מבוא למערכות לומדות

רועי זהבי 208904757

<u>תרגיל 5 - חלק תיאורטי:</u>

1) <u>Regularization</u>: יהיו $X\in\mathbb{R}^{n\times d}$ ו-$y\in\mathbb{R}^{n}$ ונניח כי X^TX הפיכה.

נסמן ב- $\hat{w}$ את פתרון הרגרסיה הלינארית הפשוטה ו- $\hat{w}_\lambda$ את פתרון Ridge (כאשר $\lambda=0$ $\hat{w}=\hat{w}_\lambda$)

(a) נראה כי $\hat{w}_\lambda = A_\lambda \hat{w}$ כאשר $A_\lambda=(X^TX+\lambda I)^{-1}(X^TX)$

<u>פתרון</u>: בתרגול ראינו כי

$$\hat{w}_\lambda = (X^TX+\lambda I)^{-1}\cdot X^Ty$$

כמו כן נזכר כי

$$\hat{w} = X^+y = V\Sigma^+U^Ty \quad , \quad V,U$$

מט' אורתו'

<u>נזכיר</u>:

$$X = U\Sigma V^T$$

$$A_\lambda \hat{w} = (X^+X+\lambda I)^{-1}(X^TX)\cdot \hat{w} =$$

$$= (X^TX+\lambda I)\cdot(V\Sigma^TU^TU\Sigma V^TV\Sigma^+U^Ty)$$

נרצה להגיע למצב בו יש X^Ty ויסתדר!

ונזכר כי

$$\Sigma^+_{i,i} = \begin{cases} 1/\sigma_i & \sigma_i \neq 0 \\ 0 & else \end{cases}$$

$$\Sigma_{i,i} = \sigma_i$$

⇐

$$\Sigma\Sigma^+_{i,i} = \begin{cases} 1 & \sigma_i \neq 0 \\ 0 & else \end{cases}$$

נתון כי X^TX הפיכה, לכן אנו במקרה ההפיך כאשר ראינו כבר כי אין ערכים סינגולריים שווים ל-0!

⇐ $\Sigma^+_{i,i} = 1/\sigma_i$ לכל i ו-σ_j

$$\Sigma\Sigma^+ = I$$

$$\Rightarrow (V\Sigma^TU^T\cancel{U\Sigma V^TV\Sigma^+}U^Ty) = X^Ty$$

⇐ $$A_\lambda \hat{w} = (X^TX+\lambda I)^{-1}X^Ty = \hat{w}_\lambda$$

כנדרש!

6) נראה כי רידג' הינו אומד מוטה, קרי

$\mathbb{E}(\hat{w}_\lambda) \neq w$ לכל סכת.

פתרון:

$$\mathbb{E}[\hat{w}_\lambda] = \mathbb{E}(A_\lambda \hat{w}) = \mathbb{E}((X^TX+\lambda I)^{-1}X^TX\hat{w})$$

$$(X^TX+\lambda I)^{-1}X^TX\cdot\mathbb{E}(\cancel{\hat{w}}) \neq w$$

w כאשר ראינו כי זה אומד לא מוטה

מכיוון ש- $\lambda > 0$ אזי $(X^TX+\lambda I)^{-1}$ שונה מההופכי של X^TX, לכן זו כופלת את w בקבוע שונה לא יחידה ולכן $w \neq \hat{w}_\lambda$.

ב) נראה כי הוא מוטה:

נזכיר את ה
בורונית!

$$Var(\hat{w}_\lambda) = Var(A_\lambda \hat{w}) =$$

$$= A_\lambda Var(\hat{w}) \cdot A_\lambda^T = A_\lambda \cdot \sigma^2 (X^TX)^{-1} \cdot A_\lambda^T$$

סקלר!

$$= \sigma^2 A_\lambda (X^TX)^{-1} A_\lambda^T$$

ג) נפתח כיוון מפורט פירוק Bias-Variance כפונקציה של λ.

