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

Lecture 8: Introduction to Bayesian Modeling

Lecturer: Jacob Steinhardt

8.1 Review: Regression and Decision Theory

In the last lecture, we saw two identification theorems and discussed the ways they help us understand what regression is doing and the conditions under which it will work well. First, we discussed the Gauss-Markov theorem, through which we saw that linear regression (specifically, ordinary least squares) is robust to zero-mean errors in the output, y, but not necessarily to errors in the covariates, x. In fact, noise in x leads to implicit regularization that is closely related to ridge regression. Next, we looked at logistic regression, and saw that it matches the observed $\mathbb{E}[y\phi(x)]$ to the predicted expectation $\mathbb{E}[y\phi(x)]$.

How does regression fit into the decision theory framework from our first three lectures? The data are $(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)})$. There are two different ways to define the decision rule δ depending on what it is we care about. If the parameters themselves are of interest, then the true reality is β^* and our decision rule is $\hat{\beta}$, which we can think of as a function of all the data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$,

$$\hat{\beta} = \operatorname*{argmax}_{\beta} \frac{1}{n} \sum_{i=1}^{n} f(\beta, x^{(i)}, y^{(i)}),$$

where $f(\beta, x^{(i)}, y^{(i)}) = \langle \beta, y^{(i)}\phi(x^{(i)})\rangle - \log(1 + \exp(\langle \beta, \phi(x^{(i)})\rangle))$ for logistic regression or $f(\beta, x^{(i)}, y^{(i)}) = (y^{(i)} - \langle \beta, \phi(x^{(i)})\rangle)^2$ for linear regression. However, if we care about the model's predictions, instead of taking δ to be the training procedure, we can think of δ as being our fitted model. Specifically, the observation is some new data point $x^{(\text{new})}$ that we did not see during training, the true reality is $y^{(\text{new})}$, and the decision rule is

$$\delta(x) = \begin{cases} 1 & \beta^\top \phi(x) > 0 \\ 0 & \text{otherwise} \end{cases}$$

for logistic regression, or $\delta(x) = \beta^\top \phi(x)$ for linear regression.

It is often useful to be able to reason about how the decision rule changes when we modify the parameters β. For example, in a logistic regression setting, we may want to think about how to modify β to move predictions towards different outcomes in a fairness setting.

Example 8.1. Suppose we have a logistic regression model with parameters β, so our decision rule is

$$\delta(x) = \begin{cases} 1 & \beta^\top \phi(x) > 0 \\ 0 & \text{otherwise} \end{cases}.$$

What happens if we double β? In this case, all that matters for determining the value of $\delta(x)$ is the sign of the dot product $\beta^\top \phi(x)$, so the decision rule will stay the same. If, instead, we were to replace β with $-\beta$, the decision rule is flipped.

8.2 Bayesian Modeling

Today, we will consider a new approach: Bayesian modeling. Throughout the lecture, we will return to the following motivating example.

Example 8.2. Suppose you are looking at two products on Amazon and want to decide which of the two to buy. Product 1 has 5 positive reviews and no negative reviews, while Product 2 has 14 positive reviews and 1 negative review. Which one of these two products should we feel better about buying? In the first case, there is more uncertainty since we saw fewer reviews. In the second, we saw one bad review, but there may be less uncertainty. Which of the two products you prefer also depends on your own prior beliefs about the world. If you believe most products are good and it is difficult to get negative reviews, you may prefer Product 1. On the other hand, if you believe it would not be too difficult to get 5 good reviews just by chance, you may feel more confident purchasing Product 2.

In decision making problems, we often want to consider both uncertainty and prior knowledge about the world when making decisions. Bayesian modeling incorporates both of these ideas, and provides a flexible way of incorporating prior beliefs into our statistical modeling.

8.3 Maximum Likelihood Estimation

For comparison, we will first consider one simple frequentist estimator which arises from considering the likelihood function over all our data: $\mathcal{L}(\theta) = \mathbb{P}(x_1, x_2, \ldots, x_n; \theta)$.[1] We can think of the likelihood as the probability that the data we observed was generated by the distribution that has its parameter set to θ. One natural thing to do would be to find the value of θ 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 define a likelihood model $\mathbb{P}(x; \theta)$ where x 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 and must instead approximate it to the best of our ability.
2. We collect our data $x_1, \ldots, x_n$.

[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.

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

   $$\mathcal{L}(\theta) = \mathbb{P}(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} \mathbb{P}(x_i; \theta),$$

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

   (b) Take the log of the likelihood

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

   This step isn't compulsory, but it usually significantly simplifies the next step by changing products into sums.

