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 $J_\lambda(\beta) = g(y) + \sum_1^d f(X_i, y, \beta_i, \lambda)$, $\lambda > 0$:

   $J_\lambda(\beta) = \frac{1}{2}\left\|y - X\beta\right\|^2 + \lambda\left\|\beta\right\|$

   $= \frac{1}{2}(y - X\beta)^T(y - X\beta) + \lambda\left\|\beta\right\|$

   $= \frac{1}{2}[y^Ty - 2y^TX\beta + (X\beta)^TX\beta] + \lambda\left\|\beta\right\|$

   $= \frac{1}{2}[y^Ty - 2y^TX\beta + \beta^TX^TX\beta] + \lambda\left\|\beta\right\|$

   $= \frac{1}{2}[y^Ty - 2y^TX\beta + \beta^T\beta] + \lambda\left\|\beta\right\| \qquad (X^TX = I)$

   $= \frac{1}{2}y^Ty - y^TX\beta + \frac{1}{2}\beta^T\beta + \lambda\left\|\beta\right\|$

   $= \frac{1}{2}y^Ty + \sum_{i=1}^d \frac{1}{2}\beta_i^T\beta_i - y^TX_i\beta_i + \lambda\left\|\beta_i\right\|$

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

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

2. $\beta_i^* > 0$:

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

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

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

   $$\boldsymbol{\beta_i^* = y^TX_i - \lambda} \tag{1}$$

3. $\beta_i^* < 0$:

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

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

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

$$\boldsymbol{\beta_i^* = y^T X_i + \lambda} \tag{2}$$

4. In both equations (1) and (2), as we increase λ, β_i^* gets closer and closer to zero (this is because β_i^* and $y^T X_i$ are the same sign since $\lambda > 0$). 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, so if β_i^* passed 0 then the equation would be inconsistent with the ones we just derived).

5. Calculating the derivative of $f(X_i, y, \beta_i^*, \lambda)$, (where i in $\{1...d\}$), with the regularization term $\frac{1}{2}\left\|\beta_i^*\right\|_2^2$ 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 alpha 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) Derive the posterior distribution:

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

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

Find the distributions of w and $p(Y|X, w)$:

$w \sim N(0, \Sigma_p) = N(0, \sigma_0^2 I)$

$\epsilon I = Y - f(X) = Y - \Phi^T w \sim N(0, \sigma_n^2 I)$

$Y|X, w \sim N(\Phi^T w, \sigma_n^2 I)$

Multiply the distributions:

$$
\begin{aligned}
p(w|Y,X) &\propto p(Y|X,w)p(w) \\
&\propto N(\Phi^T w, \sigma_n^2 I)N(0, \sigma_0^2 I) \\
&\propto exp[-\frac{1}{2}(y-\Phi^T w)^T(\sigma_n^2 I)^{-1}(y-\Phi^T w)]exp[-\frac{1}{2}w^T\Sigma_p^{-1}w] \\
&\propto exp[-\frac{1}{2}(\sigma_n^{-2}y^Ty - \sigma_n^{-2}y^T\Phi^T w - \sigma_n^{-2}w^T\Phi y + \sigma_n^{-2}w^T\Phi\Phi^T w + w^T\Sigma_p^{-1}w)]
\end{aligned}
$$

Remove constants that do not depend on w:

$$\propto exp[-\frac{1}{2}(-\sigma_n^{-2}y^T\Phi^T w - \sigma_n^{-2}w^T\Phi y + \sigma_n^{-2}w^T\Phi\Phi^T w + w^T\Sigma_p^{-1}w)]$$

Complete the square:

$$\propto exp(\frac{-1}{2}(w-(\sigma_n^{-2}\Phi\Phi^T+\Sigma_p^{-1})^{-1}\Phi y)(\sigma_n^{-2}\Phi\Phi^T+\Sigma_p^{-1})(w-(\sigma_n^{-2}\Phi\Phi^T+\Sigma_p^{-1})^{-1}\Phi y))$$

$$\sim N(\sigma_n^{-2}(\sigma_n^{-2}\Phi\Phi^T+\Sigma_p^{-1})^{-1}\Phi y,\ \sigma_n^{-2}\Phi\Phi^T+\Sigma_p^{-1})$$

(b)

$$p(f_*|X_*,X,Y) = \int p(f_*|X_*,w)p(w|X,Y)dw$$

Using Gaussian mean and covariance identities:

$$\sim N(\sigma_n^{-2}\Phi_*^T(\sigma_n^{-2}\Phi\Phi^T+\Sigma_p^{-1})^{-1}\Phi y,\ \Phi_*^T(\sigma_n^{-2}\Phi\Phi^T+\Sigma_p^{-1})\Phi_*)$$

2. Using problem 1.d:

$$f_*|X_*,X,Y \sim N(\sigma_o^2\Phi_*^T\Phi(\sigma_o^2\Phi^T\Phi+\sigma_n^2 I)^{-1}y,\ \sigma_o^2\Phi_*^T\Phi_* - \sigma_o^2\Phi_*\Phi(\sigma_o^2\Phi^T\Phi+\sigma_n^2 I)^{-1}\sigma_o^2\Phi^T\Phi_*)$$

3. Show $\sigma_n^{-2}\Phi_*^T(\sigma_n^{-2}\Phi\Phi^T+\Sigma_p^{-1})^{-1}\Phi y = \sigma_o^2\Phi_*^T\Phi(\sigma_o^2\Phi^T\Phi+\sigma_n^2 I)^{-1}y$

   Multiply σ_n^{-2} and σ_o^2 through on right and left side:

   $\Phi_*^T(\Phi\Phi^T+\sigma_n^2\Sigma_p^{-1})^{-1}\Phi y = \Phi_*^T\Phi(\Phi^T\Phi+\sigma_n^2\Sigma_p^{-1})^{-1}y$

   Multiply both sides through by y^{-1} from the left and Φ_*^T from the right:

   $(\Phi\Phi^T+\sigma_n^2\Sigma_p^{-1})^{-1}\Phi = \Phi(\Phi^T\Phi+\sigma_n^2\Sigma_p^{-1})^{-1}$

   Multiply through on each side to get rid of the inverses:

   $(\Phi^T\Phi+\sigma_n^2\Sigma_p^{-1})\Phi = \Phi(\Phi\Phi^T+\sigma_n^2\Sigma_p^{-1})$

   Multiply the Φ on the RHS though and then pull it out on the other side:

   $(\Phi^T\Phi+\sigma_n^2\Sigma_p^{-1})\Phi = (\Phi^T\Phi+\sigma_n^2\Sigma_p^{-1})\Phi$

This shows they are equal.

4. To do the prediction, we must invert either an nxn matrix (in equation 1.b) or a DxD matrix (in equation 2) which is computationally expensive. For this reason, if $D > n$ then we should use equation 1.b, and if $n > D$ then we should use equation 2.