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?

Solution: Never Increases.

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.

Solution: Look at HW3 solutions

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.

Solution:

1. By independence

$$\begin{aligned} \mathrm{Var}(\tilde{x}[1]) &= \mathrm{Var}(\tilde{b}) + \mathrm{Var}(\tilde{m}) + \mathrm{Var}(\tilde{z}_1) && (3)\\ &= 12, && (4)\\ \mathrm{Cov}(\tilde{b}\tilde{x}[1]) &= 1 && (5) \end{aligned}$$

so

$$\Sigma_{\tilde{x}[1]} = 12 \tag{6}$$

$$\Sigma_{\tilde{b}\tilde{x}[1]} = 1 \tag{7}$$

and by Theorem 2.3 in the notes the estimate equals

$$\hat{b}(\tilde{x}[1]) = \frac{\tilde{x}[1]}{12}, \tag{8}$$

and the corresponding MSE equals

$$E((\hat{b}(\tilde{x}[1]) - \tilde{b})^2) = \mathrm{Var}(\tilde{b}) - \frac{1}{12} \tag{9}$$

$$= 0.92. \tag{10}$$

The estimate just shrinks the signal.

2. By independence

$$\mathrm{Var}(\tilde{x}[1]) = \mathrm{Var}(\tilde{b}) + \mathrm{Var}(\tilde{m}) + \mathrm{Var}(\tilde{z}_1) \tag{11}$$

$$= 12, \tag{12}$$

$$\mathrm{Var}(\tilde{x}[2]) = \mathrm{Var}(\tilde{m}) + \mathrm{Var}(\tilde{z}_2) \tag{13}$$

$$= 11, \tag{14}$$

$$\mathrm{Cov}(\tilde{x}[1]\tilde{x}[2]) = E(\tilde{m}^2) \tag{15}$$

$$= 10, \tag{16}$$

$$\mathrm{Cov}(\tilde{b}\tilde{x}[1]) = 1 \tag{17}$$

$$\mathrm{Cov}(\tilde{b}\tilde{x}[2]) = 0 \tag{18}$$

$$\tag{19}$$

so

$$\Sigma_{\tilde{x}} = \begin{bmatrix} 12 & 10 \\ 10 & 11 \end{bmatrix} \tag{20}$$

$$\Sigma_{\tilde{b}\tilde{x}} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{21}$$

and by Theorem 2.3 in the notes the estimate equals

$$\hat{b}(\tilde{x}) = \tilde{x}^T \begin{bmatrix} 12 & 10 \\ 10 & 11 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{22}$$

$$= \frac{11\tilde{x}[1] - 10\tilde{x}[2]}{32}, \tag{23}$$

and the corresponding MSE equals

$$E((\hat{b}(\tilde{x}) - \tilde{b})^2) = \mathrm{Var}(\tilde{b}) - \begin{bmatrix} 1 \\ 0 \end{bmatrix}^T \begin{bmatrix} 12 & 10 \\ 10 & 11 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{24}$$

$$= 1 - \frac{11}{32} \tag{25}$$

$$= 0.66. \tag{26}$$

The estimate approximately cancels out the signal from the mum, in a way that is optimal with respect to MSE.

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

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

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.

Solution: Compute the SVD $A = USV^T$. The maximizers are given by $\vec{x} = U_{:,1}$ and $\vec{y} = V_{:,1}$ with maximum value given by $\sigma_1 = S_{11}$.

To see this is the maximum, note that

$$\vec{x}^T A \vec{y} \leq \|\vec{x}\| \|A\vec{y}\| = \|A\vec{y}\| \leq S_{11},$$

by Cauchy-Schwarz.

$A^{-1} = VS^{-1}U^T$. Therefore, The maximizers are given by $\vec{y} = U_{:,n}$ and $\vec{x} = V_{:,n}$ with maximum value given by $1/\sigma_n = S_{nn}$.

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)?

Solution: It always exists since $X^T X$ is positive semidefinite and λI is positive definite.

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

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

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?]

Solution: Note that

$$\begin{aligned} E\left[\vec{\tilde{\beta}}_{\text{RR}}\right] &= E\left[(X^T X + \lambda I)^{-1} X^T \vec{\tilde{y}}\right] && (27) \\ &= E\left[(X^T X + \lambda I)^{-1} X^T (X\vec{\beta}_{\text{True}} + \sigma\vec{\tilde{z}})\right] && (28) \\ &= (X^T X + \lambda I)^{-1} X^T X \vec{\beta}_{\text{True}}. && (29) \end{aligned}$$

If X has SVD $X = USV^T$ then $X^TX = VS^2V^T$ giving

$$\|(X^TX + \lambda I)^{-1}X^TX\| = \max_j \frac{S_{jj}^2}{S_{jj}^2 + \lambda} \leq \frac{1}{2}$$

proving the result.

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 $\tilde{z}_2 :=< \alpha\vec{u}_1 + \tilde{\alpha}\vec{u}_2, \tilde{v}z >$?

