DSC 40A

DSC 40A Midterm Review

Harshi, Nick, Yosen, & Zoe

Midterm Logistics

Format

80 minutes, on paper, no calculators or electronics

- Allowed one two-sided index card (handwritten)

Content

Lectures 1-9, Homeworks 1-4, Groupworks 1-4, feature engineering, and transformations

- You can study with the practice site!

So... When is it?

Midterm is Tuesday, May 7th in **your** section

- You will receive a randomized seat assignment over the weekend

Index Card in Depth

Example Card!

- Posted on Edstem
- On the course website

- We will not print it for you
- You cannot print it either!
- However, you can reference it!
  - And we encourage it!

Important Notes

- You can write on both sides of the index card
- You must write the index card by hand
  - No typing or writing with a tablet
- You need to find/buy a 4 inch by 6 inch index card
  - OK if 3 in by 5 in
  - Worst case you cut paper to that size

Problem 1.1 from FA21 Midterm

King Triton just made an Instagram account and has been keeping track of the number of likes his posts have received so far.

His first 7 posts have received a mean of 16 likes; the specific like counts in sorted order are

$$8, 12, 12, 15, 18, 20, 27$$

King Triton wants to predict the number of likes his next post will receive, using a constant prediction rule h. For each loss function $L(h, y)$, determine the constant prediction h^* that minimizes empirical risk. If you believe there are multiplie minimizers, specify them all. If you believe you need more information to answer the question or that there is no minimizer, state that clearly. **Give a brief justification for each answer.**

$$L(h, y) = |y - h|$$

empirical risk w/ absolute loss : median :

$\boxed{15}$

Nick

Problem 1.2 from FA21 Midterm

King Triton just made an Instagram account and has been keeping track of the number of likes his posts have received so far.

His first 7 posts have received a mean of 16 likes; the specific like counts in sorted order are

$$8, 12, 12, 15, 18, 20, 27$$

King Triton wants to predict the number of likes his next post will receive, using a constant prediction rule h. For each loss function $L(h, y)$, determine the constant prediction h^* that minimizes empirical risk. If you believe there are multiplie minimizers, specify them all. If you believe you need more information to answer the question or that there is no minimizer, state that clearly. **Give a brief justification for each answer.**

$$L(h, y) = (y - h)^2$$

empirical risk for square loss: minimized by mean.

16

Nick

Problem 1.3 from FA21 Midterm

King Triton just made an Instagram account and has been keeping track of the number of likes his posts have received so far.

His first 7 posts have received a mean of 16 likes; the specific like counts in sorted order are

$$8, 12, 12, 15, 18, 20, 27$$

King Triton wants to predict the number of likes his next post will receive, using a constant prediction rule h. For each loss function $L(h, y)$, determine the constant prediction h^* that minimizes empirical risk. If you believe there are multiplie minimizers, specify them all. If you believe you need more information to answer the question or that there is no minimizer, state that clearly. **Give a brief justification for each answer.**

$$L(h, y) = 4(y - h)^2$$

empirical risk for scaled squared error

: still minimized by mean!

$\boxed{16}$

Nick

Problem 1.4 from FA21 Midterm

King Triton just made an Instagram account and has been keeping track of the number of likes his posts have received so far.

His first 7 posts have received a mean of 16 likes; the specific like counts in sorted order are

$$8, 12, \underline{12, 12}, 15, 18, 20, 27$$

King Triton wants to predict the number of likes his next post will receive, using a constant prediction rule h. For each loss function $L(h, y)$, determine the constant prediction h^* that minimizes empirical risk. If you believe there are multiplie minimizers, specify them all. If you believe you need more information to answer the question or that there is no minimizer, state that clearly. **Give a brief justification for each answer.**

$$L(h, y) = \begin{cases} 0 & h = y \\ 100 & h \neq y \end{cases}$$

empirical risk for scaled 0-1 loss : minimized by mode.

12

Nick

Problem 1.5 from FA21 Midterm

King Triton just made an Instagram account and has been keeping track of the number of likes his posts have received so far.

His first 7 posts have received a mean of 16 likes; the specific like counts in sorted order are

$$8, 12, 12, 15, 18, 20, 27$$

King Triton wants to predict the number of likes his next post will receive, using a constant prediction rule h. For each loss function $L(h, y)$, determine the constant prediction h^* that minimizes empirical risk. If you believe there are multiplie minimizers, specify them all. If you believe you need more information to answer the question or that there is no minimizer, state that clearly. **Give a brief justification for each answer.**

$$L(h, y) = (3y - 4h)^2$$

empirical risk w/ this loss? complicated...

$(3y-4h)^2$

$(3(y-\frac{4}{3}h))^2$

$3^2(y-\frac{4}{3}h)^2 \to \underbrace{9}_{\text{does not matter}}(y-\underbrace{\tfrac{4}{3}h}_{\text{want this to be } \bar{x}})^2$

$\frac{4}{3}h = \text{mean}$

$h^* = \frac{3}{4}\text{mean}$

$\boxed{h = 12}$

Nick

Problem 1.6 from FA21 Midterm

King Triton just made an Instagram account and has been keeping track of the number of likes his posts have received so far.