תזכורת: בתרגול 4 ראינו כי

$$MSE = \mathbb{E}[\|\hat{y} - y\|^2] = Bias(\hat{y})^2 + Var(\hat{y})$$

$$\Downarrow$$

$$MSE = Bias(\hat{w}_\lambda)^2 + Var(\hat{w}_\lambda) =$$

$$= \mathbb{E}[\hat{w}_\lambda]^2 + \sigma^2 A_\lambda (X^TX)^{-1} A_\lambda^T =$$

$$\mathbb{E}[A_\lambda \hat{w}] = A_\lambda \mathbb{E}(\hat{w}) = A_\lambda w$$

$$\Rightarrow MSE = (A_\lambda w)^2 + \sigma^2 A_\lambda (X^TX)^{-1} A_\lambda^T$$

⇐ ניתן לראות כי המשתנה היחיד הפנימי הוא λ ולכן
הבלעדיו לפני כנדרש!

ב)

$$\frac{d}{d\lambda} MSE(\hat{w}_\lambda)\Big|_{\lambda=0}:$$

ראשית נשים לב:

$$Bias^2(\hat{w}_\lambda) = \| \mathbb{E}[\hat{w}_\lambda] - \mathbb{E}[\hat{w}] \|^2 =$$

מן תכונות מ"פ!

נו

$$= \| (A_\lambda - I) w \|^2 = w^T (A_\lambda - I)^T (A_\lambda - I) w$$

כמו כן, ראינו עוד כי בנוסף עבור השונות ישר.ii

$$Var(\hat{w}_\lambda) = \mathbb{E}[\| \hat{w}_\lambda - \hat{w} \|^2] =$$

סכום II
תוחלות הפרי,
הריבועים &
אפכסון גם השונות

$$= trace(Var(w_\lambda)) = \sigma^2 tr(A_\lambda (X^TX)^{-1} A_\lambda^T)$$

$$\Downarrow$$

$$\frac{d}{d\lambda} bias(\hat{w}_\lambda)^2 = \frac{d}{d\lambda} \sum_i \sum_j (A_\lambda - I)_{ij} w_j^2 \Big|_{\lambda=0} =$$

$$= 2 \sum_i \Big(\sum_j (A_\lambda - I)_{ij} w_j \frac{d}{d\lambda} \sum_j (A_\lambda - I)_{ij} w_j \Big) = 0$$

⇐ פד קיק רק λ תלוי. ולכן הם מתאפס!

$$\frac{d}{d\lambda} Var(\hat{\omega}_\lambda) = \sigma^2 \frac{d}{d\lambda} tr(A_\lambda X^T X A_\lambda^T)$$

מטרי
צו נים
הפיתוח ב-ע לפי נגזרת
לצבור טבעית

$$\frac{d}{d\lambda} Var(\hat{\omega}_\lambda^T) = \sum_{i,j} \frac{d\, Var(\hat{\omega}_\lambda^T)}{d(A_\lambda)_{i,j}} \cdot \frac{d(A_\lambda)_{i,j}}{d\lambda}\Bigg|_{\lambda=0} =$$

$$= Tr\left(\frac{d\, Var(\hat{\omega}_\lambda^T)}{dA_\lambda} \cdot \frac{dA_\lambda}{d\lambda} \right)\Bigg|_{\lambda=0} =$$

$$= -2\underset{\overset{\vee}{0}}{\sigma^2} Tr\Big(\underset{\overset{\vee}{0}}{(X^TX)^{-1}}(X^TX)^{-T}\Big) < 0$$

⇒ קיבלנו כי נגזרת ה-MSE הינה ב-0, נגטיבי!

הק

F. נסיק מן ה-E כי עבור $\hat{\omega}_\lambda$ הנגזרת של MSE שלילית!
קרי מוסף ה-bias ומקדמה את MSE יורדת, מה שאומר כי ההפסד קטן.
⇒ מכך קיבלנו כי גם בנקודה של $\lambda=0$, קרי זהו מודל LS סטנדרטי, אנו רואים כי הנגזרת שלילית ⇒ ההפסד יורד.

