A Note on Statistical Computing

Ying Nian Wu, UCLA Statistics

For STATS 202A, Fall quarter 2017

Contents

Preface

This note is mainly about the computational side of the commonly used modern statistical and machine learning methods. The main theme of the first three chapters is linear regression and its rich variations that span much of statistics and machine learning. The last chapter is on Monte Carlo methods.

Writing R and Python code to implement these methods enable us to gain first-hand experiences with these methods.

1 Linear regression: least squares, ridge, Lasso

1.1 Linear Regression

The dataset of linear regression consists of an $n \times p$ matrix $\mathbf{X} = (x_{ij})$, and a $n \times 1$ vector $\mathbf{Y} = (y_i)$. The model is of the following form:

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i,$$

for $i = 1, ..., n$, where $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ independently for $i = 1, ..., n$. Here we are deliberately ambiguous about intercept term. If $x_{i1} = 1$ for all i, then β_1 will be the the intercept term. $[\mathbf{X}, \mathbf{Y}]$ is called the training data. y_i is called response variable, outcome, dependent variable. x_{ij} is called predictor, regressor, covariate, independent variable, or simple variable. In the experimental design setting, $\mathbf{X}$ is called the design matrix.

The process of estimating β is called learning from the training data. The purpose is two-fold.

(1) Explanation: understanding the relationship between y_i and $(x_{ij}, j = 1, ..., p)$.

(2) Prediction: learn to predict y_i based on $(x_{ij}, j = 1, ..., p)$, so that in the testing stage, if we are given the predictor variables, we should be able to predict the outcome.

obs	$\mathbf{X}_{n\times p}$	$\mathbf{Y}_{n\times 1}$
1	$x_{11}, x_{12}, ..., x_{1p}$	y_1
2	$x_{21}, x_{22}, ..., x_{2p}$	y_2
...		
n	$x_{n1}, x_{n2}, ..., x_{np}$	y_n

obs	$\mathbf{X}_{n\times p}$	$\mathbf{Y}_{n\times 1}$
1	$X_1^\top$	y_1
2	$X_2^\top$	y_2
...		
n	$X_n^\top$	y_n

We can arrange the data in terms of $X_i^\top = (x_{ij}, j = 1, ..., p)$, where $X_i^\top$ is the i-th row of $\mathbf{X}$. Here X_i is not in bold font. We can write the model as $y_i = X_i^\top \beta + \epsilon_i$, where $\beta = (\beta_j, j = 1, ..., p)^\top$.

obs	$\mathbf{X}_{n\times p}$	$\mathbf{Y}_{n\times 1}$
1 2 ... n	$\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_p$	$\mathbf{Y}$

We can also arrange the data in terms of $\mathbf{X}_j = (x_{ij}, i = 1, ..., n)$, where $\mathbf{X}_j$ is the j-th column of $\mathbf{X}$. Here $\mathbf{X}_j$ is in bold font. We can write the model as $\mathbf{Y} = \sum_{j=1}^p \mathbf{X}_j \beta_j + \epsilon$, where $\epsilon = (\epsilon_i, i = 1, ..., n)^\top \sim \mathrm{N}(0, \sigma^2 \mathbf{I}_n)$, where $\mathbf{I}_n$ is the n-dimensional identity matrix.

We can write the linear regression model as $\mathbf{Y} = \mathbf{X}^\top \beta + \epsilon$. The least squares estimate of β is

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{Y} - \mathbf{X}\beta\|_{\ell_2}^2 = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

We construct a matrix $\mathbf{Z} = [\mathbf{X}\mathbf{Y}]$, and let

$$A = \mathbf{Z}^\top \mathbf{Z} = \begin{bmatrix} \mathbf{X}^\top \mathbf{X} & \mathbf{X}^\top \mathbf{Y} \\ \mathbf{Y}^\top \mathbf{X} & \mathbf{Y}^\top \mathbf{Y} \end{bmatrix}$$

be the cross-product matrix. Then

$$\mathrm{SWP}[1:p]A = \begin{bmatrix} -(\mathbf{X}^\top \mathbf{X})^{-1} & (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \\ \mathbf{Y}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} & \mathbf{Y}^\top \mathbf{Y} - \mathbf{Y}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} \end{bmatrix} = \begin{bmatrix} -\dfrac{\mathrm{Var}(\hat{\beta})}{\sigma^2} & \hat{\beta} \\ \hat{\beta}^\top & \mathrm{RSS} \end{bmatrix},$$

where $\mathrm{RSS} = \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|_{\ell_2}^2$ is the residual sum of squares.

We shall expand our treatment of linear regression in later sections. For now, it is enough to know that the sweep operator gives us all the key results we need for linear regression.

