Homework 5

Solutions

1. (Augmented dataset) The ridge-regression cost function can be reformulated as the following least-squares problem:

$$\beta_{\text{RR}} := \arg\min_{\beta} \left|\left| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} X^T \\ \sqrt{\lambda} I \end{bmatrix} \beta \right|\right|_2^2. \tag{1}$$

There are p additional examples. For the ith example the feature vector equals a one-hot vector where the ith entry equals λ and the rest equal zero. All the response values for the additional examples equal zero. The additional examples force the linear coefficients to fit zeros, preventing them from becoming too large.

2. (Correlated features)

   (a) Let's define $\bar{\alpha} := \sqrt{1-\alpha^2}$ to alleviate notation. We have

$$\begin{aligned}
\beta_{\text{OLS}} &= (XX^T)^{-1} Xy && (2)\\
&= \begin{bmatrix} 1 & \alpha \\ \alpha & 1 \end{bmatrix}^{-1} \begin{bmatrix} w_1^T \\ \alpha w_1^T + \bar{\alpha} w_\perp^T \end{bmatrix} (\beta_{\text{true}} w_1 + z) && (3)\\
&= \frac{1}{1-\alpha^2} \begin{bmatrix} 1 & -\alpha \\ -\alpha & 1 \end{bmatrix} \begin{bmatrix} \beta_{\text{true}} + 0.1 \\ \alpha(\beta_{\text{true}} + 0.1) + 0.1\bar{\alpha} \end{bmatrix} && (4)\\
&= \frac{1}{1-\alpha^2} \begin{bmatrix} (1-\alpha^2)(\beta_{\text{true}} + 0.1) - 0.1\alpha\bar{\alpha} \\ 0.1\bar{\alpha} \end{bmatrix} && (5)\\
&= \begin{bmatrix} \beta_{\text{true}} + 0.1 - \frac{0.1\alpha}{\bar{\alpha}} \\ \frac{0.1}{\bar{\alpha}} \end{bmatrix}. && (6)
\end{aligned}$$

   The estimated coefficients tend to

$$\lim_{\alpha \to 1} \beta_{\text{OLS}} = \begin{bmatrix} -\infty \\ \infty \end{bmatrix}. \tag{7}$$

   The OLS estimate overfits the response vector, which causes the coefficient estimates to explode.

   (b) The estimate of the response equals

$$\begin{aligned}
y_{\text{OLS}} &:= X^T \beta_{\text{OLS}} && (8)\\
&= \begin{bmatrix} w_1 & \alpha w_1 + \bar{\alpha} w_\perp \end{bmatrix} \begin{bmatrix} \beta_{\text{true}} + 0.1 - \frac{0.1\alpha}{\bar{\alpha}} \\ \frac{0.1}{\bar{\alpha}} \end{bmatrix} && (9)\\
&= \left(\beta_{\text{true}} + 0.1 - \frac{0.1\alpha}{\bar{\alpha}} \right) w_1 + \frac{0.1}{\bar{\alpha}} (\alpha w_1 + \bar{\alpha} w_\perp) && (10)\\
&= (\beta_{\text{true}} + 0.1) w_1 + 0.1 w_\perp, && (11)
\end{aligned}$$

   which does not change as $\alpha \to 1$. The estimate of the response is not collinear with w_1 even when $\alpha \to 1$, even though the second feature tends to be collinear with the

first one in this limit! This is because the noise in the training response vector has a component in that direction, and the linear regression estimate tries to fit it even if the corresponding component in the second feature vector is very tiny.

(c)

$$\beta_{\text{RR}} = (XX^T + \lambda I)^{-1} Xy \tag{12}$$

$$= \begin{bmatrix} 1+\lambda & \alpha \\ \alpha & 1+\lambda \end{bmatrix}^{-1} \begin{bmatrix} w_1^T \\ \alpha w_1^T + \bar{\alpha} w_\perp^T \end{bmatrix} (\beta_{\text{true}} w_1 + z) \tag{13}$$

$$= \frac{1}{(1+\lambda)^2 - \alpha^2} \begin{bmatrix} 1+\lambda & -\alpha \\ -\alpha & 1+\lambda \end{bmatrix} \begin{bmatrix} \beta_{\text{true}} + 0.1 \\ \alpha(\beta_{\text{true}} + 0.1) + 0.1\bar{\alpha} \end{bmatrix} \tag{14}$$

$$= \frac{1}{(1+\lambda)^2 - \alpha^2} \begin{bmatrix} (1+\lambda-\alpha^2)(\beta_{\text{true}} + 0.1) - 0.1\alpha\bar{\alpha} \\ \lambda\alpha(\beta_{\text{true}} + 0.1) + 0.1\bar{\alpha}(1+\lambda) \end{bmatrix}. \tag{15}$$

The estimated coefficients tend to

$$\lim_{\alpha \to 1} \beta_{\text{RR}} = \frac{1}{2\lambda + \lambda^2} \begin{bmatrix} \lambda(\beta_{\text{true}} + 0.1) \\ \lambda(\beta_{\text{true}} + 0.1) \end{bmatrix} \tag{16}$$

$$= \frac{1}{2+\lambda} \begin{bmatrix} \beta_{\text{true}} + 0.1 \\ \beta_{\text{true}} + 0.1 \end{bmatrix}. \tag{17}$$

The coefficients no longer explode as in OLS. The two coefficients are the same, which makes sense because when $\alpha \to 1$ both features are equal to w_1.

