6.7900 Fall 2024: Lecture Notes 3

1 Recap

Recall our setup: we observe training data $\mathcal{D} = \{y^{(n)}\}_{n=1}^N$. We assume the training data and future data are i.i.d., and we would like to approximate the distribution of the next data point we might see, $p(y)$. Last time, we introduced two ideas:

1. We can use the *empirical distribution* over the training data.
2. We can introduce a parametric model $p(y \mid \theta)$, and choose the *maximum likelihood estimate* $\hat{\theta}$ as the parameter.

Throughout this note, we will consider the following running example: we have $y^{(n)} \in \{0, 1\}, y^{(n)} \sim \mathrm{Ber}(\theta)$ are i.i.d. with $\theta \in (0, 1)$. For example, θ might be the proportion of people of certain type catching a certain disease, and y could be an observation of whether a particular individual of the type caught the disease.

Now, suppose we have only 3 data points $y^{(1)} = y^{(2)} = y^{(3)} = 0$. The empirical distribution is $\hat{p}(y = 1) = 0$, and the MLE estimate $\hat{\theta}$ is arbitrarily close to 0. This is clearly problematic: after observing three negative data points, we are essentially predicting positive results are impossible, which might also be contrary to our knowledge of the disease. This toy example illustrates three problems with the MLE approach:

1. It *overfits* to the training data.
2. It has no *uncertainty quantification* - we don't know how confident we are when we say $\hat{\theta}$ is arbitrarily close to 1.
3. It does not leverage *domain knowledge* - in this case, our belief of the disease before observing any training data.

In this lecture, we will introduce the **Bayesian approach** of parameter estimation, which will address overfitting, quantify uncertainty, and incorporate domain knowledge.

2 Bayes Theorem

To start, let us recall the Bayes Theorem:

Suppose (y, θ) are realizations from a joint distribution of two random variables (Y, Θ). If $p(y) > 0$, then

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)}.$$

This can be derived easily from the relationship between joint probability and conditional probability:

$$p(\theta \mid y)p(y) = p(y, \theta) = p(y \mid \theta)p(\theta).$$

For our purpose, we will interpret y as data and θ as parameter. Thus,

1. $p(y \mid \theta)$ is the **likelihood** model,
2. $p(\theta)$ is the **prior**, which reflects domain knowledge,
3. $p(\theta \mid y)$ is the **posterior**, which can be interpreted as our updated knowledge of θ after observing y, and
4. $p(y)$ is the **evidence**, which we will ignore for our purpose.

For our purpose, we will use Bayes Theorem on the whole training dataset instead of a single data point, i.e.:

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})} \propto_\theta p(\mathcal{D} \mid \theta)p(\theta),$$

where we say $f(\theta) \propto_\theta g(\theta)$ if there exists $c \neq 0$ that is constant in θ and $f(\theta) = cg(\theta)$. Note that to perform Bayesian update, we need to use the prior and the likelihood model to compute the posterior. Going back to our running example of disease diagnosis, we already have a likelihood model of

$$p(y \mid \theta) = \mathrm{Ber}(\theta).$$

So what likelihood model should we use?

3 Beta Distribution

We will introduce the **Beta distribution**:

$$\mathrm{Beta}(\theta \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}.$$

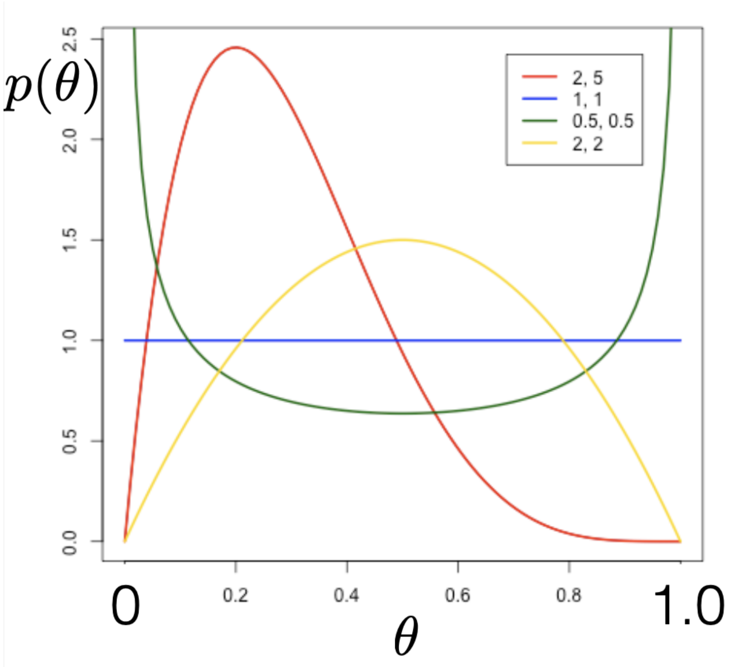

Figure 1: Beta distribution.