1.2 Gauss-Jordan elimination

For a system of linear equations $Ax = b$, where $A = (a_{ij})$ is $n \times n$, $x = (x_i)$ is $n \times 1$, and $b = (b_i)$ is $n \times 1$, we can solve $x = A^{-1} b$ by Gauss-Jordan elimination.

Specifically, for any matrix A (here we assume A can be any $n \times N$ matrix, e.g., $N = n + 1$ or $N = 2n$), let $\tilde{A} = \mathrm{GJ}[k]A$, then

$$\begin{aligned} \tilde{A}_k &= A_k / a_{kk}, \\ \tilde{A}_i &= A_i - a_{ik} \tilde{A}_k, \ i \neq k, \end{aligned}$$

where A_k is the k-th row of A. The above operation makes $\tilde{a}_{kk} = 1$, and $\tilde{a}_{ik} = 0$ for $i \neq k$. We can apply Gauss-Jordan sequentially, e.g., $\text{GJ}[1:m]$ means we apply Gauss-Jordan for $k = 1:m$. Being a row operation, Gauss-Jordan is linear, i.e., $\tilde{A} = \text{GJ}[k]A$ amounts to $\tilde{A} = G_k A$ for a matrix G_k. Thus

$$\text{GJ}[1:n][A|b] = [I|A^{-1}b] = A^{-1}[A|b],$$

$$\text{GJ}[1:n][A|I] = [I|A^{-1}] = A^{-1}[A|I].$$

That is, $\text{GJ}[1:n] = A^{-1}$, and order does not matter. Moreover,

$$\text{GJ}[1:m]\begin{bmatrix} A_{11} & A_{12} & | & I_1 & 0 \\ A_{21} & A_{22} & | & 0 & I_2 \end{bmatrix} = \begin{bmatrix} I_1 & A_{11}^{-1}A_{12} & | & A_{11}^{-1} & 0 \\ 0 & A_{22} - A_{21}A_{11}^{-1}A_{12} & | & -A_{21}A_{11}^{-1} & I_2 \end{bmatrix}.$$

We can write $\text{GJ}[1:m] = \text{GJ}[A_{11}]$, which is a matrix version of Gauss-Jordan. If we compare Gauss-Jordan $\text{GJ}[1:m]$ with the sweep operator $\text{SWP}[1:m]$, we can see that sweep operator is a space saving version of Gauss-Jordan, where we do not record the identity matrix in sweep. Because $\text{GJ}[1:m] = \text{GJ}[A_{11}]$, we have $\text{SWP}[1:m] = \text{SWP}[A_{11}]$. The above equation also leads to the identity

$$|A| = |A_{11}||A_{22} - A_{21}A_{11}^{-1}A_{12}|,$$

where $|A|$ denotes the determinant of A. Thus we can compute $|A|$ by the sweep operator, in addition to A^{-1}.

1.3 Calculation details

To flesh out the details of least squares estimation, let

$$R(\beta) = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2,$$

we have

$$\frac{R(\beta)}{\partial \beta_j} = -2\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)x_{ij}.$$

We can pack the above results in terms of $\mathbf{X}_j$ and X_i. Recall

obs	$\mathbf{X}_{n\times p}$	$\mathbf{Y}_{n\times 1}$
1	$X_1^\top$	y_1
2	$X_2^\top$	y_2
...		
n	$X_n^\top$	y_n

obs	$\mathbf{X}_{n\times p}$	$\mathbf{Y}_{n\times 1}$
1	$\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_p$	$\mathbf{Y}$
2		
...		
n		

$$\frac{\partial R(\beta)}{\partial \beta_j} = -2\langle \mathbf{Y} - \sum_{j=1}^{p}\mathbf{X}_j\beta_j, \mathbf{X}_j\rangle = -2\mathbf{X}_j^\top(\mathbf{Y} - \sum_{j=1}^{p}\mathbf{X}_j\beta_j),$$

$$R'(\beta) = \begin{bmatrix} \partial R/\partial\beta_1 \\ \partial R/\partial\beta_2 \\ ... \\ \partial R/\partial\beta_n \end{bmatrix} = -2\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)\begin{bmatrix} x_{i1} \\ x_{i2} \\ ... \\ x_{ip} \end{bmatrix} = -2\sum_{i=1}^{n}(y_i - X_i^\top\beta)X_i.$$

We can further pack the above two equations as

$$\frac{\partial R(\beta)}{\partial\beta_j} = -2\mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\beta).$$

For the second derivative,

$$\frac{\partial^2 R(\beta)}{\partial\beta_j\beta_k} = 2\sum_{i=1}^{n} x_{ij}x_{ik} = 2\langle\mathbf{X}_j, \mathbf{X}_k\rangle.$$

Define

$$R''(\beta) = \left(\frac{\partial^2 R(\beta)}{\partial\beta_j\beta_k}\right)$$

be the $p \times p$ matrix, we can write

$$R''(\beta) = 2\sum_{i=1}^{n} X_i X_i^\top = 2\mathbf{X}^\top\mathbf{X},$$

which is positive definite.