His first 7 posts have received a mean of 16 likes; the specific like counts in sorted order are

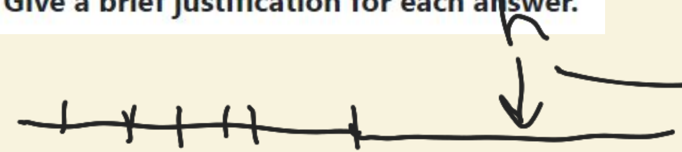
$$8, 12, 12, 15, 18, 20, 27$$

King Triton wants to predict the number of likes his next post will receive, using a constant prediction rule h. For each loss function $L(h, y)$, determine the constant prediction h^* that minimizes empirical risk. If you believe there are multiplie minimizers, specify them all. If you believe you need more information to answer the question or that there is no minimizer, state that clearly. **Give a brief justification for each answer.**

$$L(h, y) = \underline{(y - h)}^3$$

$\boxed{\text{No minimizer}}$

as h increases, the loss approaches $-\infty$



Nick

Problem 1f from Discussion 4

Problem 1.

The table below shows the softness and color of several different avocados, which we want to use to predict their ripeness. Each variable is measured on a scale of 1 to 5. For softness, 5 is softest, for color, 5 is darkest, and for ripeness, 5 is ripest.

Avocado	Softness	Color	Ripeness
1	3	4	2.5
2	1	2	2
3	4	5	5

Suppose we have decided on the following hypothesis function: given an avocado's softness and color, we average these numbers to produce a predicted ripeness.

f) Write down the *design matrix*, X.

$n = 3$

$d+1 = 3$

$\boxed{3 \times 3}$

$$X = \begin{bmatrix} 1 & 3 & 4 \\ 1 & 1 & 2 \\ 1 & 4 & 5 \end{bmatrix}$$

3×1

$$y = \begin{bmatrix} 2.5 \\ 2 \\ 5 \end{bmatrix}$$

Harshi

Problem 1g from Discussion 4

Problem 1.

The table below shows the softness and color of several different avocados, which we want to use to predict their ripeness. Each variable is measured on a scale of 1 to 5. For softness, 5 is softest, for color, 5 is darkest, and for ripeness, 5 is ripest.

Avocado	Softness	Color	Ripeness
1	3	4	2.5
2	1	2	2
3	4	5	5

$H(x) = w_0 + w_1 x_1 + \ldots + w_d x_d$

$f(x) = 0 + \underbrace{\tfrac{1}{2} x_1}_{\text{softness}} + \underbrace{\tfrac{1}{2} x_2}_{\text{color}}$

Suppose we have decided on the following hypothesis function: given an avocado's softness and color, we average these numbers to produce a predicted ripeness.

g) Write down the *parameter vector*, $\vec{w}$ that corresponds to this particular choice of hypothesis function. The parameter vector should have three components, one for the bias, and one for each of the

w $(d+1) \times 1$

X $n \times (d+1)$

$$w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0.5 \\ 0.5 \end{bmatrix}$$

3×1

a hypothesis func, not w^*

prediction $= \dfrac{\text{softness} + \text{color}}{2}$

$= \frac{1}{2}\text{softness} + \frac{1}{2}\text{color}$

Harshi

Problem 1h from Discussion 4

X 3×3

w 3×1

3×3 3×1

x w

3×1

Problem 1.

The table below shows the softness and color of several different avocados, which we want to use to predict their ripeness. Each variable is measured on a scale of 1 to 5. For softness, 5 is softest, for color, 5 is darkest, and for ripeness, 5 is ripest.

$X = n \times (d+1)$

$w = (d+1) \times 1$

prod : $n \times 1$

Avocado	Softness	Color	Ripeness
1	3	4	2.5
2	1	2	2
3	4	5	5

$H(x)$:

prediction: $\begin{bmatrix} 3.5 \\ 1.5 \\ 4.5 \end{bmatrix}$

Suppose we have decided on the following hypothesis function: given an avocado's softness and color, we average these numbers to produce a predicted ripeness.

h) Check that the entries of $X\vec{w}$ are the predicted ripenesses you found above.

3×3

$$X = \begin{bmatrix} 1 & 3 & 4 \\ 1 & 1 & 2 \\ 1 & 4 & 5 \end{bmatrix} \qquad w = \begin{bmatrix} 0 \\ 0.5 \\ 0.5 \end{bmatrix}$$

Av1 $1\times 0 + 3\times 0.5 + 4\times 0.5 = \frac{3+4}{2}$

Av2 $1\times 0 + 1\times 0.5 + 2\times 0.5 = \frac{1+2}{2}$

Av3 $1\times 0 + 4\times 0.5 + 5\times 0.5 = \frac{4+5}{2}$

$$= \begin{bmatrix} 3.5 \\ 1.5 \\ 4.5 \end{bmatrix}$$

3×1

Harshi

Problem 4.1 from WI22 Midterm

Consider the dataset shown below.

$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	y
0	6	8	-5
3	4	5	7
5	-1	-3	4
0	2	1	2

We want to use multiple regression to fit a prediction rule of the form

$$H(x^{(1)}, x^{(2)}, x^{(3)}) = w_0 + w_1 x^{(1)} x^{(3)} + w_2 (x^{(2)} - x^{(3)})^2.$$

Write down the design matrix X and observation vector $\vec{y}$ for this scenario. No justification needed.

