The Exercise of Chap 2

Yanwei Fu

September 24, 2017

1 Simple Linear Regression

We have a quantitative response Y on the basis of a single regression predictor variable X. It assumes that there is approximately a linear relationship between X and Y,

$$Y \approx \beta_0 + \beta_1 X$$

In practice, $\beta_0$and β_1 are unknown; and to estimate the coefficients. We have n observations

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$$

Suppose we define the loss function as the *residual sum of squares (RSS)*,

$$RSS = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

Prove that the miniseries are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $\bar{y} \equiv \frac{1}{n} \sum_{i=1}^{n} y_i$, and $\bar{x} \equiv \frac{1}{n} \sum_{i=1}^{n} x_i$ are the sample means.(hint: $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$).

2 Regularized Least Squares (prove Eq (3.28), Page 145, Bishop book)

Given the dataset with n observations

$$(x_1, t_1), (x_2, t_2), ..., (x_n, t_n)$$

and the loss function is

$$\frac{1}{2}\sum_{n=1}^{n}\left\{t_n - w^T\phi\left(x_n\right)\right\}^2 + \frac{\lambda}{2}w^T w$$

please prove that $w = \left(\lambda I + \Phi^T\Phi\right)^{-1}\Phi^T t$

$$\Phi = \begin{bmatrix} \phi_0\left(x_1\right) & \phi_1\left(x_1\right) & \cdots & \phi_{M-1}\left(x_1\right) \\ \phi_0\left(x_2\right) & \phi_1\left(x_2\right) & \cdots & \phi_{M-1}\left(x_2\right) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0\left(x_N\right) & \phi_1\left(x_N\right) & \cdots & \phi_{M-1}\left(x_N\right) \end{bmatrix}$$

More details, please refer to "Sec. 3.1.4 Regularized least squares"(Bishop book);

3 Reducing the cost of linear regression for large d small n

The ridge method is a regularized version of least squares, with objective function:

$$\min_{\theta\in\mathbb{R}^d} \parallel y - X\theta \parallel_2^2 + \delta^2 \parallel \theta \parallel_2^2$$

Here δ is a scalar, the input matrix $X \in \mathbb{R}^{n\times d}$ and the output vector $y \in \mathbb{R}^n$. The parameter vector $\theta \in \mathbb{R}^d$ is obtained by differentiating the above cost function, yeilding the normal equations

$$\left(X^T X + \delta^2\mathbf{I}_d\right)\theta = X^T y$$

where $\mathbf{I}_d$ is the $d\times d$ identify matrix. The predictions $\hat{y} = \hat{y}\left(X_\star\right)$ for new test poitns $X_\star \in \mathbb{R}^{n\star\times d}$ are obtained by evaluating the hyperplane

$$\hat{y} = X_\star\theta = X_\star\left(X^T X + \delta^2\mathbf{I}_d\right)^{-1}X^T y = \mathbf{H}y$$

The matrix $\mathbf{H}$ is known as the hat matrix because it puts a "hat" on y.

Questions:

1. show that the solution can be written as $\theta = X^T\alpha$, where $\alpha = \delta^{-2}\left(y - X\theta\right)$.

2. show that α can also be written as follows: $\alpha = \left(XX^T + \delta^2\mathbf{I}_n\right)^{-1}y$ and, hence, the predictions can be written as follows,

$$\hat{y} = X_\star\theta = X_\star X^T\alpha = \left[X_\star X^T\right]\left(\left[XX^T + \delta^2\mathbf{I}_n\right]\right)^{-1}y$$

(Note that this is an awesome trick! For example, if $n = 20$ patients with $d = 10,000$ gene measurements, the computation of only requires inverting the $n\times n$ matrix, while the direct computation of would have required the inversion of a $d\times d$ matrix.)