Linear Regression

Rebecca C. Steorts
Bayesian Methods and Modern Statistics: STA 360/601

Module 12

Setup

Let's assume that $D_i = (x_i, y_i)$ for all i.

Assume

$$Y_i \overset{iid}{\sim} N(w^T x_i, \sigma^2).$$

Assume σ^2 known and $\theta = w$.

What is the MLE?

$$\theta_{MLE} = \arg\max_{\theta \in \Theta} p(D \mid \theta)$$

What is the likelihood? (Want to get to the MLE).

Define $y = (y_1, \ldots, y_n)$. Note that $w^T x_i = x_i^T w$. Define $A = (x_1^T, \ldots, x_n^T)$. (A is often called the design matrix).

$$
\begin{align}
p(D \mid \theta) &= p(y \mid x, \theta) \tag{1} \\
&= \prod_i p(y_i \mid x_i, \theta) \tag{2} \\
&= \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-1/(2\sigma^2)(y_i - w^T x_i)^2\} \tag{3} \\
&= (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp\{-1/(2\sigma^2) \sum_i (y_i - w^T x_i)^2\} \tag{4} \\
&= (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp\{-1/(2\sigma^2)(y - Aw)^T (y - Aw)\} \tag{5}
\end{align}
$$

Goal: minimize

$$(y - Aw)^T (y - Aw)$$

(Think about why we're minimizing).

Goal: minimize

$$(y - Aw)^T (y - Aw)$$

Expand what we have above.

$$g := (y - Aw)^T (y - Aw) = y^T y - 2w^T A^T y + w^T A^T A w$$

Now take the gradient or derivative with respective to w.

$$\frac{\partial g}{\partial w} = -2A^T y + 2A^T A w =: 0.$$

This implies that

$$A^T y = A^T A w \implies \hat{\theta} = (A^T A)^{-1} A^T y$$

Why is $(A^T A)^{-1}$ invertible? (exercise). Hint: this also shows that $\hat{\theta}$ is unique!

Matrix Facts on previous slide

Note: We're using the fact above from matrix algebra that

$$\frac{\partial}{\partial w_j} a^T w = \sum_i a_i w_i = a_j.$$

The second fact we use is known as a quadratic form. Assume B is symmetric.

$$\frac{\partial}{\partial w_k} w^T B w = \frac{\partial}{\partial w_k} \sum_{i,j=1}^{n} w_i w_j b_{ij} \tag{6}$$

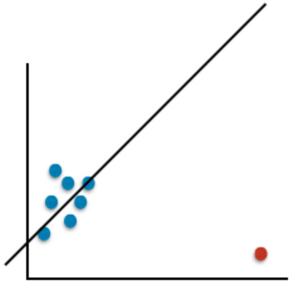
$$= \begin{cases} 2w_i b_{ij}, \text{ if } i = j = k \\ w_i b_{ij} \text{ if } j = k, i \neq j \end{cases} \tag{7}$$

We picked up some nice tricks for working with gradients.
Also, we can identify that $\hat{\theta}$ is unbiased. (exercise).
What is the variance of $\hat{\theta}$? (exercise).

Bayesian linear regression

We derived the MLE. Why not use the MLE?

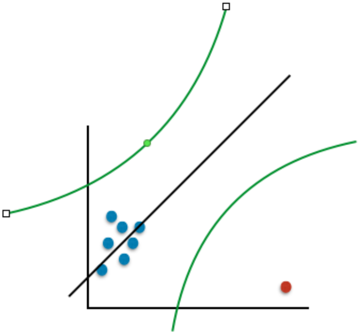
The MLE often overfits the data. Also, no notion of uncertainty.



Now suppose we want to predict a new point but what if this is the diagnostic for a patient. Or an investment for a stock portfolio.

How certain are you? (Let's put in error bars).

Bayesian linear regression



Now suppose we want to predict a new point but what if this is the diagnostic for a patient. Or an investment for a stock portfolio.

How certain are you? We're not certain at all!

Why Bayesian?

Bayesian approach allows you to say, I don't know!

We can tie back to decision theory and optimize a loss function by optimizing the predictive distribution

$$p(y \mid x, D)$$

Setup

$D = (x_i, y_i)$ for all i. Let $a^{-1} = 1/\sigma^2$.

$$y_i \mid w \overset{ind}{\sim} N(w^T x_i, a^{-1}) \tag{8}$$

$$w \sim MVN(0, b^{-1}, I) \tag{9}$$

$$\tag{10}$$

We assume that a,b are known. Here, $\theta = w$.
Recall: Look at the Multivariate model as these are needed to understand this module.

Computing the Posterior

What is the likelihood?

$$p(D \mid w) \propto P(D \mid w) \propto \exp\{-a/2(y - Aw)^T(y - Aw)\} \quad (11)$$

What is the posterior?

$$p(w \mid D) \propto p(D \mid w)p(w) \quad (12)$$

$$\propto \exp\{-a/2(y - Aw)^T(y - Aw)\} \times \exp\{-b/2w^T w)\} \quad (13)$$

Just like in the Multivariate modules, we just simplify. (Check these details on your own).

$$p(w \mid D) \propto MVN(w \mid \mu, \Lambda^{-1})$$

where $\Lambda = aA^T A + bI$ and $\mu = a\Lambda^{-1}A^T y$.

You can show (exercise that the Maximum a Posterior estimate of w is

$$a(aA^TA + bI)^{-1}A^Ty = (A^TA + b/aI)^{-1}A^Ty$$

How does this compare to the MLE estimate? Think about this on your own!

You will see more about Bayesian linear regression in lab. (For more on this, see Hoff).