$d = 2$

$d+1 = 3$

X 4×3

$$X = \begin{bmatrix} 1 & 0 & 4 \\ 1 & 15 & 1 \\ 1 & -15 & 4 \\ 1 & 0 & 1 \end{bmatrix} \qquad y = \begin{bmatrix} -5 \\ 7 \\ 4 \\ 2 \end{bmatrix}$$

$w_0 + w_1 b_1 + w_2 b_2$

$b_1 \downarrow x^1 x^3$ $\quad b_2 \downarrow (x^2 - x^3)^2$

Harshi

Problem 4.2 from WI22 Midterm

For the X and $\vec{y}$ that you have written down, let $\vec{w}$ be the optimal parameter vector, which comes from solving the normal equations $X^TX\vec{w} = X^T\vec{y}$. Let $\vec{e} = \vec{y} - X\vec{w}$ be the error vector, and let e_i be the ith component of this error vector. Show that

$$4e_1 + e_2 + 4e_3 + e_4 = 0.$$

$$\boxed{X^Te = 0}$$

$$X = \begin{bmatrix} 1 & 0 & 4 \\ 1 & 15 & 1 \\ 1 & -15 & 4 \\ 1 & 0 & 1 \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 15 & -15 & 0 \\ 4 & 1 & 4 & 1 \end{bmatrix}$$

$$X^Te = 0$$

$$X^Te = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 15 & -15 & 0 \\ 4 & 1 & 4 & 1 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{bmatrix} = 0$$

$$= \begin{bmatrix} e_1 + e_2 + e_3 + e_4 \\ 15e_2 - 15e_3 \\ 4e_1 + e_2 + 4e_3 + e_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$\star \rightarrow 4e_1 + e_2 + 4e_3 + e_4 = 0$

Harshi

$$\underbrace{X^T X w = X^T y}$$

$$\vec{e} = \vec{y} - X\vec{w}$$

$$\underbrace{X^T e} = 0 \qquad X \text{ orth } e$$

$$0 = X^T y - X^T X w$$

$$0 = X^T (\underline{y - X w})$$

$$0 = X^T e$$

Problem 2 Background from Homework 3

In the National Football League (NFL), the highest paid position is the quarterback. The job of the quarterback on an (American) football team is, among other things, to throw (or "pass") the football to other players, who then score "touchdowns." Each time a quarterback throws the ball to another player and that other player scores, we say the quarterback made a "touchdown pass."

Suppose that we have access to a dataset containing information about a random sample of 50 quarterbacks. For each quarterback, we have the number of touchdown passes they threw, along with their salary in 2023. In the 2023 dataset, the number of touchdown passes for all quarterbacks has a mean of 17 and a standard deviation of 3.

We minimize mean squared error to fit a linear hypothesis function, $H(x) = w_0 + w_1 x$, to this dataset. We will use the hypothesis function to help other players predict their 2023 salary in millions of dollars (y) based on their number of touchdown passes (x).

Yosen

Problem 2a from Homework 3

a) CJ Stroud was one of the quarterbacks in our 2023 dataset. Suppose that in 2023, he had 26 touchdown passes and his salary was only 10 million, the smallest salary in our sample.

In 2024, Stroud signed a new contract based on his performance. In 2024, he again threw 26 touchdowns, but his salary shot up to 50 million!

Suppose we create two linear hypothesis functions, one using the dataset from 2023 when Stroud had a salary of 10 million and another using the dataset from 2024 when Stroud had a salary of 50 million. Assume that all other players threw the same amount of touchdowns and had the same salary in both datasets. **That is, only this one data point is different between these two datasets.**

Suppose the optimal slope and intercept fit on the first dataset (2023) are w_1^* and w_0^*, respectively, and the optimal slope and intercept fit on the second dataset (2024) are w_1' and w_0', respectively.

What is the difference between the new slope and the old slope? **That is, what is $w_1' - w_1^*$?** The answer you get should be a number with no variables.

Note: Since we want to salary in millions of dollars, use 10 instead of 10,000,000 for Stroud's salary in 2023.

$$w_1^* = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Yosen

$$w_1^* = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Stroud: player j

$$w_1^* = \frac{\sum_{i \neq j}^{n}(x_i - \overline{x})y_i + (x_j - \overline{x})y_j}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$w_1^{*\prime} = \frac{\sum_{i \neq j}^{n}(x_i - \overline{x})y_i + (x_j - \overline{x})y_j'}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$w_1^{*\prime} - w_1^* = \frac{\left(\cancel{\sum_{i \neq j}^{n}(x_i - \overline{x})y_i} + (x_j - \overline{x})y_j'\right) - \left(\cancel{\sum_{i \neq j}^{n}(x_i - \overline{x})y_i} + (x_j - \overline{x})y_j\right)}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$= \frac{(x_j - \overline{x})(y_j' - y_j)}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$= \frac{(x_j - \bar{x})(y_j' - y_j)}{n\sigma^2}$$

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

$$x_j = 26$$

$$= \frac{(26-17)(50-10)}{50(3)^2}$$

$$= \frac{9 \cdot 40}{50 \cdot 9}$$

$$= \frac{4}{5}$$

Problem 2b from Homework 3

b) Let $H^*(x)$ be the linear hypothesis function fit on the 2023 dataset (i.e. $H^*(x) = w_0^* + w_1^* x$) and $H'(x)$ be the linear hypothesis function fit on the 2024 dataset (i.e. $H'(x) = w_0' + w_1' x$).

Consider two other players, Davis Mills and Tom Brady, neither of whom were part of our original sample in 2023. Suppose that in 2021, Mills had 24 touchdowns and Brady had 43 touchdowns.

