Bayesian Least Squares Regression

Donald E. Brown

School of Data Science
University of Virginia
Charlottesville, VA 22904

Bayesian Formulation

- Regression with response $\mathbf{y}$, data $\mathbf{X}$, and parameter $\boldsymbol{\theta}$:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \epsilon$$

- The parameters are random variables, so $\boldsymbol{\theta}$ is a random vector
- Given the prior is $p(\boldsymbol{\theta})$ and data $\{\mathbf{X}, \mathbf{y}\}$, then the posterior is

$$p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{X}, \mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{X}, \mathbf{y})}$$

- $\mathbf{X}$ is known (i.e., not random) & $\mathbf{y}$ is a function of $\mathbf{X}$ made random by ϵ. So will simply write $p(\boldsymbol{\theta}|\mathbf{y})$
- Next estimate $\boldsymbol{\theta}$

Maximum A Posteriori (MAP) Estimate

- Find the estimate that maximizes the posterior distribution for $\boldsymbol{\theta}$
- Conjugate prior approach:

$$p(\boldsymbol{\theta}) \sim N(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_\theta)$$
$$f(\mathbf{y}|\boldsymbol{\theta}, \sigma_\epsilon^2) \sim N(\mathbf{X}\boldsymbol{\theta}, \sigma_\epsilon^2 \mathbf{I})$$

- The conditional posterior is Gaussian, and the marginal is a t distribution; for both, the mean is the max

$$E[\boldsymbol{\theta}|\mathbf{y}] =$$

$$\boldsymbol{\theta}_0 + \left(\boldsymbol{\Sigma}_\theta^{-1} + \mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T \left(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}_0\right)$$

Bayesian Formulation as Regularization

- Prior acts to regularize the resulting estimate
- Suppose $\boldsymbol{\Sigma}_\theta = \sigma_\theta^2 \mathbf{I}$ and $\boldsymbol{\theta}_0 = \mathbf{0}$

$$\boldsymbol{\theta}_{MAP} = (\lambda \mathbf{I} + \mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T y$$

where, $\lambda = \sigma_\epsilon^2 / \sigma_\theta^2$

- This is ridge regression
- The ridge regression solution is

$$\boldsymbol{\theta}_{Ridge} = (\lambda \mathbf{I} + \mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T y$$

Regularization Parameter

- Choosing $\lambda = \sigma_\epsilon^2 / \sigma_\theta^2$ determines the performance of the regularization
- Typically use grid search with cross-validation
- Bayesian formulation provides the basis for improved understanding through the posterior, $p(\boldsymbol{\theta}|\mathbf{y})$
- Instead of just the maximum, we may want all of $p(\boldsymbol{\theta}|\mathbf{y})$, which means we also want the Bayesian denominator

$$f(\mathbf{y}) = \int f(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta})$$

Conjugate Approach

- Again, the conditional prior and likelihood are Gaussian $p(\boldsymbol{\theta}) \sim N(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_\theta)$ & $f(\mathbf{y}|\boldsymbol{\theta}) \sim N(\mathbf{X}\boldsymbol{\theta}, \sigma_\epsilon^2 \mathbf{I})$
- So, $\mathbf{y}$, $p(\mathbf{y}|\sigma_\epsilon^2) \sim N\left(\mathbf{X}\boldsymbol{\theta}_0, \sigma_\epsilon^2 \mathbf{I} + \mathbf{X}\Sigma_\theta \mathbf{X}^T\right)$
- The conditional posterior is Gaussian, $p(\boldsymbol{\theta}|\mathbf{y}, \sigma_\epsilon^2) \sim N(\boldsymbol{\mu}_{\theta|y}, \boldsymbol{\Sigma}_{\theta|y})$ where

$$\mu_{\theta|y,\sigma_\epsilon^2} = \boldsymbol{\theta}_0 + \frac{1}{\sigma_\epsilon^2}\left(\boldsymbol{\Sigma}_\theta^{-1} + \frac{1}{\sigma_\epsilon^2}\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}_0)$$

$$\boldsymbol{\Sigma}_{\theta|y,\sigma_\epsilon^2} = \left(\boldsymbol{\Sigma}_\theta^{-1} + \frac{1}{\sigma_\epsilon^2}\mathbf{X}^T\mathbf{X}\right)^{-1}$$

- This gives the distribution for the estimate, $\hat{\boldsymbol{\theta}}$

Prediction

- Goal is to predict y given new $\mathbf{x}$
- From the conditional posterior, $p(\boldsymbol{\theta}|\mathbf{y}, \sigma_\epsilon^2)$, obtain

$$p(y|\mathbf{x}, \mathbf{y}, \sigma_\epsilon^{\mathbf{2}}) = \int p(y|\mathbf{x}, \theta, \sigma_\epsilon^{\mathbf{2}}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$$

- From the regression model

$$p(y|\mathbf{x}, \theta, \sigma_\epsilon^{\mathbf{2}}) \sim N(\mathbf{x}\boldsymbol{\theta}, \sigma_\epsilon^2)$$

Distribution of the Prediction

- For $\mathbf{\Sigma}_\theta = \sigma_\theta^2 \mathbf{I}$ the posterior parameters are

$$\mu_{\theta|y,\sigma_\epsilon^2} = \boldsymbol{\theta}_0 + \frac{1}{\sigma_\epsilon^2}\left(\frac{1}{\sigma_\theta^2}\mathbf{I} + \frac{1}{\sigma_\epsilon^2}\mathbf{x}^T\mathbf{x}\right)^{-1}\mathbf{x}^T(\mathbf{y} - \mathbf{x}\boldsymbol{\theta}_0)$$

$$\sigma^2_{\theta|y,\sigma_\epsilon^2} = \left(\frac{1}{\sigma_\theta^2}\mathbf{I} + \frac{1}{\sigma_\epsilon^2}\mathbf{x}^T\mathbf{x}\right)^{-1}$$

- So $p(y|\mathbf{x},\mathbf{y}) \sim N(\mu_y, \sigma_y^2)$

$$\mu_y = \mathbf{x}\mu_{\theta|y,\sigma_\epsilon^2}$$

$$\sigma_y^2 = \sigma_\epsilon^2 + \sigma_\epsilon^2\sigma^2_{\theta|y,\sigma_\epsilon^2}\mathbf{x}\left(\sigma_\epsilon^2\mathbf{I} + \sigma^2_{\theta|y,\sigma_\epsilon^2}\mathbf{x}^T\mathbf{x}\right)^{-1}\mathbf{x}$$