

HOMework 3

REGRESSION, GAUSSIAN PROCESSES, AND BOOSTING

DANA VAN AKEN

Problem 2: Regression

2.1 Why Lasso Works

1. Write $J_\lambda(\beta)$ in the form $J_\lambda(\beta) = g(y) + \sum_1^d f(X_i, y, \beta_i, \lambda)$, $\lambda > 0$:

$$\begin{aligned} J_\lambda(\beta) &= \frac{1}{2} \|y - X\beta\|^2 + \lambda \|\beta\| \\ &= \frac{1}{2} (y - X\beta)^T (y - X\beta) + \lambda \|\beta\| \\ &= \frac{1}{2} [y^T y - 2y^T X\beta + (X\beta)^T X\beta] + \lambda \|\beta\| \\ &= \frac{1}{2} [y^T y - 2y^T X\beta + \beta^T X^T X\beta] + \lambda \|\beta\| \\ &= \frac{1}{2} [y^T y - 2y^T X\beta + \beta^T \beta] + \lambda \|\beta\| \quad (X^T X = I) \\ &= \frac{1}{2} y^T y - y^T X\beta + \frac{1}{2} \beta^T \beta + \lambda \|\beta\| \\ &= \frac{1}{2} y^T y + \sum_{i=1}^d \frac{1}{2} \beta_i^T \beta_i - y^T X_i \beta_i + \lambda \|\beta_i\| \end{aligned}$$

Let $g(y) = \frac{1}{2} y^T y$ and $f(X_i, y, \beta_i, \lambda) = \frac{1}{2} \beta_i^T \beta_i - y^T X_i \beta_i + \lambda \|\beta_i\|$, then:

$$J_\lambda(\beta) = g(y) + \sum_{i=1}^d f(X_i, y, \beta_i, \lambda)$$

2. $\beta_i^* > 0$:

Calculating the derivative of $f(X_i, y, \beta_i^*, \lambda)$, (where i in $\{1 \dots d\}$), we get:

$$\frac{f(X_i, y, \beta_i^*, \lambda)}{d\beta_i^*} = \beta_i^* - y^T X_i + \lambda$$

Setting the LHS equal to zero and solving for β_i^* gives:

$$\beta_i^* = y^T X_i - \lambda \quad (1)$$

3. $\beta_i^* < 0$:

Calculating the derivative of $f(X_i, y, \beta_i^*, \lambda)$, (where i in $\{1 \dots d\}$), we get:

$$\frac{f(X_i, y, \beta_i^*, \lambda)}{d\beta_i^*} = \beta_i^* - y^T X_i - \lambda$$

Setting the LHS equal to zero and solving for β_i^* gives:

$$\beta_i^* = y^T X_i + \lambda \quad (2)$$

4. In both equations (1) and (2), as we increase λ , β_i^* gets closer and closer to zero (assuming β_i^* and $y^T X_i$ are the same sign). Once you increase λ enough that β_i^* reaches zero, it sticks there because moving it below zero increases the L1 penalty and moves it further away from the least squares term (mathematically, λ switches its sign at this point because of the characteristics of the absolute value function).

5. Calculating the derivative of $f(X_i, y, \beta_i^*, \lambda)$, (where i in $\{1 \dots d\}$), with the regularization term $\frac{1}{2} \|\beta_i^*\|_2^2$ we get:

$$\frac{f(X_i, y, \beta_i^*, \lambda)}{d\beta_i^*} = \beta_i^* - y^T X_i + \lambda \beta_i^*$$

Setting the LHS equal to zero and solving for β_i^* gives:

$$\beta_i^* = \frac{y^T X_i}{1 + \lambda}$$

Unlike equations (1) and (2), there is no value of λ that can drive β_i^* to zero. This demonstrates why Lasso regression often results in “sparser” solutions whereas Ridge regression does not.

2.2 Bayesian regression and Gaussian process

1. (a)

(b)

2.

3.

4.