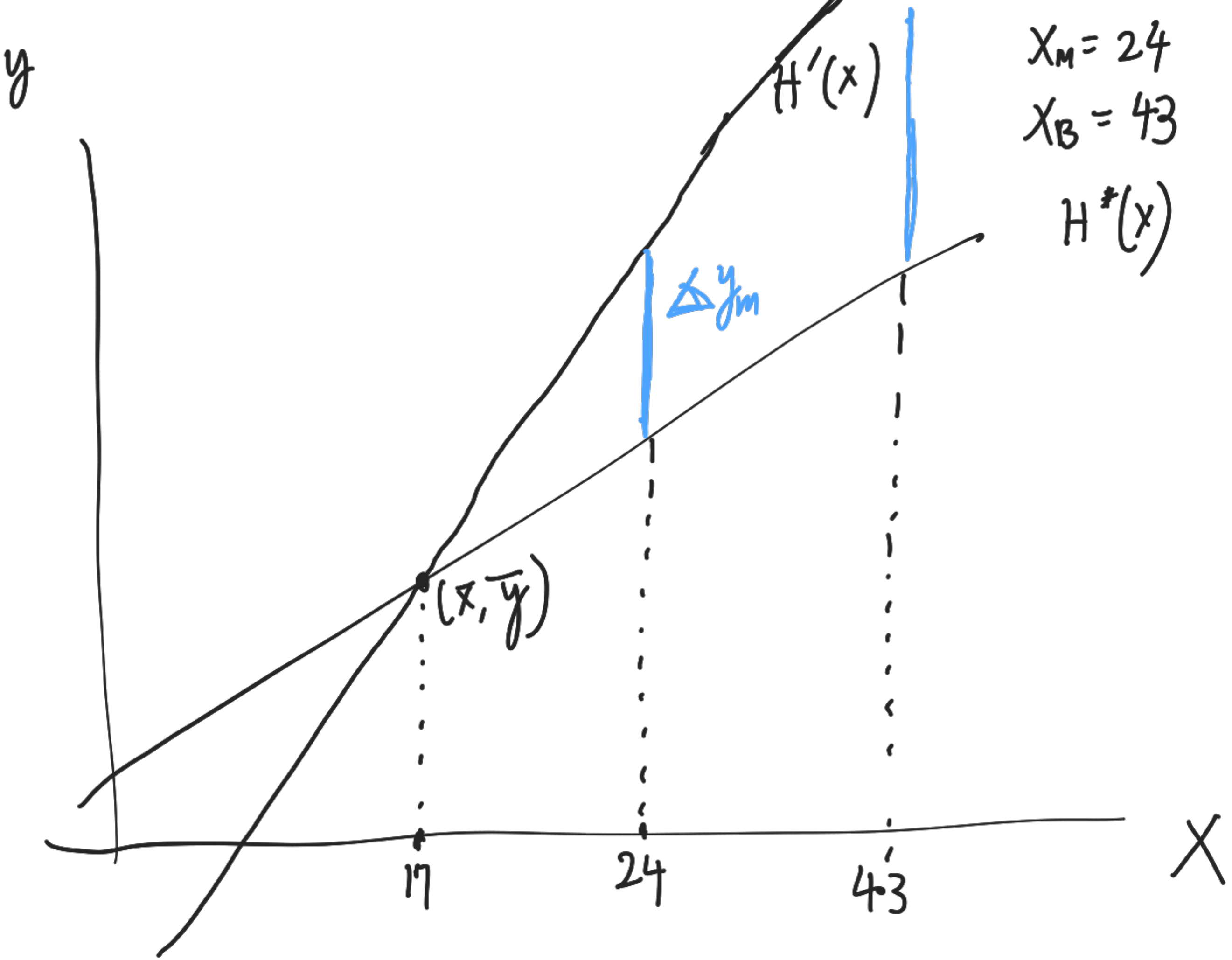
Both Mills and Brady want to try and use one of our linear hypothesis functions to predict their salary for next year.

Suppose they both first use $H^*(x)$ to determine their predicted yields as per the first rule (when Stroud had a salary of 10 million). Then, they both then use $H'(x)$ to determine predicted yields as per the second rule (when Stroud had a salary of 50 million).

Whose prediction changed more by switching from $H^*(x)$ to $H'(x)$ – Mills' or Brady's?

Δy_B

Yosen



$X_M = 24$
$X_B = 43$
y
$H'(x)$
$H^*(x)$
Δy_m
$(\bar{x}, \bar{y})$
17
24
43
X

Problem 2c from Homework 3

c) In this problem, we'll consider how our answer to part (b) might have been different if Stroud had fewer touchdowns in both 2023 and 2024.

- Suppose Stroud instead had 17 touchdowns in both 2023 and 2024. If his salary increased from 2023 to 2024, and everyone else's data stayed the same, which slope would be larger: $H^*(x)$ or $H'(x)$?
- Suppose Stroud instead had 10 touchdowns in both 2023 and 2024. If his salary increased from 2023 to 2024, and everyone else's data stayed the same, which slope would be larger: $H^*(x)$ or $H'(x)$?

You don't have to actually calculate the new slopes, but given the information in the problem and the work you've already done, you should be able to answer the questions and give brief justification.

