DS 102 Data, Inference, and Decisions **Fall 2019**

Lecture 8: Bayesian Hierarchical Models I

Lecturer: Ramesh Sridharan

1 Parameter Estimation

Say we observe data $y_1, y_2, \ldots, y_n$. Let's assume the datapoints are independently and identically distributed (i.i.d.) from the same distribution. The datapoints are independently distributed if

$$\mathbb{P}\left(\bigcap_{i \in \mathcal{J}} y_i\right) = \prod_{i \in \mathcal{J}} \mathbb{P}(y_i)$$

for all $\mathcal{J} \subseteq \{1, 2, \ldots, n\}$. The datapoints are identically distributed if

$$y_i \sim \mathcal{P}(\theta),$$

for all $i \in \{1, 2, \ldots, n\}$, where $\mathcal{P}$ is some family of distributions and $\theta \in \mathbb{R}^d$ is the parameter for that distribution. For example we could have

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

for all $i \in \{1, 2, \ldots, n\}$. Here $\theta = (\mu, \sigma^2)^\top$.

Let's say we don't know θ, or a subset of θ's coordinates. In previous lectures we looked at hypothesis testing which we can use to test specific conditions on θ. For example, in the case where we know that the observed data comes from a Gaussian distribution with either mean $\mu = 0$ or $\mu = 5$ we might conduct a hypothesis test to differentiate between the two cases. Or we might try to determine if $\mu > 1$ or not.

However, instead of simply testing whether conditions on θ are true or not, we might want to directly estimate the parameter instead. You saw this idea being partially developed in Lectures 6 and 7, we will further expand upon this idea in this Lecture. From here on we will denote any estimator of a parameter by putting a hat above it. For example, an estimator of θ will be $\hat{\theta}$.

2 Frequentist Estimation

One simple frequentist estimator comes about by considering the likelihood function over all our data: $\mathcal{L}(\theta) = \mathbb{P}(y_1, y_2, \ldots, y_n; \theta)^1$. We can think of the likelihood as the probability that the data

[1] Note that we use a semi-colon to indicate a specific value of θ. It has much the same meaning as the vertical bar used for conditioning, except that it implies we think of θ as a fixed unknown variable instead of a random variable. Although the two symbols are often used interchangeably.

was generated by a distribution assuming that it has its parameter set to θ. One natural thing to do would be to find the value that maximizes this probability. We call this value the maximum likelihood estimate (MLE) and denote it by $\hat{\theta}_{MLE}$. We usually proceed as follows when doing maximum likelihood estimation

1. We then define a likelihood model $\mathbb{P}(y;\theta)$ where y will be an observed datapoint and θ is a fixed unknown parameter. Note that this is often a subjective (albeit informed) decision. We very rarely know the true distribution from which the data comes from and must instead approximate to the best of our ability.

2. We collect our data $y_1, \ldots, y_n$.

3. Finally we attempt to better understand the world by finding the maximum likelihood estimate (MLE). The process involves the following steps:

   (a) Calculate the likelihood function over all our data

   $$\mathcal{L}(\theta) = \mathbb{P}(y_1, \ldots, y_n; \theta) = \prod_{i=1}^{n} \mathbb{P}(y_i; \theta),$$

   where the second equality used the i.i.d assumption on the data.

   (b) Take the log of the likelihood

   $$\ell(\theta) = \log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log \mathbb{P}(y_i; \theta).$$

   This step isn't compulsory but it usually significantly simplifies the next step.

   (c) Find the MLE which is defined as

   $$\hat{\theta}_{MLE} = \operatorname*{argmax}_{\theta} \mathcal{L}(\theta) = \operatorname*{argmax}_{\theta} \ell(\theta)$$

   where the second equality used the fact that $\operatorname{argmax}_x f(x) = \operatorname{argmax}_x g(f(x))$ whenever g is a strictly increasing function, and that the $\log$ is a strictly increasing function. Assuming the likelihood function is well behaved we can find this maximum argument by taking a derivative, setting it to 0, and solving for θ.

For concrete examples of maximum likelihood estimation look at Question 1 of Discussions 1 and 3.

