Homework 5

Due March 15 at 11 pm

1. (Augmented dataset) Ridge regression is equivalent to applying OLS on an expanded dataset that has additional examples. Describe these additional examples in detail. Intuitively, what effect do these additional examples have?

2. (Correlated features) Consider a regression problem where the response only depends on one feature, but we don't know it, so we incorporate an additional feature into the model that happens to be very correlated with the first feature. More specifically, let $y \in \mathbb{R}^n$ be defined by

$$y := \beta_{\text{true}} w_1 + z, \tag{1}$$

where $\beta_{\text{true}} \in \mathbb{R}$ is the true coefficient, $w_1 \in \mathbb{R}^n$ is the first feature vector, and $z \in \mathbb{R}^n$ is additive noise. The second feature vector is given by $w_2 \in \mathbb{R}^n$ and can be decomposed into

$$w_2 = \alpha w_1 + \sqrt{1-\alpha^2} w_\perp, \tag{2}$$

where $w_\perp$ is orthogonal to w_1. The vectors w_1, w_2, $w_\perp$ and z all have unit ℓ_2 norm. In addition, we assume

$$w_1^T z = 0.1, \tag{3}$$
$$w_\perp^T z = 0.1. \tag{4}$$

We fit a linear regression model to y using the feature matrix

$$X = \begin{bmatrix} w_1^T \\ w_2^T \end{bmatrix}. \tag{5}$$

   (a) What does the OLS estimator of the coefficients β_{OLS} equal to when $\alpha \to 1$? Explain what is happening.
   Hint: Use the fact that for any a, b, c, and d such that $ad \neq bc$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}. \tag{6}$$

   (b) What does the corresponding estimate of the response $y_{\text{OLS}} := X^T \beta_{\text{OLS}}$ equal to when $\alpha \to 1$? Is it collinear with the true feature w_1 when $\alpha \to 1$? Explain what is happening.

   (c) What does the ridge regression estimator of the coefficients β_{RR} equal to when $\alpha \to 1$ and the regularization parameter $\lambda > 0$ is fixed? Describe the difference with the OLS estimate.

   (d) What does the corresponding estimate of the response $y_{\text{RR}} := X^T \beta_{\text{RR}}$ equal to when $\alpha \to 1$? Is it collinear with the true feature w_1?

3. (Prior knowledge) Consider a linear regression problem where we have prior information indicating that the coefficients should be close to a certain value β_{prior}.

   (a) How can you incorporate this prior knowledge if you are using ridge regression? Write the corresponding optimization problem.

   (b) Assume that the data are generated according to a linear model $\tilde{y} := X^T\beta_{\text{true}} + \tilde{z}$, where $\beta_{true} \in \mathbb{R}^p$ and $X \in \mathbb{R}^{p\times n}$ are fixed and $\tilde{z}$ is an iid Gaussian random vector with zero mean and variance σ^2. Does the modification change the mean or the covariance matrix of the estimator? If so, report the new value.

   (c) How can you incorporate this prior knowledge if you are using gradient descent with early stopping? Write the corresponding update equation as a function of β_{prior}.

   (d) Assume that the data are generated according to the linear model described above. Does the modification change the mean or the covariance matrix of the estimator? If so, report the new value.

4. The code you will implement in this question is located in the `regress.py` file in the time folder of `hw5.zip`. Define a sequence of random variables as follows:

$$\begin{aligned}
\vec{\tilde{x}}[0] &= 1 \\
\vec{\tilde{x}}[1] &= \vec{\tilde{x}}[0] + \vec{\tilde{z}}[1] \\
\vec{\tilde{x}}[2] &= \vec{\tilde{x}}[1] + \vec{\tilde{z}}[2] \\
\vec{\tilde{x}}[3] &= \vec{\tilde{x}}[2] + \vec{\tilde{z}}[3] \\
\vec{\tilde{x}}[4] &= \vec{\tilde{x}}[3] + \vec{\tilde{z}}[4],
\end{aligned}$$

   where $\vec{\tilde{z}}[1], \vec{\tilde{z}}[2], \vec{\tilde{z}}[3], \vec{\tilde{z}}[4]$ are independent, $\vec{\tilde{z}}[1] \sim \mathcal{N}(0,1)$ and $\vec{\tilde{z}}[2], \vec{\tilde{z}}[3], \vec{\tilde{z}}[4] \sim \mathcal{N}(0, 0.01^2)$. There is a function $f : \mathbb{R}^5 \to \mathbb{R}$ of the form $f(x) = \vec{\beta}^T x$ where $\vec{\beta}$ is unknown. We are given a training sample of independent draws

$$(\vec{\tilde{x}}_1, f(\vec{\tilde{x}}_1) + \tilde{w}_1), \ldots, (\vec{\tilde{x}}_n, f(\vec{\tilde{x}}_n) + \tilde{w}_n) \in \mathbb{R}^5 \times \mathbb{R},$$

   where $\tilde{w}_i$ are iid standard normal random variables corrupting our measurements of f. Using this training data, we will estimate $\vec{\beta}$ and test our performance on a validation set drawn from the same distribution. Below we refer to the square loss function $L : \mathbb{R}^5 \times \mathbb{R}^{n\times 5} \times \mathbb{R}^n \to \mathbb{R}$ defined by

$$L(\hat{\beta}, X, y) = \sum_{i=1}^{n} (X[i,:]\hat{\beta} - y[i])^2$$

   where $X \in \mathbb{R}^{n\times 5}$ denotes a matrix of data (training or validation; each row is a data point), and y is the corresponding vector of f-values.

   (a) Using least squares (i.e., minimizing the square loss on the training set) compute an estimate for $\vec{\beta}$. Include your estimate for $\vec{\beta}$, your square loss on the training set, and your square loss on the validation set in your submission. [Hint: If computed correctly your training loss should be larger than 30 and your validation loss should be larger than 10.]

(b) Compute the singular values of the training data matrix $X \in \mathbb{R}^{n \times 5}$.

(c) The true value of $\vec{\beta}$ can be found at the top of `regress.py`. Give an explanation as to why the least squares estimates aren't close to the true $\vec{\beta}$-values.

(d) Use ridge regression to produce a new estimate of $\vec{\beta}$ and report the resulting estimate of $\vec{\beta}$, and your square loss on the training and validation sets. Here $\hat{\beta}$ should solve

$$\text{minimize}_{\vec{\eta}} \quad \|X\vec{\eta} - \vec{y}\|_2^2 + 0.5\|\vec{\eta}\|_2^2.$$

You're not required to include your code in your submission, but you are free to do so.