SHEET - An Introduction to Statistical Learning
Chapter 5 - Classification

23 september 2024

1 Linear Model Selection and Regularization

1.1 Courses' Demonstrations

Suppose that we wish to invest a fixed sum of money in two financial assets that yield returns of X and Y , respectively, where X and Y are random quantities. We will invest a fraction α of our money in X, and will invest the remaining $1-\alpha$ in Y . Since there is variability associated with the returns on these two assets, we wish to choose α to minimize the total risk, or variance, of our investment. In other words, we want to minimize $Var(\alpha X + (1-\alpha)Y)$. One can show that the value that minimizes the risk is given by

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}} \tag{5.6}$$

Demonstration : $\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$

$Var(\alpha X + (1-\alpha)Y) = \alpha^2 Var(X) + (1-\alpha)^2 Var(Y) + 2\alpha(1-\alpha)Cov(XY)$

We search α **that minimizes** $Var(\alpha X + (1-\alpha)Y)$

$\dfrac{d}{d\alpha} Var(\alpha X + (1-\alpha)Y) = 0$

$2\alpha Var(X) - 2(1-\alpha)Var(Y) + 2(1-\alpha)Cov(XY) - 2\alpha Cov(XY) = 0$

$2\alpha Var(X) - 2(1-\alpha)Var(Y) + 2Cov(XY) - 2\alpha Cov(XY) - 2\alpha Cov(XY) = 0$

$\alpha Var(X) + (\alpha - 1)Var(Y) + Cov(XY) - 2\alpha Cov(XY) = 0$

$\alpha(Var(X) + Var(Y) - 2Cov(XY)) = Var(Y) - Cov(XY)$

Finally

$alpha = \dfrac{Var(Y)^2 - Cov(XY)}{Var(X)^2 + Var(Y)^2 - 2Cov(XY)}$

$\alpha = \dfrac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$

In order to obtain a better intuition about the behavior of ridge regression and the lasso, consider a simple special case with $n = p$, and X a diagonal matrix with 1's on the diagonal and 0's in all off-diagonal elements. To simplify the problem further, assume also that we are performing regression without an intercept. With these assumptions, the usual least squares problem simplifies to finding $\beta_1, \ldots, \beta_p$ that minimize

$$\sum_{j=1}^{p}(y_j - \beta_j)^2. \quad (6.11)$$

In this case, the least squares solution is given by $\hat{\beta}_j = y_j$. And in this setting, ridge regression amounts to finding $\beta_1, \ldots, \beta_p$ such that

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \quad (6.12)$$

is minimized, and the lasso amounts to finding the coefficients such that

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \quad (6.13)$$

is minimized. One can show that in this setting, the ridge regression estimates take the form

$$\hat{\beta}_j^R = \frac{y_j}{1 + \lambda}, \quad (6.14)$$

and the lasso estimates take the form

$$\hat{\beta}_j^L = \begin{cases} y_j - \frac{\lambda}{2} & \text{if } y_j > \frac{\lambda}{2}; \\ y_j + \frac{\lambda}{2} & \text{if } y_j < -\frac{\lambda}{2}; \\ 0 & \text{if } |y_j| \leq \frac{\lambda}{2}. \end{cases} \quad (6.15)$$

Demonstration : The ridge regression estimates take the form : $\hat{\beta}_j^R = \frac{y_j}{1+\lambda}$, (6.14)

We have

$$L_R = \sum_{j=1}^{p}(y_j - \hat{\beta}_j)^2 + \lambda \sum_{j=1}^{p} \hat{\beta}_j^2 \quad (6.12)$$

$$\frac{d}{d\hat{\beta}_j} L_R = 0$$

Then

$$-2(y_j - \hat{\beta}_j) + 2\lambda\hat{\beta}_j = 0$$

$$= -y_j + \hat{\beta}_j(1+\lambda) = 0$$

$$\hat{\beta}_j = \frac{y_j}{1+\lambda}$$

Demonstration : The lasso regression estimates take the form :

$$\hat{\beta}_j^L = \begin{cases} y_j - \frac{\lambda}{2} & \textbf{if } y_j > \frac{\lambda}{2}; \\ y_j + \frac{\lambda}{2} & \textbf{if } y_j < -\frac{\lambda}{2}; \\ 0 & \textbf{if } |y_j| \leq \frac{\lambda}{2}. \end{cases} \qquad (6.15)$$

We have

$$L_L = \sum_{j=1}^{p} (y_j - \hat{\beta}_j)^2 + \lambda \sum_{j=1}^{p} |\hat{\beta}_j| \qquad (6.13)$$

$$\frac{d}{d\hat{\beta}_j} L_R = 0$$

Then

$$-2(y_j - \hat{\beta}_j) + \lambda sign(\hat{\beta}_j) = 0$$

$$2\hat{\beta}_j = 2y_j - \lambda sign(\hat{\beta}_j)$$

$$\hat{\beta}_j = y_j - \frac{\lambda}{2} sign(\hat{\beta}_j)$$

We must have $\hat{\beta}_j \geq 0$

Then

$$\hat{\beta}_j^L = \begin{cases} y_j - \frac{\lambda}{2} & \text{if } y_j > \frac{\lambda}{2}; \\ y_j + \frac{\lambda}{2} & \text{if } y_j < -\frac{\lambda}{2}; \\ 0 & \text{if } |y_j| \leq \frac{\lambda}{2}. \end{cases}$$

We now show that one can view ridge regression and the lasso through a Bayesian lens. A Bayesian viewpoint for regression assumes that the coefficient vector β has some prior distribution, say $p(\beta)$, where $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$. The likelihood of the data can be written as $f(Y|X, \beta)$, where $X = (X_1, \ldots, X_p)$. Multiplying the prior distribution by the likelihood gives us (up to a proportionality constant) the posterior distribution, which takes the form

$$p(\beta|X, Y) \propto f(Y|X, \beta)p(\beta|X) = f(Y|X, \beta)p(\beta),$$

where $p(\beta|X, Y)$ denotes the posterior distribution, $f(Y|X, \beta)$ is the likelihood, and $p(\beta)$ is the prior distribution.

Demonstration : $p(\beta|X, Y) \propto f(Y|X, \beta)p(\beta|X) = f(Y|X, \beta)p(\beta)$

We have

$$p(A|B) = \frac{P(B|A)P(A)}{p(B)}$$

Then

$$p(\beta|X, Y) = \frac{p(Y|\beta, X)p(\beta|X)}{p(Y|X)}$$

Because we have

$$p(Y \mid X) = \int p(Y \mid X, \beta)p(\beta)\, d\beta$$

Then $p(Y|X)$ **indepedent of** β

Then $p(Y|X)$ **is a constant for** $p(\beta|X, Y)$

Then

$$p(\beta|X, Y) = \frac{p(Y|\beta, X)p(\beta|X)}{p(Y|X)}$$

$$\propto p(Y|\beta, X)p(\beta|X)$$

$$= p(Y|\beta, X)\frac{p(X|\beta)p(\beta)}{p(X)}$$

And because $p(X|\beta)$ **independent of** β

$$= p(Y|\beta, X)\frac{p(X)p(\beta)}{p(X)}$$

$$= p(Y|\beta, X)p(\beta)$$