3 Bayesian Estimation

Unlike frequentist estimation, in Bayesian estimation we treat the unknown parameter as a random variable instead of a fixed (but unknown) quantity. Since θ is often reserved for fixed parameters we will now denote our random parameter as x instead. Here the primary quantity of interest will no longer be the likelihood $\mathbb{P}(y_1, y_2, \ldots, y_n|x)$ but instead the posterior distribution $\mathbb{P}(x|y_1, y_2, \ldots, y_n)$. The posterior can be thought of as our belief on the value of our parameter x given that we have observed our data. Now once again a natural estimator comes about by maximizing the posterior probability, this is called the maximum a posteriori (MAP) estimate and is denoted by $\hat{x}_{MLE}$. To find the MAP estimate we go through the following steps

1. Define a likelihood model $\mathbb{P}(y|x; \theta_y)$ where x is the unknown parameter (equivalent to θ in the frequentist setting), y is an observed datapoint, and θ_y is a known hyperparameter that we set.

2. Define a prior distribution $\mathbb{P}(x; \theta_x)$ on our unknown parameter x. Here θ_x is also a known hyperparameter that we set. As in the frequentist setting there is a lot of subjectivity in the last two steps, we almost never know the true prior nor the true likelihood but we hope to make an informed decision. It's also worth noting that our hyperparameters θ_x and θ_y can also be considered as part of the subjective task of picking the right model to represent our prior and likelihood.

3. Collect our data $y_1, \ldots, y_n$.

4. Finally we attempt to better understand the world by finding the maximum a posteriori (MAP) estimate. The process involves the following steps:

   (a) Calculate the likelihood function over all our data

   $$\mathbb{P}(y_1, \ldots, y_n|x; \theta_y) = \prod_{i=1}^{n} \mathbb{P}(y_i|x; \theta_y),$$

   where the second equality used the i.i.d assumption on the data.

   (b) Write down the posterior distribution, this can be done through Bayes' theorem

   $$\mathbb{P}(x|y_1, y_2, \ldots y_n) = \frac{\mathbb{P}(y_1, y_2, \ldots, y_n|x; \theta_y)\mathbb{P}(x; \theta_x)}{\mathbb{P}(y_1, y_2, \ldots, y_n)} = \frac{\prod_{i=1}^{n} \mathbb{P}(y_i|x; \theta_y)\mathbb{P}(x; \theta_x)}{\mathbb{P}(y_1, y_2, \ldots, y_n)}.$$

   (c) Find the MAP which is defined as

   $$\hat{x}_{MAP} = \operatorname*{argmax}_x \mathbb{P}(x|y_1, y_2, \ldots y_n) = \operatorname*{argmax}_x \prod_{i=1}^{n} \mathbb{P}(y_i|x; \theta_y)\mathbb{P}(x; \theta_x)$$

   where the second equality uses the fact that $\operatorname{argmax}_x f(x)/c = \operatorname{argmax}_x f(x)$ for a positive constant c, and that $\mathbb{P}(y_1, y_2, \ldots, y_n)$ is a positive constant with respect to x. Assuming the posterior is well behaved enough we can take its derivative with respect to x, set it to 0 and solve for x to find the MAP estimator, you may also use the log trick of Section 2 if it facilitates calculation.

3.1 Example: Beta prior and Binomial likelihood

As an example of computing the posterior distribution consider the situation where we have $y \sim Binom(n, x)$, where n is fixed, and $x \sim Beta(r, s)$ for fixed and known $r > 0$ and $s > 0$. Our hyperparameters in this case are $\theta_y = n$ and $\theta_x = (r, s)^\top$. In that case our prior is

$$\mathbb{P}(x) = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)} x^{r-1}(1-x)^{s-1},$$

where Γ is the gamma function defined as

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx.$$

You can think of the gamma function as the generalization of the factorial function to all real numbers, although for the purpose of this example it will only play a relatively minor role since it will only serve to normalize the distribution. Then we can express the prior distribution on x as

$$\mathbb{P}(x; r, s) = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)} x^{r-1}(1-x)^{s-1}$$

the likelihood as

$$\mathbb{P}(y|x; n) = \binom{n}{y} x^y (1-x)^{n-y}.$$

Now let’s try to find the posterior distribution of x by using Bayes’ theorem,

$$\begin{aligned}\mathbb{P}(x|y) &= \frac{\mathbb{P}(y|x; n)\mathbb{P}(x; r, s)}{\mathbb{P}(y)} \\ &= \binom{n}{y} \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)} \frac{x^y(1-x)^{n-y}x^{r-1}(1-x)^{s-1}}{\int_0^1 \mathbb{P}(y|z; n)\mathbb{P}(z; r, s)dz,}\end{aligned}$$

where we have expanded the denominator using the fact that $\mathbb{P}(y) = \int \mathbb{P}(y|x)\mathbb{P}(x)dx$ when x is a continuous random variable, and we have expanded the definition of the likelihood and the prior in the numerator. This quickly got out of hand with many hard to evaluate terms! But one key observation we can make is that what we really only care about the shape of the posterior here not its scale. This is because

1. we are ultimately trying to find the argmax (with respect to x) which doesn’t change as we scale the posterior by a constant,

2. if we can show that a function has the same shape as a known distribution, it can only correspond to that distribution once we scale it down so it integrates to 1 (see Figure 8.1).