Yosen

Stroud: player j

$$w_1^* = \frac{\sum_{i\neq j}^{n} (x_i-\bar{x})y_i + (x_j-\bar{x})y_j}{\sum_{i=1}^{n} (x_i-\bar{x})^2} \qquad w_1^{*\prime} = \frac{\sum_{i\neq j}^{n} (x_i-\bar{x})y_i + (x_j-\bar{x})y_j'}{\sum (x_i-\bar{x})^2}$$

$$w_1^{*\prime} - w_1^* = \frac{\left(\cancel{\sum_{i\neq j}^{n} (x_i-\bar{x})y_i} + (x_j-\bar{x})y_j'\right) - \left(\cancel{\sum_{i\neq j}^{n} (x_i-\bar{x})y_i} + (x_j-\bar{x})y_j\right)}{\sum_{i=1}^{n} (x_i-\bar{x})^2}$$

$$= \frac{(x_j-\bar{x})(y_j'-y_j)}{\sum_{i=1}^{n} (x_i-\bar{x})^2}$$

1. $x_j' = 17, \quad x_j = 10 \qquad x_j - \bar{x}$

$$= \frac{(\overset{10}{\cdot} - \overset{17}{17})(\overset{50-10}{y_j' - y_j})}{\boxed{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \qquad \boxed{\sum_{i=1}^{n} (x_i - \bar{x})^2} \to n\sigma^2$$

2. $x_j' = 10$

5 min Break Time!

Problem 4.1 from FA22 Midterm

Consider a dataset that consists of $y_1, \cdots, y_n$. In class, we used calculus to minimize mean squared error, $R_{sq}(h) = \frac{1}{n}\sum_{i=1}^{n}(h - y_i)^2$. In this problem, we want you to apply the same approach to a slightly different loss function defined below:

$$L_{\text{midterm}}(y, h) = (\alpha y - h)^2 + \lambda h$$

Write down the empiricial risk $R_{\text{midterm}}(h)$ by using the above loss function.

$\star\ R(h) = \frac{1}{n}\sum_{i=1}^{n} L(y, h)$ $\quad R(h) = \frac{1}{n}\sum_{i=1}^{n}\left[(\alpha y - h)^2 + \lambda h\right]$

$R(h) = \frac{1}{n}\sum_{i=1}^{n}(\alpha y - h)^2 + \left(\frac{\lambda h \cancel{n}}{\cancel{n}}\right) \rightarrow \frac{1}{n}\sum_{i=1}^{n}(\alpha y - h)^2 + \lambda h$

Zoe

Problem 4.2 from FA22 Midterm

Consider a dataset that consists of $y_1, \cdots, y_n$. In class, we used calculus to minimize mean squared error, $R_{sq}(h) = \frac{1}{n}\sum_{i=1}^{n}(h - y_i)^2$. In this problem, we want you to apply the same approach to a slightly different loss function defined below:

$$L_{\text{midterm}}(y, h) = \underbrace{(\alpha y - h)^2} + \lambda h$$

The mean of dataset is $\bar{y}$, i.e. $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Find h^* that minimizes $R_{\text{midterm}}(h)$ using calculus. Your result should be in terms of $\bar{y}$, α and λ.

$$\frac{d}{dh}\left[f(g(x))\right] = f'(g(x))\,g'(x)$$

$f(x) = x^2 \quad f'(x) = 2x$

$(\alpha y - h)^2$

$g(x) = \alpha y - h \quad g'(x) = -1$

$$R(h) = \frac{1}{n}\sum_{i=1}^{n}(\alpha y - h)^2 + \lambda h$$

$$-\frac{2}{n}\sum_{i=1}^{n}(\alpha y_i - h) + \lambda$$

$$-2\alpha\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) + \frac{2}{\cancel{n}}(h\cancel{n}) + \lambda$$

Zoe

derivative
$R(h) = -2\alpha\bar{y} + 2h + \lambda$

$0 = -2\alpha y + 2h + \lambda$

$\dfrac{2\alpha y - \lambda = 2h}{2} \rightarrow h^* = \alpha\bar{y} - \dfrac{\lambda}{2}$

6.2 $\quad \vec{w}^* = \begin{array}{c} \begin{bmatrix} 2000 \\ 10\,000 \\ -1000 \end{bmatrix} \end{array} \begin{array}{l} w_0 \\ w_1 \\ w_2 \end{array}$ carats: 0.65
length: 4 cm

predicted price $= w_0 + w_1(\text{carat}) + w_2(\text{length})$

$2000 + 10\,000(\overset{\text{carat}}{0.65}) + (-1000)(\overset{\text{length}}{4})$

$2000 + 6500 - 4000 = 4500$

Problem 6.1 from FA21 Midterm

Billy's aunt owns a jewellery store, and gives him data on 5000 of the diamonds in her store. For each diamond, we have:

- **carat**: the weight of the diamond, in carats
- **length**: the length of the diamond, in centimeters
- **width**: the width of the diamond, in centimeters
- **price**: the value of the diamond, in dollars

The first 5 rows of the 5000-row dataset are shown below:

	carat	length	width	price
1	0.40	4.81	4.76	1323
2	1.04	6.58	6.53	5102
3	0.40	4.74	4.76	696
4	0.40	4.67	4.65	798
5	0.50	4.90	4.95	987

Billy has enlisted our help in predicting the price of a diamond given various other features.

Suppose we want to fit a linear prediction rule that uses two features, carat and length, to predict price. Specifically, our prediction rule will be of the form

$$\text{predicted price} = w_0 + w_1 \cdot \overset{x_1}{\underline{\text{carat}}} + w_2 \cdot \overset{x_2}{\underline{\text{length}}}$$

each row $= [\underset{w_0}{1} \quad \underline{\text{carat}} \quad \text{length}]$

We will use least squares to find $\vec{w}^* = \begin{bmatrix} w_0^* \\ w_1^* \\ w_2^* \end{bmatrix}$.

