Recitation 10

1 Topic One: Bayesian Regression

Today, I will go over Bayesian regression in a slightly less mathematically intensive way than in the lectures.

1.1 Definitions

A Bayesian model has four inputs:

1. Data: x
2. Parametric Model: A *parametric model* is the assumed joint distribution of the data, given a parameter value:

$$f(x|\theta), \quad \theta \in \Theta.$$

3. Prior Distribution: The *prior distribution* is the distribution that we think the true parameter value might come from. From knowledge prior to seeing the data! This distribution is denoted $\pi(\theta)$.
4. Loss Function: The *Loss Function* is a measure of how poorly our decision rule/estimate is doing. It is small if our estimate is good, and large if our estimate is bad. In this class, and in almost all applications in the real world, we use the loss quadratic loss function:

$$L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)'(\hat{\theta} - \theta)$$

The quadratic loss function is so widely used, that we can almost forgot about the loss function input and assume we are using the quadratic. **I do so for the rest of this recitation.**

1.2 The posterior

The steps to Bayesian Analysis are:

1. Assume a model $f(x|\theta)$ and a prior $\pi(\theta)$.
2. Calculate the posterior distribution $\pi(\theta|x)$.
3. Generate point estimates of parameters by taking the expectation of the posterior distribution:

$$\hat{\theta}(x)^* = E_{\theta|x}[\theta|x] = \int \theta\pi(\theta|x)d\theta$$

   Note: in the general case, we define our bayesian point estimate as the value that minimizes expected loss over the prior distribution.

$$\hat{\theta}(x)^* = min_{\hat{\theta}} \int L(\hat{\theta}, \theta)\pi(\theta|x)d\theta$$

   , but this is equal to the above when we use the quadratic loss function.

4. Generate the $1-\alpha\%$ credibility set of θ by finding the $\frac{\alpha}{2}th$ and $(1-\frac{\alpha}{2})th$ quantile of the posterior, $\pi(\theta|x)$

Note that the $1-\alpha\%$ *credibility set* is defined as the set of values that θ has a $1-\alpha\%$ probability of being drawn from. For our purposes, you can think about this as analagous to the $1-\alpha\%$ confidence set, which is a set of values that contains the true parameter value $1-\alpha\%$ of the time. Note that comparing these two things is very difficult, because they are using two different mental frameworks! Dont worry about it too much if you are very confused!

1.3 Bayesian Regression

The goal of the Bayesian regression is to attain estimates of the beta coefficients in a linear model. Keep this in mind!

First, to translate the above to the bayesian regression, note that $x = \{y, X\}$ and $\theta = \{\beta, \sigma^2\}$.

Step One: the inputs to the model are:

1. Data: y, X

2. Parametric Model: $f(y, X|\beta, \sigma^2)$

$$Y_i = \beta' X_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2) \text{ iid}$$

$$\Rightarrow Y_i \sim N(\beta' X_i, \sigma^2) \text{ iid}$$

3. Prior Distribution: For the Bayesian regression, we split the prior into two parts: $\pi(\beta, \sigma^2) = \pi(\beta|\sigma^2)\pi(\sigma^2)$
   Since we only care about estimating β, not σ^2 and it turns out that $\pi(\sigma^2)$ is not important to estimating β, we only need to assume a form of $\pi(\beta|\sigma^2)$. The prior we assume is:
   $$\beta \sim N(m, V)$$
   where m and V are called *hyperparameters*, which are simply inputs into the prior.

4. Loss Function: quadratic.

   Step Two: to find the posterior distribution of β, σ^2, $\pi(\beta, \sigma^2|y, X)$, we use Bayes' theorem:
   $$f(w|z) = \frac{f(w, z)}{f(z)} = \frac{f(w, z)}{\int f(w, z) dw}$$

   $$\begin{aligned} \pi(\beta, \sigma^2|y, X) &= \frac{f(\beta, \sigma^2, y, X)}{\int f(\beta, \sigma^2, y, X) d\beta, \sigma^2} \\ &= \frac{f(y, X|\beta, \sigma^2)\pi(\beta, \sigma^2)}{\int f(y, X|\beta, \sigma^2)\pi(\beta, \sigma^2) d\beta, \sigma^2} \end{aligned}$$

   Finally, to show that the $\pi(\sigma^2)$ indeed does not matter if we only care about estimating a point estimate of β:
   $$\pi(\beta|y, X, \sigma^2) = \frac{f(y, X|\beta, \sigma^2)\pi(\beta|\sigma^2)}{\int f(y, X|\beta, \sigma^2)\pi(\beta|\sigma^2) d\beta, \sigma^2}$$

   Ok, well what is this $\pi(\beta|y, X, \sigma^2)$? We did all the algebra last class, but the big picture is:

$$\beta|\sigma^2, y, X \sim \mathcal{N}_k\left(\left(\frac{1}{T}V^{-1} + \frac{X'X}{T}\right)^{-1}\frac{(X'y + V^{-1}m)}{T}\,,\ \frac{\sigma^2}{T}\left(\frac{1}{T}V^{-1} + \frac{X'X}{T}\right)^{-1}\right)$$

Note that $X'X$ and $X'y$ grow as T increases, but V^{-1}, m, and σ^2 do not.

Step Three:

Recall that our Bayesian point estimate is simply the mean of the posterior distribution.

$$\hat{\beta}^*_{Bayes} = E[\beta|y, X] = \left(\frac{1}{T}V^{-1} + \frac{X'X}{T}\right)^{-1} \frac{(X'y + V^{-1}m)}{T}$$

Things to remember about this estimate:

1. It is, in a sense, a weighted average of our prior and the OLS estimate.

2. It is biased, but consistent.

3.

$$\hat{\beta}^*_{Bayes} - \hat{\beta}_{OLS} \to 0$$

, as $T \to \infty$, since

$$\frac{1}{T}V^{-1} \Rightarrow 0$$

Step Four:

In order to find the credibility set, note that we need to know ^{2}, so we cannot estimate it only using the conditional priors.

1.4 General Methods

In this very special model, we were able to solve for an analytical solution of the point estimate. However, in most Bayesian analysis cases, we are not able to do so. In these cases, how would you generate a point estimate of a given parameter $\hat{\theta_1}$? How would you estimate the credibility set for that parameter? By simulation from the posterior!

Once we have a posterior, $\pi(\theta|x)$, we can do the following:

(a) Take S draws of from $\pi(\theta|x)$. Record draw of the parameter you care about: $\theta_1^{(s)}$

(b) To attain a Bayesian point estimate, take the average of these draws:

$$\hat{\theta}_{1Bayes} = \frac{1}{S}\sum \theta_1^{(s)}$$

(c) To attain the credibility set, order the $\theta_1^{(s)}$ and take the $\frac{\alpha}{2}$th largest to be the lower bound and the $1-\frac{\alpha}{2}$th largest to be the upper bound of your credibility set.

1.5 A few general results

For a wide class of models, $\pi(\theta|x)$ is asymptotically equal to the pdf of of a normal distribution: $N(\hat{\theta}_{ML}, Var(\hat{\theta}_{ML}))$