Solution:

1. We have

$$X^TX = \begin{bmatrix} 1 & \alpha \\ \alpha & \alpha^2 + \tilde{\alpha}^2 \end{bmatrix}. \tag{34}$$

By Theorem 4.6 in the lecture notes on the SVD and the formula for the matrix inverse,

$$\begin{aligned} \vec{\beta}_{\text{LS}} - \vec{\beta}_{\text{true}} &= \left(X^TX\right)^{-1}X^T\vec{z}_{\text{noise}} && (35)\\ &= \frac{1}{\tilde{\alpha}^2}\begin{bmatrix} \alpha^2 + \tilde{\alpha}^2 & -\alpha \\ -\alpha & 1 \end{bmatrix}\begin{bmatrix} z_1 \\ \alpha z_1 + \tilde{\alpha}z_2 \end{bmatrix} && (36)\\ &= \frac{1}{\tilde{\alpha}^2}\begin{bmatrix} \tilde{\alpha}^2 z_1 - \alpha\tilde{\alpha}z_2 \\ \tilde{\alpha}z_2 \end{bmatrix} && (37)\\ &= \begin{bmatrix} z_1 - \frac{\alpha z_2}{\tilde{\alpha}} \\ \frac{z_2}{\tilde{\alpha}} \end{bmatrix}. && (38) \end{aligned}$$

2. When α is close to one, the coefficient error becomes very large. This occurs because the columns of X are highly correlated, so there is a vector $\vec{v}$ (approximately $\vec{e}_1 - \vec{e}_2$) that is nearly in the nullspace of X. When fitting the noise, the resulting coefficients will have an correspondingly large error component in the direction of $\vec{v}$,

3. By Lemma 4.4 in the Linear Regression notes, the training error only depends on the column space of X, which does not change with the value of α as long as it is not equal to one.

4. By Lemma 4.4 in the Linear Regression notes, the training error is equal to the projection of the noise onto the orthogonal complement of the column space of X. If α is set to one then the column space only contains $\vec{u}_1$, as opposed to $\vec{u}_1$ and $\vec{u_2}$, so the training error will increase.

5. By Lemma 4.4 in the PCA notes, the variance equals

$$\begin{aligned}
\mathrm{Var}(\tilde{z}_1) &= \mathrm{Var}(\vec{u}_1^T \tilde{v} z) && (39)\\
&= \vec{u}_1^T \sigma^2 I \vec{u}_1 && (40)\\
&= \sigma^2, && (41)\\
\mathrm{Var}(\tilde{z}_2) &= \mathrm{Var}((\alpha\vec{u}_1 + \tilde{\alpha}\vec{u}_2)^T \tilde{v} z) && (42)\\
&= (\alpha\vec{u}_1 + \tilde{\alpha}\vec{u}_2)^T \sigma^2 I (\alpha\vec{u}_1 + \tilde{\alpha}\vec{u}_2) && (43)\\
&= \sigma^2(\alpha^2 + \tilde{\alpha}^2). && (44)
\end{aligned}$$

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.]

Solution: As $D^{-1/2}\vec{\tilde{x}} \sim \mathcal{N}(\vec{0}, I)$ we see that

$$D^{-1/2}\vec{\tilde{y}} = D^{-1/2}X\vec{\beta} + D^{-1/2}\vec{\tilde{z}}$$

is a standard (homoscedastic) least squares problem with optimal linear unbiased estimator

$$((D^{-1/2}X)^T(D^{-1/2}X))^{-1}(D^{-1/2}X)^T D^{-1/2}\vec{\tilde{y}} = (X^T D^{-1} X)^{-1} X^T D^{-1}\vec{\tilde{y}}$$

for $\vec{\beta}$ (see Homework 4 question 3). Letting $W = (X^T D^{-1} X)^{-1} X^T D^{-1}$, this implies that

$$E[(W\vec{\tilde{y}})(W\vec{\tilde{y}})^T] \preceq E[(CD^{-1/2}\vec{\tilde{y}})(CD^{-1/2}\vec{\tilde{y}})^T]$$

for any unbiased estimator $CD^{-1/2}\vec{\tilde{y}}$ of $\vec{\beta}$ where $C \in \mathbb{R}^{d\times n}$ and $\preceq$ is the semidefinite ordering. But any unbiased estimator $C'\vec{\tilde{y}}$ with of $\vec{\beta}$ with $C' \in \mathbb{R}^{d\times n}$ can be written as $C'D^{1/2}D^{-1/2}\vec{\tilde{y}}$ giving

$$E[(W\vec{\tilde{y}})(W\vec{\tilde{y}})^T] \preceq E[(C'D^{1/2}D^{-1/2}\vec{\tilde{y}})(C'D^{1/2}D^{-1/2}\vec{\tilde{y}})^T] = E[(C'\vec{\tilde{y}})(C'\vec{\tilde{y}})^T]$$

completing the proof.