Write out the first 5 rows of the design matrix, X. Your matrix should not have any variables in it.

$$X = \begin{bmatrix} 1 & 0.4 & 4.81 \\ 1 & 1.04 & 6.58 \\ 1 & 0.4 & 4.74 \\ 1 & 0.4 & 4.67 \\ 1 & 0.5 & 4.90 \end{bmatrix}_{r \times c}$$

$r=5 \quad c=3$

$(d+1)=3$

$d=2$

Zoe

Problem 6.3 from FA21 Midterm

Billy's aunt owns a jewellery store, and gives him data on 5000 of the diamonds in her store. For each diamond, we have:

- **carat**: the weight of the diamond, in carats
- **length**: the length of the diamond, in centimeters
- **width**: the width of the diamond, in centimeters
- **price**: the value of the diamond, in dollars

The first 5 rows of the 5000-row dataset are shown below:

carat	length	width	price
0.40	4.81	4.76	1323
1.04	6.58	6.53	5102
0.40	4.74	4.76	696
0.40	4.67	4.65	798
0.50	4.90	4.95	987

Billy has enlisted our help in predicting the price of a diamond given various other features.

Suppose $\vec{e} = \begin{bmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{bmatrix}$ is the error/residual vector, defined as

$$\vec{e} = \vec{y} - X\vec{w}^*$$

where $\vec{y}$ is the observation vector containing the prices for each diamond.

For each of the following quantities, state whether they are guaranteed to be equal to 0 the scalar, $\vec{0}$ the vector of all 0s, or neither. No justification is necessary.

- $\sum_{i=1}^{n} e_i$
- $||\vec{y} - X\vec{w}^*||^2$
- $X^T X\vec{w}^*$
- $2X^T X\vec{w}^* - 2X^T\vec{y}$

Zoe

$\sum_{i=1}^{n} e_i$

↑ sum of entries

2b

guaranteed to equal 0

$\vec{e} = \vec{y} - X\vec{w}$

$\|\vec{e}\|^2$ — length of error vector

$||\vec{y} - X\vec{w}^*||^2$

$2 - 2 = 0$

└ zero because same val

$\vec{y} = X\vec{w}^* = 0$

actual pred

Not guaranteed if $\vec{y} \neq X\vec{w}^*$

$X^T X \vec{w}^*$

$\vec{w}^* = (X^T X)^{-1} X^T y$

Not guaranteed

$\underbrace{X^T}X \vec{w} = X^T \vec{y}$

$\vec{y}$ orthogonal to every column of X

$X^T X \vec{w} = 0$?

$2X^T X \vec{w}^* - 2X^T \vec{y}$

hint: $X^T X \vec{w} = X^T \vec{y}$

$$\underset{3\times 1}{\begin{bmatrix} \\ \\ \end{bmatrix}} - \underset{3\times 1}{\begin{bmatrix} \\ \\ \end{bmatrix}} = \underset{3\times 1}{\begin{bmatrix} \\ \\ \end{bmatrix}}$$

$2(X^T X \vec{w} - X^T \vec{y} = \vec{0})$

yes! guaranteed to be $\vec{0}$

$2X^T X \vec{w} - 2X^T \vec{y} = \vec{0}$

$$\underset{d+1 \times \cancel{n}}{\begin{bmatrix} \\ \\ \end{bmatrix}} \underset{\cancel{n} \times \cancel{d} \times 1}{\begin{bmatrix} \\ \\ \end{bmatrix}} \underset{\cancel{d} \times 1 \times n}{\begin{bmatrix} \\ \\ \end{bmatrix}} = \underset{d+1 \times n}{\begin{bmatrix} \\ \\ \end{bmatrix}}$$

Problem 6.4 from FA21 Midterm

Suppose we introduce two more features:

- width alone, and
- area, which is defined as length times width

Suppose we also decide to remove the intercept term of our prediction rule. With all of these changes, our prediction rule is now

carat	length	width	price
0.40	4.81	4.76	1323
1.04	6.58	6.53	5102
0.40	4.74	4.76	696
0.40	4.67	4.65	798
0.50	4.90	4.95	987

no w_0?

$$\text{predicted price} = w_1 \cdot \text{carat} + w_2 \cdot \text{length} + w_3 \cdot \text{width} + w_4 \cdot (\text{length} \cdot \text{width})$$

0+

each row of X:

[carat length width l·w]

$$X = \begin{bmatrix} 0.4 & 4.81 & 4.76 & 4.81(4.76) \\ 1.04 & 6.58 & 6.53 & 6.58(6.53) \end{bmatrix}$$

- Write out just the first 2 rows of the design matrix X for this new prediction rule. You do **not** need to simplify the numbers in your matrix, it is fine if they involve the multiplication symbol.
- Is the optimal coefficient for carat, w_1^*, for this new prediction rule guaranteed to be equal to 10000, the optimal coefficient for carat in our original prediction rule? No justification is necessary.

$$w^* = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix}$$

$w_1 = 1000$ 6.2

$$\text{old } w = \begin{bmatrix} 2000 \\ 10\,000 \\ -1000 \end{bmatrix} \begin{matrix} \leftarrow w_0 \\ \leftarrow w_1 \\ \leftarrow w_2 \end{matrix}$$

$$\text{old} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} \quad \text{new} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ \cancel{w_5} \end{bmatrix} \rightarrow 0$$

Zoe

Problem 4a from Homework 2

Problem 4. Slippery Slope

In Lecture 2, we found that $h^* = \text{Median}(y_1, y_2, ..., y_n)$ is the constant prediction that minimizes mean absolute error:

$$R_{\text{abs}}(h) = \frac{1}{n}\sum_{i=1}^{n} |y_i - h|$$

Suppose that we have a dataset of numbers $y_1, y_2, ..., y_n$ such that n **is odd** and the values are arranged in increasing order. That is, $y_1 \leq y_2, \leq \ldots \leq y_n$.

