Notes on Bayesian Linear Regression

CS 6957: Probabilistic Modeling

February 11, 2013

Linear Regression Model

We are considering a random variable y as a function of a (typically non-random) vector-valued variable $x \in \mathbb{R}^k$. This is modeled as a linear relationship, with coefficients β_j, plus i.i.d. Gaussian random noise:

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ik}\beta_k + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2).$$

In matrix form, this looks like

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Or, simply

$$y = X\beta + \epsilon.$$

It is common to set the first column of X to a constant column of 1's, so that β_1 is an intercept term.

Ordinary Least-Squares (OLS)

Our goal is to estimate the unknown parameters in β. The maximum-likelihood estimate (MLE) of β is based on the Gaussian likelihood:

$$p(y \mid X, \beta; \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|^2\right).$$

Keep in mind that this is a product of likelihoods for each of the individual components of $y = (y_1, \ldots, y_n)$. Taking the log of this likelihood and then taking the derivative w.r.t. β, we get

$$\nabla_\beta \ln p(y \mid X, \beta; \sigma^2) = -\frac{1}{\sigma^2} X^T(y - X\beta).$$

Setting this derivative equal to zero and solving for β gives the MLE or OLS estimate,

$$\hat{\beta} = (X^T X)^{-1} X^T y,$$

where the inverse here is the Moore-Penrose pseudoinverse (same as the inverse when it exists). The MLE for β is Gaussian distributed and unbiased:

$$\hat{\beta} \sim \mathrm{N}\left(\beta, \sigma^2 (X^T X)^{-1}\right).$$

Bayesian Linear Regression

As seen in the polynomial regression code example (`BayesianLinearRegression.r`), the entries of β can be overinflated for higher-order coefficients, as the model tries to overfit the data with a "wiggly" curve. To counteract this, we may inject our prior belief that these coefficients should not be so large. So, we introduce a conjugate Gaussian prior, $\beta \sim \mathrm{N}(0, \Lambda^{-1})$. Here we are parameterizing the Gaussian using the inverse covariance, or precision matrix, Λ, which will make computations easier. A common choice is $\Lambda = \lambda I$, for a positive scalar parameter λ.

Now, the posterior for β is

$$p(\beta \,|\, y; X, \sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|^2 - \frac{1}{2}\beta^T \Lambda \beta\right).$$

Just as we worked out in the univariate case, the conjugate prior for β results in the posterior also being a multivariate Gaussian. Completing the square inside the exponent, we see that the posterior for β has the following distribution:

$$\beta \sim \mathrm{N}(\mu_n, \Sigma_n),$$

where

$$\mu_n = \left(X^T X + \sigma^2 \Lambda\right)^{-1} X^T y, \tag{1}$$

$$\Sigma_n = \sigma^2 \left(X^T X + \sigma^2 \Lambda\right)^{-1}. \tag{2}$$

Compare this to the MLE, and note what happens when $\Lambda \to 0$ (and think of how this would be the noninformative Jeffreys prior on β in the limit).

Exercise: Complete the square in the posterior to derive the formulas for μ_n and Σ_n.

Posterior Predictive Distribution

Now, say we are given a new independent data point $\tilde{x}$, and we would like to predict the corresponding unseen dependent value, $\tilde{y}$. Remember, the posterior predictive distribution of $\tilde{y}$ is given by,

$$p(\tilde{y} \,|\, y; \tilde{x}, X, \sigma^2, \Lambda) = \int p(\tilde{y} \,|\, \beta; \tilde{x}, \sigma^2)\, p(\beta \,|\, y; X, \sigma^2, \Lambda)\, d\beta.$$

This is now a univariate Gaussian:

$$\tilde{y} \,|\, y \sim \mathrm{N}\left(\tilde{x}^T \mu_n, \sigma_n^2(\tilde{x})\right),$$

where

$$\sigma_n^2(\tilde{x}) = \sigma^2 + \tilde{x}^T \Sigma_n \tilde{x}.$$

The first term on the right is due to the noise (the additive ϵ), and the second term is due to the posterior variance of β, which represents our uncertainty in the parameters.

Exercise: Derive the formulas above for the mean and variance of $p(\tilde{y} \,|\, y)$. You might find this Wikipedia topic useful:
`http://en.wikipedia.org/wiki/Multivariate_normal_distribution#Conditional_distributions`