

Bayesian Linear Regression

Ryan Wang

May 2019

We consider Bayesian methods for multiple linear regression problem

$$Y = X\beta + \epsilon \quad (1)$$

Where $Y \in \mathbf{R}^n$, $X \in \mathbf{R}^{n \times d}$, $\beta \in \mathbf{R}^d$, and $\epsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$.

1 Bayesian Statistics

Before we actually start to perform the Bayesian approach to the problem, let's do a review on the idea of Bayesian statistics. Bayesian statistics is built upon the Bayes' theorem of probability:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (2)$$

Suppose we have the following statistical model:

$$x_1, \dots, x_n \text{ i.i.d.}, x_i \sim p(x|\theta) \quad (3)$$

Instead using a point estimation (e.g. M.L.E.), we firstly assume that θ follows a probability distribution $p(\theta|\alpha)$, which is called the prior distribution.

The idea is to consider how the observations help us to update our prior belief on θ , which can be considered as $p(\theta|x, \alpha)$. We are going to use the Bayes' theorem to do that:

$$p(\theta|x, \alpha) = \frac{p(x|\theta, \alpha)p(\theta, \alpha)}{p(x, \alpha)} = \frac{p(x|\theta, \alpha)p(\theta|\alpha)p(\alpha)}{p(x|\alpha)p(\alpha)} = \frac{p(x|\theta, \alpha)p(\theta|\alpha)}{p(x|\alpha)} \quad (4)$$

Here, $p(x|\theta, \alpha)$ is the likelihood function of data. Notice that $p(x|\alpha)$ is a constant term with respect to the distribution of θ . Thus, we can conclude that:

$$\text{posterior} \propto \text{likelihood} * \text{prior} \quad (5)$$

likelihood * prior will be the kernel of the posterior distribution, thus it's sufficient for us to determine the p.d.f. of the posterior distribution.

2 Updating β : Conjugate Multivariate Normal

Let's look back to the multiple linear regression problem $Y = X\beta + \epsilon$. We know that the observation $Y \in \mathbf{R}^n$ follows a multivariate normal $\mathcal{N}(X\beta, \sigma^2 I_n)$. The fact is that multivariate normal is a conjugate prior of itself, which we will see in the following derivation.

Because of the conjugate property of multivariate normal, we let the prior be $\mathcal{N}(\mu_0, \Lambda_0)$:

$$p(\theta) \propto \exp\left(-\frac{1}{2}(\theta - \mu_0)^T \Lambda_0^{-1}(\theta - \mu_0)\right) \quad (6)$$

Since $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$, we write the likelihood function of Y as:

$$p(Y|\theta, \alpha) \propto \exp\left(-\frac{1}{2\sigma^2}(Y - X\theta)^T(Y - X\theta)\right) \quad (7)$$

Therefore

$$\begin{aligned} p(\theta|Y) &\propto (6) * (7) \\ &\propto \exp\left(-\frac{1}{2}\left(-\frac{2}{\sigma^2}Y^T X\theta + \frac{1}{\sigma^2}\theta^T X^T X\theta\right) + \theta\Lambda_0^{-1}\theta - 2\mu_0^T \Lambda_0^{-1}\theta\right) \end{aligned} \quad (8)$$

Which tells us that

$$\theta|Y \sim \mathcal{N}(\mu, \Lambda) \quad (9)$$

Where $\Lambda = \left(\frac{X^T X}{\sigma^2} + \Lambda_0^{-1}\right)^{-1}$, and $\mu = \Lambda\left(\frac{X^T Y}{\sigma^2} + \Lambda_0^{-1}\mu_0\right)$.

(It has been shown above that multivariate normal is a conjugate prior to itself.)