Note: Parts (a) and (b) are independent of each other.

a) Suppose that $R_{\text{abs}}(\alpha) = V$, where V is the minimum value of $R_{\text{abs}}(h)$ and α is one of the numbers in our dataset.

Let $\alpha + \beta$ be the smallest value greater than α in our dataset, where $\beta > 0$. Another way of thinking about this is that $\beta = (\text{smallest value greater than } \alpha) - \alpha$.

Suppose we modify our dataset by replacing the value α with the value $\alpha + \beta + 1$. In our new dataset of n values, what is the new minimum value of $R_{\text{abs}}(h)$ and at what value of h is it minimized? Your answers to both parts should only involve the variables V, α, β, n, and/or one or more constants.

Harshi

Absolute error

Raho(α) = V $\rightarrow$ min value

$\alpha \longrightarrow$ median   old

$\alpha + \beta$ $\beta > 0$

new

$\alpha \longrightarrow \alpha + \beta + 1$

median ?

Raho(h) new value ?

old

$y_1 \ldots \alpha \quad \underline{\alpha+\beta} \ldots y_n$

$\uparrow$

median

new

$y_1 \ldots \underline{\alpha+\beta} \quad \alpha+\beta+1 \ldots y_n$

new median is $\underline{\alpha+\beta}$

new minimizer

old

$y_1 \ldots \alpha \quad \alpha+\beta \ldots y_n$

median

imagine new

$+\beta$

$y_1 \ldots \alpha+\beta \quad \alpha+\beta+1 \ldots y_n$

median

old left abs dev sum

olads total of n values.

$\frac{n-1}{2}$ values to the right

$y_1 + \beta$
$y_2 + \beta$
… y just before new median
$\left(\frac{n-1}{2}\right)$ of these

olads → old.

$\underbrace{\text{olads} + \beta\frac{(n-1)}{2}}$ → new − olads.

$\frac{\beta(n-1)}{2}$

change: } left values

old $\alpha+\beta-\alpha \to \beta$

new $(\alpha+\beta+1)-(\alpha+\beta) \to 1$

$1-\beta$

0 → middle or median position

values to the right → moved β closer to each $\xrightarrow{\text{these many.}}$ $\underbrace{\left(\frac{n-1}{2}\right)-1}$

rlads
d

$\text{rlads} - \beta\left(\frac{n-1}{2}-1\right)$

$-\beta\left(\frac{n-1}{2}\right)-1$

$\text{Rabs}(\underset{\alpha}{\text{old}}) = V$

Vn → new $V + \frac{1}{n}$

new sum of abs dev ol

$$\frac{1}{n}\left(Vn + \underbrace{\frac{\beta(n-1)}{2}}_{\text{left vals}} + \underset{\text{med}}{0} + \underset{\text{val imm r of med}}{(1-\beta)} + \left(\underset{\text{val to right except imm}}{-\beta\left(\frac{n-1}{2}-1\right)}\right)\right)$$

Problem 4b from Homework 2

Problem 4. Slippery Slope

In Lecture 2, we found that $h^* = \text{Median}(y_1, y_2, ..., y_n)$ is the constant prediction that minimizes mean absolute error:

$$R_{\text{abs}}(h) = \frac{1}{n}\sum_{i=1}^{n} |y_i - h|$$

Suppose that we have a dataset of numbers $y_1, y_2, ..., y_n$ such that n **is odd** and the values are arranged in increasing order. That is, $y_1 \le y_2, \le \ldots \le y_n$.

Note: Parts (a) and (b) are independent of each other.

b) 🥑🥑🥑 Let y_a and y_b be two values in our dataset such that $y_a < y_b$ and that the slope of $R_{\text{abs}}(h)$ is the same between $h = y_a$ and $h = y_b$. Specifically, let d be the slope of $R_{\text{abs}}(h)$ between y_a and y_b.

Suppose we introduce a new value q to our dataset such that $q > y_b$. In our new dataset of $n+1$ values, the slope of $R_{\text{abs}}(h)$ is still the same between $h = y_a$ and $h = y_b$, but it's no longer equal to d. What is the slope of $R_{\text{abs}}(h)$ between $h = y_a$ and $h = y_b$ in our new dataset? Your answer should depend on d, n, q, and/or one or more constants.

Harshi

$R_{abs} \rightarrow$ absolute loss.

new

new datapoint q
such that $q > y_b$.