In order to solve the least squares problem, we only need to solve $R'(\beta) = 0$. Geometrically, it means $\mathbf{X}_j \perp \mathbf{R}$, where $\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}}$, and $\hat{\mathbf{Y}} = \sum_{j=1}^{p}\mathbf{X}_j\hat{\beta}_j$ is the projection of $\mathbf{Y}$ onto the subspace spanned by $(\mathbf{X}_j, j = 1, ..., p)$.

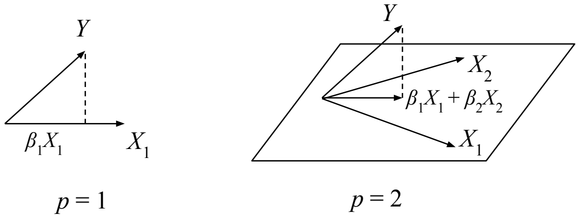

Figure 1: Least squares projection

The calculation for ridge regression is similar. Let $f(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|_{\ell_2}^2 + \lambda\|\beta\|_{\ell_2}^2$, then

$$f'(\beta) = -2\mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\beta) + 2\lambda\beta,$$

$$f''(\beta) = 2(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p).$$

1.4 Matrix calculus

Suppose $Y = (y_i)_{m\times1}$, and $X = (x_j)_{n\times1}$. Suppose $Y = h(X)$. We can define

$$\frac{\partial Y}{\partial X^\top} = \left(\frac{\partial y_i}{\partial x_j}\right)_{m\times n}.$$

If Y is a scaler, then the gradient $h'(X) = \partial Y/\partial X$ is a $n \times 1$ column vector, and $\partial Y/\partial X^\top$ is a $1 \times n$ row vector. For scaler Y, we can define the Hessian or second derivative

$$h''(X) = \frac{\partial^2 Y}{\partial X\partial X^\top} = \left(\frac{\partial^2 Y}{\partial x_i\partial x_j}\right)_{n\times n}.$$

If $Y = AX$, then $y_i = \sum_k a_{ik} x_k$. Thus $\partial y_i / \partial x_j = a_{ij}$. So $\partial Y / \partial X^\top = A$.

Chain rule. If $Y = h(X)$ and $X = g(Z)$, then $\partial y_i / \partial z_j = \sum_k (\partial y_i / \partial x_k)(\partial x_k / \partial z_j)$. Thus

$$\frac{\partial Y}{\partial Z^\top} = \frac{\partial Y}{\partial x^\top} \frac{\partial X}{\partial Z^\top}.$$

Product rule. If $Y = \langle h(X), g(X) \rangle = \sum_i h_i(X) g_i(X)$, then $\partial Y / \partial x_j = \sum_i [\partial h_i / \partial x_j g_i + h_i \partial g_i / \partial x_j]$. So

$$\frac{\partial Y}{\partial X^\top} = h(X)^\top \frac{\partial g(X)}{\partial X^\top} + g(X)^\top \frac{\partial h(X)}{\partial X^\top}.$$

Taylor expansion

$$f(X) = f(X_0) + \langle f'(X_0), X - X_0 \rangle + \frac{1}{2}(X - X_0)^\top f''(X_0)(X - X_0) + o(|X - X_0|^2).$$

Jacobian. Let $Y = h(X)$ where both X and Y are $n \times 1$. Assume that h is one-to-one differentiable mapping. Let D_X be a local region around X in the domain of X. Suppose h maps D_X to a region D_Y in the domain of Y. Then as the size of D_X goes to 0, $|D_Y|/|D_X| \to |h'(X)|$, where $|h'(X)|$ is the determinant of $h'(X) = \partial Y / \partial X^\top$.

For $R(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|^2$, $R'(\beta) = -2\mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\beta)$ and $R''(\beta) = 2\mathbf{X}^\top\mathbf{X}$. We can derive these by the chain rule. Let $e = \mathbf{Y} - \mathbf{X}\beta$. Then

$$\frac{\partial R}{\partial \beta^\top} = \frac{\partial R}{\partial e^\top} \frac{\partial e}{\partial \beta^\top} = -2e^\top \mathbf{X}.$$

$R'(\beta) = \partial R / \partial \beta$, which is obtained by transposing $-2e^\top \mathbf{X}$.

$$R''(\beta) = \frac{\partial^2 R}{\partial \beta \partial \beta^\top} = \partial(-2\mathbf{X}^\top e)/\partial \beta^\top = -2\mathbf{X}^\top \mathbf{X}.$$