Note that a and b are parameters for the distribution of our parameter of interest (θ) - we call them *hyperparameters*. The second term $\theta^{a-1}(1-\theta)^{b-1}$ is a function of θ and is called the *kernel*; the first term $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ is a normalization function constant in θ, where Γ is the *Gamma function*. For our purpose, it is sufficient to know that Gamma functions behave similar to the factorials; in particular, for any $t > 0$, we have

$$\Gamma(t+1) = t\Gamma(t).$$

Figure 1 shows plots of several Beta distributions. Some observations we can draw here:

1. $\mathrm{Beta}(1,1)$ is uniform distribution.
2. When a and b get very small ($a, b << 1$), the distribution has high density on tails, i.e., near 0 and 1.
3. When a and b get very large, the distribution peaks somewhere in the middle and has small density near 0 and 1.
4. When $a > b$, the distribution is right-skewed; when $a < b$, the distribution is left-skewed.

It will be useful to know the expectation and variance for Beta distributions:

Exercise: Show that if $\theta \sim \mathrm{Beta}(a, b)$, then

$$\mathbb{E}(\theta) = \frac{a}{a+b}$$

and

$$\mathrm{Var}(\theta) = \frac{ab}{(a+b)^2(a+b+1)}.$$

We make two remarks about the expectation and variance of Beta distributions:

1. The expectation matches the observations we drew from Figure 1.
2. From the variance expression, we can see that if either a or b gets very big, the variance becomes very small, so we will have very little uncertainty about θ.

4 From Prior to Posterior

We will assume our prior $p(\theta) = \mathrm{Beta}(\theta \mid a, b)$ is a Beta distribution for some hyperparameters a and b. Then with Bayes Theorem, we can compute the posterior:

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})} \propto_\theta p(\mathcal{D} \mid \theta)p(\theta) = p(\theta)\prod_{n=1}^{N} p(y^{(n)} \mid \theta).$$

Going back to our running example, we have

$$\begin{aligned}
p(\theta \mid \mathcal{D}) &\propto_\theta \mathrm{Beta}(\theta \mid a, b)\prod_{n=1}^{N} \mathrm{Ber}(y^{(n)} \mid \theta) \\
&\propto_\theta \theta^{a-1}(1-\theta)^{b-1}\prod_{n=1}^{N} \mathrm{Ber}(y^{(n)} \mid \theta) \\
&= \theta^{a-1}(1-\theta)^{b-1}\prod_{n=1}^{N} \left(\theta^{y^{(n)}}(1-\theta)^{1-y^{(n)}}\right) \\
&= \theta^{a-1+\sum_{n=1}^{N} y^{(n)}}(1-\theta)^{b-1+\sum_{n=1}^{N}(1-y^{(n)})} \\
&\propto_\theta \mathrm{Beta}\left(\theta \,\middle|\, a + \sum_{n=1}^{N} y^{(n)}, b + \sum_{n=1}^{N}(1-y^{(n)})\right).
\end{aligned}$$

Now, because the Beta distribution integrates to 1, we can replace $\propto$ with equality:

$$p(\theta \mid \mathcal{D}) = \mathrm{Beta}\left(\theta \middle| a + \sum_{n=1}^{N} y^{(n)}, b + \sum_{n=1}^{N} (1 - y^{(n)})\right).$$

Several remarks:

1. If we do not have any data, then the posterior is just the prior, i.e., $\mathrm{Beta}(\theta \mid a, b)$.

2. As we get more data, the hyperparameters of the Beta distribution becomes larger, so the variance for θ becomes smaller. In particular, we can use the variance of the posterior distribution to *quantify uncertainty*.

3. Just like the prior, the posterior will always be a Beta distribution. Because the posterior will always be in the same family as the prior, we say our prior distribution is a *conjugate prior*.

