Bayesian Linear Regression Notes

Nicholas Hoell

May 10, 2021

1 Linear Regression

1.1 The Basic Formulation

Consider the simple multivariate linear regression problem where in data of the form $(\mathbf{x}_i, y_i)_{i=1}^m$ are drawn from an unknown data-generating distribution $\mathcal{D}$ and we wish to find a tidy linear estimator for the y_i's as responses to the input $\mathbf{x}_i$'s. In keeping with standard convention assume that $\mathbf{x}_i = (1, \tilde{\mathbf{x}}_i) \in \mathbb{R}^d$ holds for all i.

We then assume that the observables $(\mathbf{x}_i, y_i)_{i=1}^m$ satisfy

$$y_i = \mathbf{x}_i \cdot \theta + \epsilon_i \tag{1}$$

Inspired by the classical central limit theorem (saying, very roughly, that independent errors accumulate as Gaussians) we propose that the residuals terms are distributed according to a normal distribution,

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

each with known variance σ^2. Assume further that the errors are drawn independent of each other and of the data.

1.2 The Random Variable Viewpoint

The above outline is the standard framework for multivariate linear regression.

Consider now that the above scenario is simply the multiple *realizations* of an equation involving random variables $(X, Y) \sim \mathcal{D}$, i.e.

$$Y = X \cdot \theta + \epsilon \tag{2}$$

It is as though this random variable equation is being sampled m times to produce the simultaneous equations (**??**). Supposing that the latent variables θ were fixed, (**??**) tells us that

$$\begin{aligned} p(y \mid \mathbf{x}, \theta) &= p(\mathbf{x} \cdot \theta + \epsilon \mid \theta) \\ &\propto p(\epsilon) \\ &= \mathcal{N}(0, \sigma^2)(\epsilon) \\ &= \mathcal{N}(0, \sigma^2)(y - \mathbf{x} \cdot \theta) \\ &\propto e^{-\frac{|y - \mathbf{x} \cdot \theta|^2}{2\sigma^2}} \end{aligned}$$

where, in the above, $p(z) = \mathbb{P}(Z = z)$ etc. Recall that given a probability density function $f(x \mid \theta)$ parameterized by parameters θ the **likelihood** function is defined as $L(\theta) \doteq f(x \mid \theta)$. Namely, a likelihood function of a data set is the probability of obtaining that data set under a given choice of parameters. For our regression problem, therefore, we have

$$\begin{aligned} L(\theta) &= p((y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), ..., (y_n, \mathbf{x}_n) \mid \theta) \\ &= \prod_{i=1}^{m} p((y_i, \mathbf{x}_i) \mid \theta) \\ &\propto \prod_{i=1}^{m} e^{-\frac{|y_i - \mathbf{x}_i \cdot \theta|^2}{2\sigma^2}} \\ &= e^{-\frac{||\mathbf{y} - \mathbf{X} \cdot \theta||^2}{2\sigma^2}} \end{aligned}$$

In the above, $\mathbf{y} = (y_1, ..., y_m)$ and $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_m^T \end{bmatrix}$.

The maximum likelihood principle suggests selecting parameters which make the data most likely to obtain. So in the case we select

$$\begin{aligned} \theta^* &= \arg\max_{\theta} L(\theta) \\ &= \arg\max_{\theta} \log L(\theta) \\ &= \arg\max_{\theta} -\frac{||\mathbf{y} - \mathbf{X} \cdot \theta||^2}{2\sigma^2} \\ &= \arg\min_{\theta} \frac{||\mathbf{y} - \mathbf{X} \cdot \theta||^2}{2\sigma^2} \end{aligned}$$

Solving $\arg\min_\theta \frac{||\mathbf{y}-\mathbf{X}\cdot\theta||^2}{2\sigma^2}$ will yield the standard, familar, linear regression equations.

1.3 Including Priors

Suppose that we spice up the previous scenario by not assuming that the latent parameters θ are static, but rather that they are drawn themselves from a distribution. For simplicity, assume that $\theta_i \sim \mathcal{N}(0, \tau^2)$ with samples being iid and independent of ϵ and of the data. We then have the fully randomized model

$$Y = X \cdot \Theta + \epsilon$$

Performing a similar calculation as before, we have

$$\begin{aligned}
p(y \mid \mathbf{x}, \theta, \epsilon) &= p(\mathbf{x}\cdot\theta + \epsilon \mid \theta, \epsilon) \\
&= p(\mathbf{x}\cdot\theta + \epsilon \mid \theta)p(\mathbf{x}\cdot\theta + \epsilon \mid \epsilon) \\
&\propto p(\epsilon)p(\mathbf{x}\cdot\theta) \\
&\propto p(\epsilon)p(\theta) \\
&= \mathcal{N}(0,\sigma^2)(\epsilon)\mathcal{N}(0,\tau^2)(\theta) \\
&= \mathcal{N}(0,\sigma^2)(y-\mathbf{x}\cdot\theta)\mathcal{N}(0,\tau^2)(\theta) \\
&\propto e^{-\frac{|y-\mathbf{x}\cdot\theta|^2}{2\sigma^2}} e^{-\frac{\theta^2}{2\tau^2}}
\end{aligned}$$

Putting these together, we have

$$\begin{aligned}
L(\theta) &= p((y_1,\mathbf{x}_1),(y_2,\mathbf{x}_2),....,(y_n,\mathbf{x}_n) \mid \theta) \\
&= \prod_{i=1}^{m} p((y_i,\mathbf{x}_i) \mid \theta) \\
&\propto \prod_{i=1}^{m} e^{-\frac{|y_i-\mathbf{x}_i\cdot\theta|^2}{2\sigma^2}} e^{-\frac{\theta^2}{2\tau^2}} \\
&= e^{-\frac{||\mathbf{y}-\mathbf{X}\cdot\theta||^2}{2\sigma^2} - \frac{\theta^2}{2\tau^2}}
\end{aligned}$$

from which we see that the maximum likelihood estimator solution solves

$$\theta^* = \arg\min_\theta ||\mathbf{y} + \mathbf{X}\cdot\theta||^2 + \frac{\sigma}{\tau}||\theta||^2 \tag{3}$$

which the literature refers to as the so-called *ridge regression* solution. A few comments about (**??**) are in order.

1. The hyperparameter $\lambda = \frac{\sigma}{\tau}$ controls the size of the so-called penalty term in the optimization. It controls how much "fidelity" we are willing to trade for taming the parameters.

2. The larger σ is, the larger λ is and thus, the more uncertainty we have in our model, the more we should enforce tightening of the coefficients and vice versa.

3. The larger τ is the larger λ is and thus, the more uncertainty we have on what to expect of our coefficients, the more we should enforce tightening of them and vice versa.

4. Prior assumptions on the latent variables *and thus prior assumptions on the data generating distribution* cause the optimization problem to take on different form, resulting in a different final regressor function.