$y_a < y_b$
slope b/w $= d$

★ Lecture 2

$\frac{d}{dh} R_{abs}(\underline{h}) = \frac{1}{n}$(# pts to left of h − # pts to right of h)

@ pt c $\uparrow$ c b/w y_a and y_b slope is d

(# pts left of c − # pts to the right of c)

slope old $\rightarrow d = \frac{1}{n}$ (# pts left of c − # pts of right of c)

$\rightarrow nd =$ (# pts left of c − # pts of right of c + 1))

new slope $\rightarrow d' = \frac{1}{n+1}$ (# pts left of c − (# pts right of c + 1))

$$d' = \frac{1}{n+1}\left(\underbrace{\#\text{pts left of } c - \#\text{ pts right of } c}_{} - 1 \right)$$

$$d' = \frac{1}{n+1}\left(nd - 1 \right) \Rightarrow \boxed{\frac{nd-1}{n+1}}$$

Problem 4.1 from WI24 Midterm

Albert collected 400 data points from a radiation detector. Each data point contains 3 features: feature A, feature B and feature C. The true particle energy E is also reported. Albert wants to design a linear regression algorithm to predict the energy E of each particle, given a combination of one or more of feature A, B, and C. As the first step, Albert calculated the correlation coefficients among A, B, C and E. He wrote it down in the following table, where each cell of the table represents the correlaton of two terms:

	A	B	C	E
A	1	-0.99	0.13	0.8
B	-0.99	1	0.25	-0.95
C	0.13	0.25	1	0.72
E	0.8	-0.95	0.72	1

Albert wants to start with a simple model: fitting only a single feature to obtain the true energy (i.e. $y = w_0 + w_1 x$). Which feature should he choose as x to get the lowest mean square error?

- ○ A
- ● B
- ○ C

Yosen

Problem 4.2 from WI24 Midterm

Albert collected 400 data points from a radiation detector. Each data point contains 3 features: feature A, feature B and feature C. The true particle energy E is also reported. Albert wants to design a linear regression algorithm to predict the energy E of each particle, given a combination of one or more of feature A, B, and C. As the first step, Albert calculated the correlation coefficients among A, B, C and E. He wrote it down in the following table, where each cell of the table represents the correlaton of two terms:

	A	B	C	E
A	1	-0.99	0.13	0.8
B	-0.99	1	0.25	-0.95
C	0.13	0.25	1	0.72
E	0.8	-0.95	0.72	1

Albert wants to add another feature to his linear regression in part (a) to further boost the model's performance. (i.e. $y = w_0 + w_1 x + + w_2 x_2$) Which feature should he choose as x_2 to make additional improvements?

- ○ A
- ○ B
- ● C

Yosen

Problem 4.3 from WI24 Midterm

Albert collected 400 data points from a radiation detector. Each data point contains 3 features: feature A, feature B and feature C. The true particle energy E is also reported. Albert wants to design a linear regression algorithm to predict the energy E of each particle, given a combination of one or more of feature A, B, and C. As the first step, Albert calculated the correlation coefficients among A, B, C and E. He wrote it down in the following table, where each cell of the table represents the correlaton of two terms:

	A	B	C	E
A	1	-0.99	0.13	0.8
B	-0.99	1	0.25	-0.95
C	0.13	0.25	1	0.72
E	0.8	-0.95	0.72	1

Albert further refines his algorithm by fitting a prediction rule of the form:

$$H(A,B,C) = w_0 + w_1 \cdot \overset{b_1}{\boxed{A \cdot C}} + w_2 \cdot \overset{b_2}{\boxed{B^{C-7}}}$$

Given this prediction rule, what are the dimensions of the design matrix X?

$n \times (d+1)$ ✓

$$\begin{bmatrix} & \\ & \end{bmatrix}_{r \times c}$$

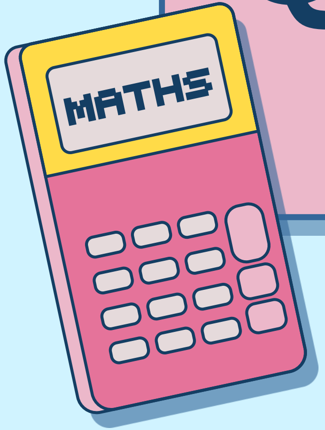
So, what are r and c in r rows $\times$ c columns?

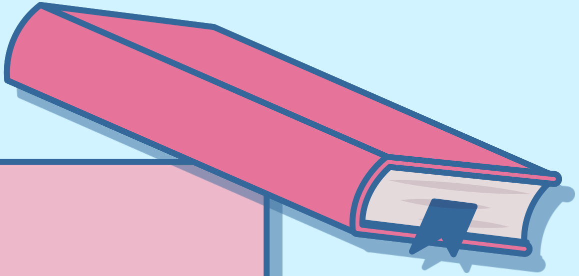
r: 400, c: $d+1 = 3$

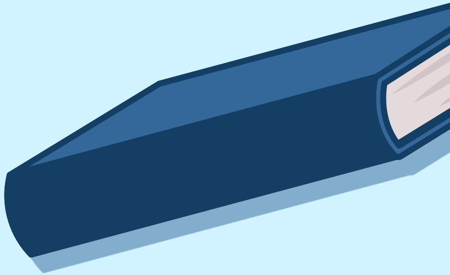
$$\begin{bmatrix} 1 & b_1^{(1)} & b_1^{(2)} \\ \vdots & b_2^{(1)} & b_2^{(2)} \\ \vdots & \vdots & \vdots \\ 1 & b_{400}^{(1)} & b_{400}^{(2)} \end{bmatrix}$$

Yosen

Questions?







DSC 40A

Thanks for Coming!

Good luck everyone!