2\sigma_0^2} + \text{const} \\
&= -\sum_{i=1}^{N} \frac{x_i^2}{2\sigma^2} - \mu^2 \frac{N}{2\sigma^2} + \mu \sum_{i=1}^{N} \frac{2x_i}{2\sigma^2} - \mu^2 \frac{1}{2\sigma_0^2} - \frac{\mu_0^2}{2\sigma_0^2} + \mu \frac{2\mu_0}{2\sigma_0^2} + \text{const} \\
&= -\mu^2 \left[\frac{N}{2\sigma^2} + \frac{1}{2\sigma_0^2} \right] + \mu \left[\sum_{i=1}^{N} \frac{2x_i}{2\sigma^2} + \frac{2\mu_0}{2\sigma_0^2} \right] + \text{const}
\end{aligned}$$

Now, let

$$\begin{aligned}
\sigma_1 &= \left[\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \right]^{-1} \\
\mu_1 &= \left[\sum_{i=1}^{N} \frac{x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right] \sigma_1
\end{aligned}$$

We now have

$$\begin{aligned}
\log p(\mu|\mathcal{D}) &= -\frac{\mu^2}{2\sigma_1} + \frac{\mu\mu_1}{\sigma_1} + \text{const} \\
&= -\frac{\mu^2 - 2\mu\mu_1}{2\sigma_1} + \text{const} \\
&= -\frac{\mu_1^2 + \mu^2 - 2\mu\mu_1}{2\sigma_1} + \text{const} \\
&= -\frac{(\mu - \mu_1)^2}{2\sigma_1} + \text{const} \\
&= -\log \sigma_1 - \frac{1}{2}\log 2\pi - \frac{(\mu - \mu_1)^2}{2\sigma_1} + \text{const}
\end{aligned}$$

The last line is precisely the equation for a Gaussian with mean μ_1 and standard deviation σ_1. We know therefore that the constant on the last line is zero, because a Gaussian integrates to one, and therefore we have recovered the form of the posterior. It is again Gaussian.

If we need to estimate the standard deviation σ, we typically put a prior instead on the variance σ^2, and the conjugate prior is an inverse-gamma distribution (you do not need to know this for homeworks or exams). This is a distribution over positive real numbers, and is given by

$$p(\sigma^2) = \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-\alpha-1} \exp\left(\frac{-\beta}{\sigma^2} \right)$$

If we need to estimate *both* σ and μ, we use the normal inverse-gamma distribution, which is simply the product of a normal distribution on μ and an inverse-gamma on σ^2. The posterior will be normal inverse-gamma.