   (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.

Example 8.2 continued. Let $\theta \in [0, 1]$ be the probability of a positive review. Further let x_i be the i^{th} customer review, where $x_i = 1$ indicates a positive review and $x_i = 0$ indicates a negative review. In this case,

$$\mathbb{P}(x_i; \theta) = \theta^{x_i}(1 - \theta)^{1 - x_i}$$

and the likelihood function for Product 1 is

$$\mathbb{P}(x_1, \ldots, x_5; \theta) = \mathbb{P}(x_1; \theta) \cdots \mathbb{P}(x_5; \theta) = \theta^{x_1 + \cdots + x_5}(1 - \theta)^{5 - (x_1 + \cdots + x_5)}.$$

Plugging in our data gives the following result for each product:

- Product 1 has likelihood θ^5, so $\hat{\theta}_{MLE} = 1$.
- Product 2 has likelihood $\theta^1 4(1 - \theta)$, so $\hat{\theta}_{MLE} = \frac{14}{15}$.

Based on the MLE we might argue that we should always pick the first product, which may not be a satisfying solution to most people. We will see that Bayesian modeling might yield a different answer, depending on how we incorporate our prior beliefs.

8.4 Bayesian Estimation

Unlike frequentist estimation, in Bayesian estimation we treat the unknown parameter as a random variable instead of a fixed (but unknown) quantity. Here, the primary quantity of interest will no longer be the likelihood $\mathbb{P}(x_1, x_2, \ldots, x_n|\theta)$, but instead the posterior distribution $\mathbb{P}(\theta|x_1, x_2, \ldots, x_n)$. The posterior can be thought of as our belief about the value of our parameter θ after having observed our data.

To set up a Bayesian estimation problem, we need to define a likelihood model $\mathbb{P}(x|\theta)$ where θ is the unknown parameter and x is an observed datapoint. We also need to define a prior distribution $\mathbb{P}(\theta)$ on our unknown parameter θ. As in the frequentist setting, there is a lot of subjectivity in these first two steps: we almost never know the true prior nor the true likelihood, but we hope to make an informed decision. It is also worth noting that our prior and our likelihood may have additional hyperparameters that we set, which can also be considered part of the subjective task of picking the right model. Once we have these components and have collected our data, we can write down the posterior distribution using Bayes' theorem:

$$\mathbb{P}(\theta|x_1, x_2, \ldots x_n) = \frac{\mathbb{P}(x_1, x_2, \ldots, x_n|\theta)\mathbb{P}(\theta)}{\mathbb{P}(x_1, x_2, \ldots, x_n)} = \frac{\prod_{i=1}^n \mathbb{P}(x_i|\theta)\mathbb{P}(\theta)}{\mathbb{P}(x_1, x_2, \ldots, x_n)}.$$

Notice that $\mathbb{P}(x_1, x_2, \ldots, x_n)$ is just a positive constant with respect to θ; in fact, it must be the unique normalizing constant that makes the posterior distribution $\mathbb{P}(\theta|x_1, x_2, \ldots x_n)$ integrate to one (see Figure 8.1). Thus, the term $\prod_{i=1}^n \mathbb{P}(x_i|\theta)\mathbb{P}(\theta)$ determines the shape of the posterior distribution, and we often write

$$\mathbb{P}(\theta|x_1, x_2, \ldots x_n) \propto_\theta \prod_{i=1}^n \mathbb{P}(x_i|\theta)\mathbb{P}(\theta),$$

where $\propto$ is the symbol for "proportional to," and $\propto_\theta$ indicates proportionality with respect to the variable θ.