We can also compute the mean of the posterior:

$$\begin{aligned}\mathbb{E}(\theta \mid \mathcal{D}) &= \frac{\sum_{n=1}^{N} y^{(n)} + a}{N + a + b} \\ &= \frac{N}{N + a + b} \frac{\sum_{n=1}^{N} y^{(n)}}{N} + \frac{a + b}{N + a + b} \frac{a}{a + b}.\end{aligned}$$

Looking at the first equation, we can interpret a and b as “pseudocounts”: from the posterior perspective, adding 3 to a is equivalent to adding three positive samples to $\mathcal{D}$. Looking at the last equation, we notice that $\frac{a}{a+b}$ is the mean of the prior distribution, and $\frac{\sum_{n=1}^{N} y^{(n)}}{N}$ is the MLE of θ from $\mathcal{D}$. Thus, we can interpret the posterior mean as a weighted average of the prior mean and the MLE. By examining the weights, we can notice that

1. We weigh prior higher with fewer data points (thus leveraging *domain knowledge*), and we weigh the MLE higher as we get more data points;

2. If $a, b > 0$, the posterior will never be 0 or 1, which reduces *overfitting*,

which addresses the three problems we had with MLE.

5 From Posterior to Predictive

In the end, our goal is to approximate the distribution of a future data point. With the Law of Total Probability, we can obtain the **posterior predictive distribution** from the posterior distribution:

$$\begin{aligned} p(y^{(N+1)} \mid \mathcal{D}) &= \int_\theta p(y^{(N+1)}, \theta \mid \mathcal{D}) d\theta \\ &= \int_\theta p(y^{(N+1)} \mid \theta, \mathcal{D}) p(\theta \mid \mathcal{D}) d\theta \\ &= \int_\theta p(y^{(N+1)} \mid \theta) p(\theta \mid \mathcal{D}) d\theta. \end{aligned}$$

Note that the last equation uses the fact that all data points are i.i.d. conditioned on the parameter. In our running example, we have

$$\begin{aligned} p(y^{(N+1)} = 1 \mid \mathcal{D}) &= \int_\theta p(y^{(N+1)} = 1 \mid \theta) p(\theta \mid \mathcal{D}) d\theta \\ &= \int_\theta \theta p(\theta \mid \mathcal{D}) d\theta = \mathbb{E}(\theta \mid \mathcal{D}) \\ &= \frac{\sum_{n=1}^N y^{(n)} + a}{N + a + b}. \end{aligned}$$

6 Application: Streaming Data

Bayesian update is very convenient when we are dealing with **streaming data**: we get some batch of data at a time, and we do not have the space to store all data, so we want to draw some conclusions on the existing data and perform updates as the new batch comes in.

Formally, suppose we first get one batch of data $\mathcal{D}_1 = \{y^{(n)}\}_{n=1}^{N_1}$, and after a while we get another batch $\mathcal{D}_2 = \{y^{(N+n)}\}_{n=1}^{N_2}$. The key observation is that we can treat our posterior after the first batch as our prior for the new batch. We will show this is equivalent to updating using both batches at once:

$$\begin{aligned} p(\theta \mid \mathcal{D}_2, \mathcal{D}_1) &\propto_\theta p(\mathcal{D}_2 \mid \theta, \mathcal{D}_1) p(\theta \mid \mathcal{D}_1) \\ &= p(\mathcal{D}_2 \mid \theta) p(\theta \mid \mathcal{D}_1) \\ &\propto_\theta p(\mathcal{D}_2 \mid \theta) p(\mathcal{D}_1 \mid \theta) p(\theta) \\ &= \left(\prod_{n=1}^{N_1} p(y^{(n)} \mid \theta) \right) \left(\prod_{n=1}^{N_2} p(y^{(n+N_1)} \mid \theta) \right) p(\theta) \end{aligned}$$

$$= \left(\prod_{n=1}^{N_1+N_2} p(y^{(n)} \mid \theta) \right) p(\theta)$$

$$= p(\mathcal{D}_2, \mathcal{D}_1 \mid \theta) p(\theta).$$

As an example, in our running example, the posterior hyperparameters are simply obtained by adding the number of positive / negative observations to the two prior hyperparameters, so updating two batches one by one is obviously equivalent to treating the two batches as a single batch.

7 Extension: MAP Estimation and Prediction

In our running example, we conveniently chose Beta distribution for our prior and Bernoulli distribution for our model. However, with more complex models, Bayesian posteriors and posterior predictives might not be in closed form and might be difficult to compute. Even though MLE have the problems we discussed, MLE are generally easier to compute. Thus, in practice, an intermediate choice is the **maximum a posteriori** (MAP) estimate, where we choose the parameter to maximize the density under the posterior distribution:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta \mid \mathcal{D}).$$

In our running example, the MAP estimate would be

$$\hat{\theta}_{MAP} = \frac{\sum_{n=1}^{N} y^{(n)} + a - 1}{N + (a-1) + (b-1)}.$$

Exercise: Prove this!

Note that because this is the MLE of the posterior distribution, it still uses the prior distribution instead of blindly follow the empirical distribution as the MLE estimator.