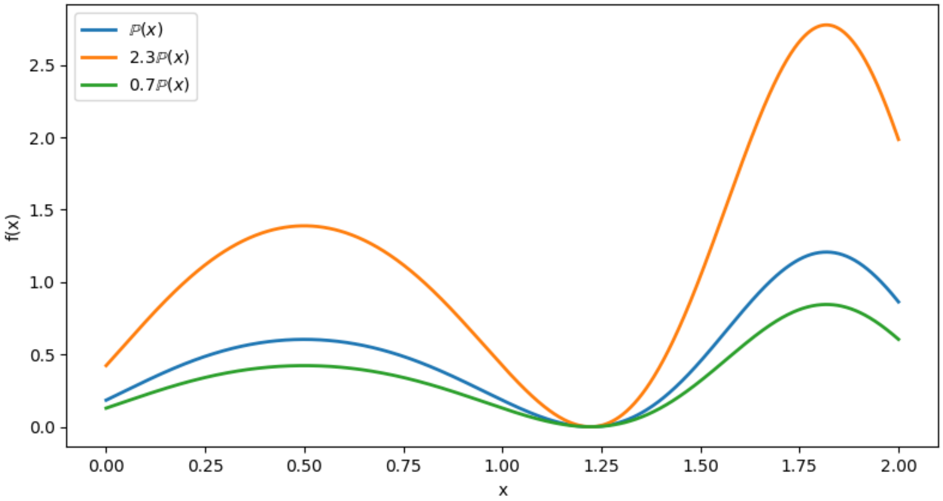

Figure 8.1: Different scalings of some distribution ℙ(x). Even though only one of these curves integrate to 1 we can see that they all uniquely identify the same distribution since they all have the same shape.

With this observation in mind we can discard any scaling term that doesn't depend on x, hence

$$\begin{aligned}\mathbb{P}(x|y) &\propto_x x^y(1-x)^{n-y}x^{r-1}(1-x)^{s-1}\\ &= x^{y+r-1}(1-x)^{n-y+s-1},\end{aligned}$$

where $\propto_x$ means "proportional to" when all variables other than x are treated as constants. But this is simply an unnormalized version of $Beta(y+r, n-y+s)$. So our posterior belongs to the same family of distributions as our prior and can simply be computed by adding the number of successes and failures in the binomial trial to the prior hyperparameters.

When the prior and posterior belong to the same family of distributions, we say that they are *conjugate distributions*, with the prior being called a *conjugate prior*. As we will see, we can't always have conjugate distributions in which case the posterior becomes much harder to compute.

See Discussion 4 for further details and intuition on this specific prior/likelihood pair. In particular we look at how to find the MAP estimate given the posterior.

3.2 Example: Gaussian priors and Gaussian likelihood

Let's look at a continuous example of conjugate distributions. We have $y \sim \mathcal{N}(x, \sigma^2)$ where σ^2 is fixed and known, $x \sim \mathcal{N}(\mu_x, \sigma^2)$, and μ_x is fixed and known. Then the prior on x is given by

$$\mathbb{P}(x; \mu_x, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_x)^2}{2\sigma^2}\right).$$

The likelihood is given by

$$\mathbb{P}(y|x;\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-x)^2}{2\sigma^2}\right).$$

Computing the posterior gives us

$$\begin{aligned}
\mathbb{P}(x|y) &\propto_x \exp\left(-\frac{(y-x)^2 + (x-\mu_x)^2}{2\sigma^2}\right) \\
&= \exp\left(-\frac{2(x^2 - xy - x\mu_x) + y^2 + \mu_x^2}{2\sigma^2}\right) \\
&= \exp\left(-\frac{2(x^2 - xy - x\mu_x)}{2\sigma^2}\right) \exp\left(-\frac{y^2 + \mu_x^2}{2\sigma^2}\right) \\
&\propto_x \exp\left(-\frac{x^2 - x(y+\mu_x)}{2\sigma^2/2}\right) \\
&= \exp\left(-\frac{x^2 - x(y+\mu_x) + (y+\mu_x)^2/4 - (y+\mu_x)^2/4}{2\sigma^2/2}\right) \\
&\propto_x \exp\left(-\frac{x^2 - x(y+\mu_x) + (y+\mu_x)^2/4}{2\sigma^2/2}\right) \\
&= \exp\left(-\frac{(x-(y+\mu_x)/2)^2}{2\sigma^2/2}\right).
\end{aligned}$$

So the posterior distribution is also Gaussian of the form $\mathcal{N}(\frac{y+\mu_x}{2}, \frac{\sigma^2}{2})$. So we once again have a conjugate prior.

