Recitation 6. Midterm Review

DS-GA 1013 Mathematical Tools for Data Science

1. (Recitation 5) Under what conditions will training error increase if you add a feature to your regression problem?

2. (Homework 3) Consider a dataset of n 2-dimensional data points $x_1, \ldots, x_n \in \mathbb{R}^2$. Assume that the dataset is centered. Our goal is to find a line in the 2D space that lies *closest* to the data. First, we apply PCA and consider the line in the direction of the first principal direction. Second, we fit a linear regression model where $x_i[1]$ is a feature, and $x_i[2]$ the corresponding response. Describe how each of the is different line in terms of the quantity it minimizes geometrically (e.g. sum of some distance from the points to the lines). Draw a picture with an example of a dataset where both lines are different.

3. (Homework 3) We are interested in computing the best linear estimate of the heartbeat of a fetus in the presence of strong interference in the form of the heartbeat of the baby's mother. To simplify matters, let us assume that we only want to estimate the heartbeat at a certain moment. We have available a measurement from a microphone situated near the mother's belly and another from a microphone that is away from her belly. We model the measurements as

$$\tilde{x}[1] = \tilde{b} + \tilde{m} + \tilde{z}_1 \tag{1}$$

$$\tilde{x}[2] = \tilde{m} + \tilde{z}_2, \tag{2}$$

   where $\tilde{b}$ is a random variable modeling the heartbeat of the baby, $\tilde{m}$ is a random variable modeling the heartbeat of the mother, and $\tilde{z}_1$ and $\tilde{z}_2$ model additive noise. From past data, we determine that $\tilde{b}$, $\tilde{m}$, $\tilde{z}_1$, and $\tilde{z}_2$ are all zero mean and uncorrelated with each other. The variances of $\tilde{b}$, $\tilde{z}_1$ and $\tilde{z}_2$ are equal to 1, whereas the variance of $\tilde{m}$ is much larger, it is equal to 10.

   1. Compute the best linear estimator of $\tilde{b}$ given $\tilde{x}[1]$ in terms of MSE, and the corresponding MSE. Describe in words what the estimator does.
   2. Compute the best linear estimator of $\tilde{b}$ given $\tilde{x}$ in terms of MSE, and the corresponding MSE. Describe in words what the estimator does.

4. (Recitation 4) Let $A \in \mathbb{R}^{m \times n}$. Find maximizers $\vec{x} \in \mathbb{R}^m, \vec{y} \in \mathbb{R}^n$ solving

$$\begin{aligned} \text{maximize} \quad & \vec{x}^T A \vec{y} \\ \text{subject to} \quad & \|\vec{x}\|_2 = 1, \\ & \|\vec{y}\|_2 = 1. \end{aligned}$$

   Also give the maximum value obtained. How would your answer change if the objective function was $\vec{x}^T A^{-1} \vec{y}$ assuming A is square and invertible.

5. (Recitation 5) The ridge regression estimator is given by

$$\vec{\beta}_{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T \vec{y}.$$

   Under what conditions on X is this formula valid (i.e., does the inverse exist)?

6. We are given data $X \in \mathbb{R}^{n \times d}$ and $\vec{y} \in \mathbb{R}^n$ that satisfy the linear model

$$\vec{y} = X\vec{\beta}_{\text{True}} + \sigma\vec{z}$$

with parameters $\vec{\beta}_{\text{True}} \in \mathbb{R}^d$, $\sigma > 0$ and $\vec{\tilde{z}} \sim \mathcal{N}(\vec{0}, I)$. Let $\vec{\tilde{\beta}}_{\text{RR}}$ denote the ridge regression estimator of $\vec{\beta}_{\text{True}}$ with regularization parameter λ:

$$\vec{\tilde{\beta}}_{\text{RR}} = \arg\min_{\beta} \|X\vec{\beta} - \vec{\tilde{y}}\|_2^2 + \lambda\|\vec{\beta}\|_2^2.$$

Prove that if $\lambda \geq \|X\|^2$ then

$$\left\| E\left[\vec{\tilde{\beta}}_{\text{RR}}\right] \right\|_2 \leq \frac{1}{2}\|\vec{\beta}_{\text{True}}\|_2.$$

[Hint: $E\left[\vec{\tilde{\beta}}_{\text{RR}}\right]$ is a linear function of $\vec{\beta}_{\text{True}}$. What is the norm of that function?]

7. We consider a dataset for linear regression where the training matrix of features equals

$$X := \begin{bmatrix} \vec{u}_1 & \alpha\vec{u}_1 + \tilde{\alpha}\vec{u}_2 \end{bmatrix}, \tag{30}$$

where $0 < \alpha < 1$, $\tilde{\alpha} := 1 - \alpha$, and $\vec{u}_1 \in \mathbb{R}^n$ and $\vec{u}_2 \in \mathbb{R}^n$ are orthogonal unit-norm vectors. The training response vector equals

$$\vec{y} := X\vec{\beta}_{\text{true}} + \vec{z}_{\text{noise}}, \tag{31}$$

where $\beta_{\text{true}} \in \mathbb{R}^2$ is a vector of coefficients and $\vec{z}$ is a vector of noise.

1. Express the error in the least-squares coefficient estimate

$$\vec{\beta}_{\text{LS}} := \arg\min_{\vec{\beta}} \|\vec{y} - X\vec{\beta}\|_2 \tag{32}$$

in terms of α, $\tilde{\alpha}$, $z_1 :=< \vec{u}_1, \vec{z}_{\text{noise}} >$, and $z_2 :=< \vec{u}_2, \vec{z}_{\text{noise}} >$.
Hint: Use the following expression for the inverse of a 2×2 matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}. \tag{33}$$

2. Is the coefficient error high or low when α is close to 1? Why does this happen?
3. Does the training error depend on α? (Notice that $0 < \alpha < 1$.)
4. Does the training error change if we set $\alpha = 1$? If so, does it increase or decrease?
5. If we model the noise as a random vector $\tilde{v}z$ with independent Gaussian entries with mean zero and standard deviation σ, what is the variance of $\tilde{z}_1 :=< \vec{u}_1, \tilde{v}z >$ and $\tilde{z}_2 :=< \alpha\vec{u}_1 + \tilde{\alpha}\vec{u}_2, \tilde{v}z >$?

8. Suppose $X \in \mathbb{R}^{n \times d}$, $\vec{\beta} \in \mathbb{R}^d$, and $\vec{\tilde{y}} \in \mathbb{R}^n$ is defined by

$$\vec{\tilde{y}} = X\vec{\beta} + \vec{\tilde{z}},$$

where $\vec{\tilde{z}} \in \mathcal{N}(\vec{0}, D)$ and $D \in \mathbb{R}^{n \times n}$ is diagonal and known. If the diagonal of D is not constant, the standard least squares estimator for $\vec{\beta}$ is no longer optimal (in the sense of being the linear unbiased estimator with the minimum variance in every direction). Find an unbiased linear estimator of $\vec{\beta}$ that is optimal in this sense. [Hint: To obtain optimality, transform the data to a case where the least squares estimator is optimal.]