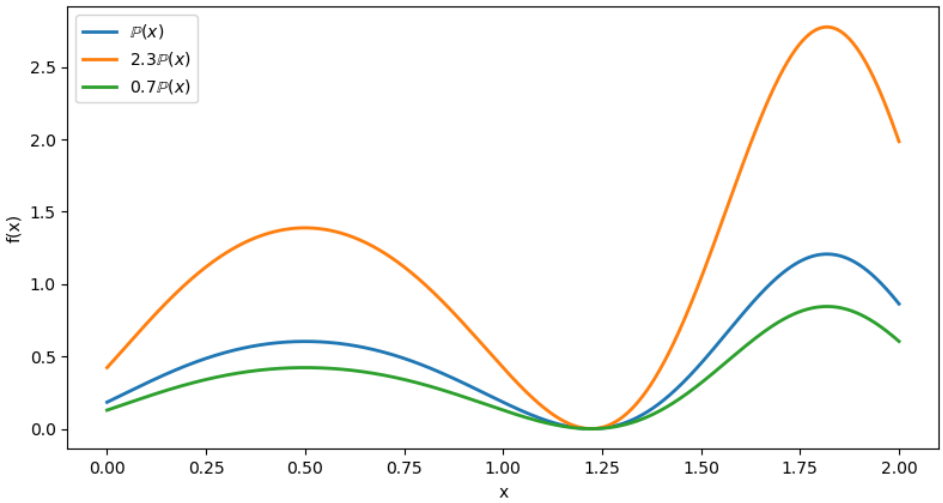

Figure 8.1: Different scalings of some distribution $\mathbb{P}$(x). Even though only one of these curves integrate to 1 we can see that they all uniquely identify the same distribution since they all have the same shape.

Example 8.3. As an example of computing the posterior distribution, consider the situation where we have $x \sim \text{Binom}(n, \theta)$ where n fixed, and $\theta \sim \text{Beta}(r, s)$ for fixed and known $r > 0$ and $s > 0$. Recalling the form of the Beta distribution, our prior is

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

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. We can express the likelihood as

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

To find the posterior distribution of θ, note that

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

is simply an unnormalized version of $\text{Beta}(x+r, n-x+s)$. So, our posterior belongs to the same family of distributions as our prior and can simply be computed by adding the number of observed successes and failures in the Binomial trial to the prior hyperparameters r and s.

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*. The Beta-Binomial of Example 8.3 is one well-known example of conjugate distributions.

Example 8.2 continued. Suppose our prior is $\theta \sim \text{Beta}(1,1)$. Note that $x_i \sim \text{Bernoulli}(\theta) = \text{Binom}(1,\theta)$. Thus, this is a special case of the Beta-Binomial example above. For example,

$$\mathbb{P}(\theta|x_1,\ldots,x_5) \propto_\theta \mathbb{P}(\theta)\mathbb{P}(x_1,\ldots,x_5|\theta) = \theta^5 = \text{Beta}(6,1).$$

Conjugate distribution also exist for continuous random variables – we will look at one such example next.

Example 8.4. Suppose we have $x \sim \mathcal{N}(\theta,\sigma^2)$ where σ^2 is fixed and known, $\theta \sim \mathcal{N}(\mu_\theta,\sigma^2)$, and μ_θ is fixed and known. Then, recalling the Normal density, the prior on θ is given by

$$\mathbb{P}(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(\theta-\mu_\theta)^2}{2\sigma^2}\right).$$

The likelihood is given by

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

Again, we can simplify our computation of the posterior by noting we only need to consider terms that depend on θ, since any terms that are constants or depend only on x are part of the normalizing constant. Computing the posterior gives us

$$\begin{aligned}
\mathbb{P}(\theta|x) &\propto_\theta \exp\left(-\frac{(x-\theta)^2+(\theta-\mu_\theta)^2}{2\sigma^2}\right)\\
&= \exp\left(-\frac{2(\theta^2-\theta x-\theta\mu_\theta)+x^2+\mu_\theta^2}{2\sigma^2}\right)\\
&= \exp\left(-\frac{2(\theta^2-\theta x-\theta\mu_\theta)}{2\sigma^2}\right)\exp\left(-\frac{x^2+\mu_\theta^2}{2\sigma^2}\right)\\
&\propto_\theta \exp\left(-\frac{\theta^2-\theta(x+\mu_\theta)}{2\sigma^2/2}\right)\\
&= \exp\left(-\frac{\theta^2-\theta(x+\mu_\theta)+(x+\mu_\theta)^2/4-(x+\mu_\theta)^2/4}{2\sigma^2/2}\right)\\
&\propto_\theta \exp\left(-\frac{\theta^2-\theta(x+\mu_\theta)+(x+\mu_\theta)^2/4}{2\sigma^2/2}\right)\\
&= \exp\left(-\frac{(\theta-(x+\mu_\theta)/2)^2}{2\sigma^2/2}\right).
\end{aligned}$$

This shows that the posterior distribution is also Gaussian of the form $\mathcal{N}(\frac{x+\mu_\theta}{2},\frac{\sigma^2}{2})$, so we once again have a conjugate prior. This result is quite intuitive: after observing the data our variance went down, and the posterior mean takes into account both the data we observed and our prior belief about the location of the mean.

