Bayesian Linear Regression

Ce Wang

May 2019

We consider Bayesian methods for multiple linear regression problem

$$Y = X\beta + \epsilon \tag{1}$$

Where $Y \in \mathbf{R}^n$, $X \in \mathbf{R}^{n*d}$,$\beta \in \mathbf{R}^d$, 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:

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

Suppose we have the following statistical model:

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

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:

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

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:

$$posterior \propto likelihood * prior \tag{5}$$

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 with 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 $\mathcal{N}(\mu_0, \Lambda_0)$:

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

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

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

Therefore

$$\begin{aligned} p(\theta|Y) &\propto (6) * (7) \\ &\propto -\frac{1}{2}(\frac{1}{{\sigma_n}^2}Y^TY - \frac{2}{{\sigma_n}^2}(XY)^T\theta + \frac{1}{{\sigma_n}^2}\theta^T XX^T\theta + \frac{1}{{\sigma_n}^2}\theta^T XX^T\theta \\ &+ \theta^T {\Lambda_0}^{-1}\theta - 2{\mu_0}^T{\Lambda_0}^{-1}\theta + {\mu_0}^T{\Lambda_0}^{-1}\mu_0) \\ &\propto -\frac{1}{2}(-\frac{2}{{\sigma_n}^2}(XY)^T\theta + \frac{1}{{\sigma_n}^2}\theta^T XX^T\theta + \theta^T{\Lambda_0}^{-1}\theta - 2\mu_0{\Lambda_0}^{-1}\theta) \end{aligned} \tag{8}$$

Let $\Lambda = (\frac{X^TX}{\sigma^2} + {\Lambda_0}^{-1})^{-1}$, and $\mu = \Lambda(\frac{X^TY}{\sigma^2} + {\Lambda_0}^{-1}\mu_0)$, we have

$$P(\theta|Y) \propto -\frac{1}{2}(\theta^T\Lambda^{-1}\theta - 2\mu^T\Lambda^{-1}\theta) \tag{9}$$

By completing the square, we know that $P(\theta|Y) \sim \mathcal{N}(\mu, \Lambda)$.