(d)

$$\lim_{\alpha \to 1} y_{\text{RR}} := (\lim_{\alpha \to 1} X)^T \lim_{\alpha \to 1} \beta_{\text{RR}} \tag{18}$$

$$= \begin{bmatrix} w_1 & w_1 \end{bmatrix} \frac{1}{2+\lambda} \begin{bmatrix} \beta_{\text{true}} + 0.1 \\ \beta_{\text{true}} + 0.1 \end{bmatrix} \tag{19}$$

$$= \frac{2(\beta_{\text{true}} + 0.1) w_1}{2+\lambda}. \tag{20}$$

In contrast to OLS, the response estimate is collinear with the true feature vector as $\alpha \to 1$.

3. (Prior knowledge)

(a) A natural way to modify the ridge-regression cost function is to incorporate β_{prior} in the regularization term, to promote solutions that are close to it:

$$\beta_{\text{RRP}} := \arg\min_{\beta} \|y - X^T\beta\|_2^2 + \lambda \|\beta_{\text{prior}} - \beta\|_2^2. \tag{21}$$

(b) The cost function can be reformulated to equal a modified least-squares problem

$$\beta_{\text{RR}} := \arg\min_{\beta} \left\| \begin{bmatrix} \tilde{y} \\ \sqrt{\lambda}\beta_{\text{prior}} \end{bmatrix} - \begin{bmatrix} X^T \\ \sqrt{\lambda} I \end{bmatrix} \beta \right\|_2^2. \tag{22}$$

By Theorem 2.4 the solution equals

$$\tilde{\beta}_{\text{RRP}} = \left(\begin{bmatrix} X & \sqrt{\lambda} I \end{bmatrix} \begin{bmatrix} X & \sqrt{\lambda} I \end{bmatrix}^T \right)^{-1} \begin{bmatrix} X & \sqrt{\lambda} I \end{bmatrix} \begin{bmatrix} \tilde{y} \\ \sqrt{\lambda} \beta_{\text{prior}} \end{bmatrix} \tag{23}$$

$$= \left(X X^T + \lambda I \right)^{-1} \left(X \tilde{y} + \lambda \beta_{\text{prior}} \right) \tag{24}$$

$$= \left(X X^T + \lambda I \right)^{-1} \left(X \left(X^T \beta_{\text{true}} + \tilde{z}_{\text{train}} \right) + \lambda \beta_{\text{prior}} \right) \tag{25}$$

$$= U(S^2 + \lambda I)^{-1} U^T \left(U S^2 U^T \beta_{\text{true}} + U S V^T \tilde{z}_{\text{train}} + \lambda \beta_{\text{prior}} \right) \tag{26}$$

$$= U(S^2 + \lambda I)^{-1} S^2 U^T \beta_{\text{true}} + \lambda U (S^2 + \lambda I)^{-1} U^T \beta_{\text{prior}} + U \left(S^2 + \lambda I \right)^{-1} S V^T \tilde{z}_{\text{train}}.$$

The mean changes to $U(S^2 + \lambda I)^{-1} S^2 U^T \beta_{\text{true}} + \lambda U (S^2 + \lambda I)^{-1} U^T \beta_{\text{prior}}$, which approaches β_{prior} as λ increases. The covariance matrix does not change.

(c) A natural way to incorporate the prior knowledge in gradient descent is to initialize the coefficient estimate at β_{prior}. By equation 144 in the notes, the update equation is

$$\beta^{(k+1)} = \left(I - \alpha X X^T \right)^{k+1} \beta_{\text{prior}} + \sum_{i=0}^{k} \left(I - \alpha X X^T \right)^i \alpha X y. \tag{27}$$

(d) To ease notation, let $\tau_j := 1 - \alpha s_j^2$. By Theorem 6.1

$$\beta^{(k)} = U \operatorname{diag}_{j=1}^{p} \left(\tau_j^k \right) U^T \beta_{\text{prior}} + U \operatorname{diag}_{j=1}^{p} \left(\frac{1 - \tau_j^k}{s_j} \right) V^T y \tag{28}$$

$$= U \operatorname{diag}_{j=1}^{p} \left(\tau_j^k \right) U^T \beta_{\text{prior}} + U \operatorname{diag}_{j=1}^{p} \left(1 - \tau_j^k \right) U^T \beta_{\text{true}} + U \operatorname{diag}_{j=1}^{p} \left(\frac{1 - \tau_j^k}{s_j} \right) V^T \tilde{z}_{\text{train}}.$$

The covariance matrix remains the same. The mean changes to

$$\sum_{j=1}^{p} \left((1 - \alpha s_j^2)^k \langle u_j, \beta_{\text{prior}} \rangle + \left(1 - (1 - \alpha s_j^2)^k \right) \langle u_j, \beta_{\text{true}} \rangle \right) u_j. \tag{29}$$

4. (a) We have

$$\hat{\beta} = [9.50608976, 39.07089651, -8.181986, -23.08385266, -4.18995163],$$

with training loss 34.743450 and validation loss 11.283276.

(b) Singular values are

$$[21.10424705, 4.55884277, 0.09244803, 0.04802712, 0.03404531].$$

(c) By the previous part the data matrix is approximately rank 2. The last 4 columns are highly correlated, and thus differences of these columns are nearly in the null space of X. Thus the coefficients can vary wildly, but their sum is still approximately 4.

(d) We have

$$\hat{\beta} = [9.25990778, 1.2406521, 0.99213936, 0.8182251, 0.71057517]$$

with training loss 43.279605 and validation loss 6.074517.