In the previous few examples, we chose priors that made the math required to compute the posterior work out nicely. Unfortunately, not all priors are conjugate priors, and not all model choices will allow us to compute a nice, closed form for the posterior distribution. In future lectures, we will talk about algorithms we can use to compute complicated functions which will help us do Bayesian estimation without closed-form posteriors. For now, we will simply look at an example of a situation where our priors are not conjugate.

Example 8.5. Assume $x \sim \mathcal{N}(\theta, \sigma^2)$ where σ^2 is fixed and known, and the pdf of θ is given by

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

Computing the posterior distribution gives us

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

Here, we see that our posterior doesn't correspond to any distribution that we know of.

Once we have a posterior distribution for our parameter θ, how do we actually estimate its value? There are multiple ways to summarize the posterior distribution $\mathbb{P}(\theta|x)$. One natural choice is the maximum a posteriori (MAP) estimate,

$$\hat{\theta}_{MAP} = \operatorname*{argmax}_\theta \mathbb{P}(\theta|x),$$

which is the value the maximizes the posterior probability. We can relate the problem of maximizing the posterior to maximizing the likelihood by Bayes' rule and taking logarithms:

$$\begin{aligned} \operatorname*{argmax}_\theta \mathbb{P}(\theta|x) &= \operatorname*{argmax}_\theta \mathbb{P}(\theta)\mathbb{P}(x|\theta) \\ &= \operatorname*{argmax}_\theta \log \mathbb{P}(\theta) + \log \mathbb{P}(x|\theta). \end{aligned}$$

Writing the problem this way exposes some structure: it is similar to an optimization problem with a regularizer ($\log \mathbb{P}(theta)$) and a loss ($\log \mathbb{P}(x|\theta)$). It is an interesting exercise to convince yourself that ℓ_2 regularization is equivalent to having a Gaussian prior.

MAP may not always be the choice for estimating θ. For example, suppose you have a posterior distribution with two modes. The MAP estimate may then lie in one very peaky mode of the distribution and not take into account more typical points. Thus, MAP is not always a good summary of the entire posterior. We could also use the posterior mean, posterior median, or posterior mode as an estimator of θ. The posterior mean, also called the least mean-square error (LMSE) estimator, is a particularly common choice.

8.5 Gaussian Mixture Models

We now turn to a particularly ubiquitous Bayesian model, the Gaussian mixture model (GMM). We will explore this model briefly here by way of an illustrative example, and will consider it in detail in the next lecture.

Example 8.6. Assume we have a dataset of i.i.d heights $x_1, x_2, \ldots, x_n$. We could model the distribution of heights as a simple Gaussian, however we know that an individual's sex 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 θ_i. Where $\theta_i = 0$ if the i^{th} participant is female and $\theta_i = 1$ if they are male. These types of unseen variables are called *latent variables*. We can put a Bernoulli prior on the sex θ of any participant as

$$\begin{aligned}\pi_0 &= \mathbb{P}(\theta = 0),\\ \pi_1 &= \mathbb{P}(\theta = 1).\end{aligned}$$

Then, the likelihood of any datapoint x is given by

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

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

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

An example of such a GMM distribution is shown in Figure 8.2.

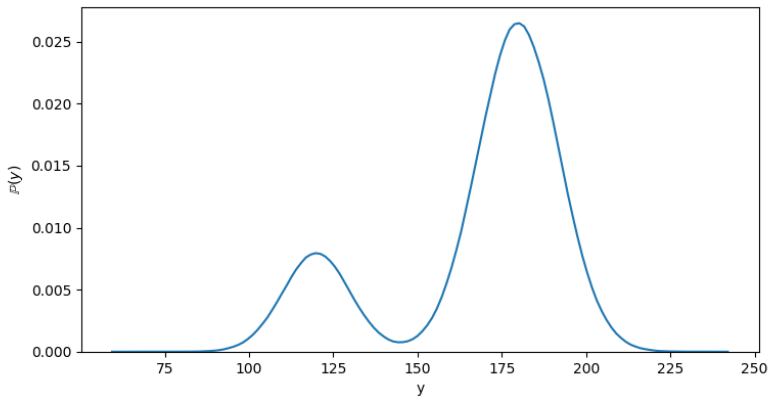

Figure 8.2: An example GMM where $\pi_0 = 0.2$, $\pi_1 = 0.8$, $\mu_0 = 120$, $\mu_1 = 180$, and $\sigma = 11$.

Often, there is hidden structure in our data and if we explicitly model the hidden structure, the remaining signal "look nice;" the GMM is one such example where the remaining signal is Gaussian once we model the right latent variables. In general, Hierarchical Bayesian Modeling, which we will discuss in the next lecture, tried to explicitly capture such hidden structure.