3.3 Example: Non-conjugate priors

Let's look at a situation where our priors aren't conjugate. Here, assume $y \sim \mathcal{N}(x, \sigma^2)$ where σ^2 is fixed and known, and the pdf of x is given by

$$\mathbb{P}(x) = \begin{cases} \cos(x), & x \in [0, \frac{\pi}{2}] \\ 0, & \text{otherwise.} \end{cases}$$

Computing the posterior distribution gives us

$$\mathcal{P}(x|y) \propto_x \begin{cases} \exp\left(\frac{1}{2\sigma^2}(x^2 - 2xy)\right) cos(x), & x \in [0, \frac{\pi}{2}] \\ 0, & \text{otherwise.} \end{cases}$$

Here we see that our posterior doesn't correspond to any distribution that we know of. This isn't a huge issue if we just want to compute the MAP since we can just take the derivative of this expression. However, if we wanted to compute some other quantity such as the mean of the posterior, we would need to evaluate $\mathbb{P}(y)$ which involves computing an integral with no closed form. In future lectures we will see how to handle this by sampling.

4 Gaussian Mixture Models

We now turn to a particularly ubiquitous Bayesian model, the Gaussian mixture model (GMM). We will look at the model in its full generality in the next lecture but for now let's consider a simple scenario.

Assume we have a dataset of i.i.d heights $y_1, y_2, \dots, y_n$. We could model the distribution of heights as a simple Gaussian, however we know that the sex of the person plays a big role in determining their height. Hence a better way to model this situation might be as a mixture of two Gaussian distributions.

Given that our dataset only includes height information (and not the sex of the participants) we instead treat their sex as a hidden binary random variable x_i. Where $x_i = 0$ if the participant is female and $x_i = 1$ if they are male. These types of unseen variables are called *latent variables*. We can put a Bernoulli prior on the sex x of any participant as

$$\pi_0 = \mathbb{P}(x = 0)$$
$$\pi_1 = \mathbb{P}(x = 1).$$

And then the likelihood on any datapoint y is given by

$$\mathbb{P}(y|x) = \begin{cases} \mathcal{N}_0 = \mathcal{N}(y; \mu_0, \sigma^2), & x = 0 \\ \mathcal{N}_1 = \mathcal{N}(y; \mu_1, \sigma^2), & x = 1. \end{cases}$$

where μ_0, μ_1, σ^2 are all fixed and known. Then the posterior is given by

$$\mathbb{P}(x_i|y_i) = \frac{(\pi_1 \mathcal{N}_1)^x (\pi_0 \mathcal{N}_0)^{1-x}}{\pi \mathcal{N}_1 + (1-\pi)\mathcal{N}_0}.$$

An example of such a GMM distribution is shown in Figure 8.2. In the next lecture we will look at the case where the Gaussian parameters are unknown and how we can estimate them.

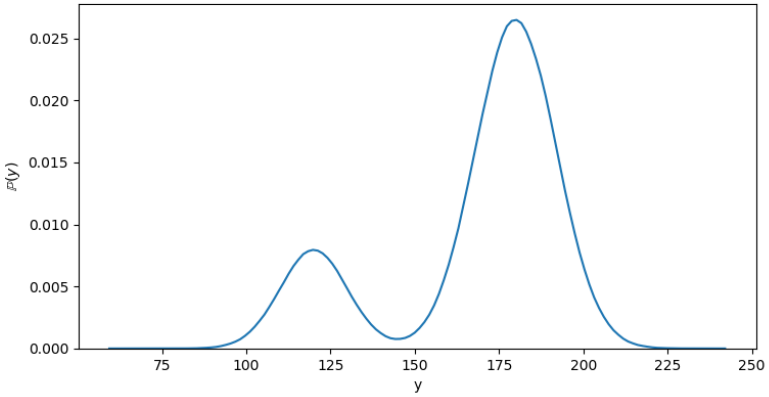

Figure 8.2: An example GMM where $\pi_0 = 0.2$, $\pi_1 = 0.8$, $\mu_0 = 120$, $\mu_1 = 180$, and $\sigma = 11$.