מכיוון שהפונק׳ רציפה (וגם גזירה) כאשר λ גדל
אך עדיין הדין מאוד נמוך כי ערך ה-MSE יהיה
=> הוספנו ערך רגולריזציה לRidge מעלה
את ה-MSE בהרבה!

PCA:

(2) פתרון: נראה כי הערך המקסימלי של השונות מתקבל ע"י
הו"ע של הע"ע הגדול ביותר $\Rightarrow$ השונות של v^Tx
תלויה על כיוון.

ראשית: יהי $v \in \mathbb{R}^d$, $\|v\|_2 = 1$

ממוצע ההטלות של x הם $\mathbb{E}(v^TX) = \frac{1}{m}\sum_{i=1}^{m} v^T x_i$

$= v^T \bar{x}$ הם כפי שראינו בכיתה

$$\mathrm{var}(v^TX) = \mathbb{E}_x\left(\left(v^Tx_i - \mathbb{E}_x[v^Tx]\right)^2\right) =$$

$$= \frac{1}{m}\sum_i \left(v^Tx_i - v^T\bar{x}\right)^2 =$$

$$= \frac{1}{m}\sum_i \left[v^T(x_i - \bar{x})\right]^2 =$$

$$= \frac{1}{m}\sum_i \left(v^T(x_i - \bar{x})\right)\left(v^T(x_i-\bar{x})\right)^T =$$

$$= \frac{1}{m}\sum_i v^T(x_i - \bar{x})(x_i - \bar{x})^T v =$$

$$= v^T \underbrace{\sum_i (x_i - \bar{x})(x_i - \bar{x})^T}_{\text{cov sample mat}} v =$$

$$= v^T S v$$

$\Rightarrow$ נרצה לפתור: $\hat{v} = \arg\max_{\|v\|=1} v^T S v$

דרך לינקס שונה!

נזכר ב- lagrange multipliers

נגדיר $g(v) = 1 - v^T v$

ש

$$L = v^T v + \lambda g(v)$$

$$\frac{d}{dv} L = 2Sv - 2\lambda v = 0$$

<= זה אומר כי המקסימום ב- $v^T v$ הינו ו"ע

ניתן להסיק כי המקסימום הוא ו"ע λ ע"ע גדול
ביותר!

<= קיבלנו כי הבעיה שלנו, ואם המספר ע"י ו"ע
של הע"ע הגדול ביותר מקסימום של וקטור
$X^T v$ => ואכן זו של יכולה להיות בעיה
מהותית, כלומר!

∎

בונוסים:

(3) יהי k הקרנל המקורי הנתון שלנו, כפי שראינו כי עם פונק' f

$$\tilde{k}(x,x') = f(x)\cdot k(x,x')\cdot f(x')$$

הינו קרנל PSD תקין!

ונבחר

$$f(x) = \begin{cases} 1/\sqrt{k(x,x)} & k(x,x) \geq 0 \\ -1/\sqrt{-k(x,x)} & \text{else} \end{cases}$$

⇐ נבדוק:

$k(x,x) \geq 0$

$$\tilde{k}(x,x) = f(x)\,k(x,x)\cdot f(x) = \frac{1}{k(x,x)}\,k(x,x) = 1$$

אחרת $k(x,x) < 0$

$$\tilde{k}(x,x) = \frac{-1}{\sqrt{-k(x,x)}}\,k(x,x)\cdot\frac{-1}{\sqrt{-k(x,x)}} = 1$$

קרי באופן זה קיבלנו את הנדרש!

מש"ל

ב.) נכונה ל כ"ל: $x \in \mathbb{R}^2$ כאשר $x_i = (x_i, y_i)$

נקודה ב-$\mathbb{R}^2$ שלו יוקצה כי ה-label y_i יקבע
באופן הבא:

אם $\sqrt{x_i^2+y_i^2}$ זוגי אז נתייג 1 אחרת
-1.

⇐ וזה ברור כי לא ניתן להפריד את S
הדאטה ב-$\mathbb{R}^2$ באופן לינארי.

נגדיר $\Psi(x) = (x_i, y_i, \sqrt{x_i^2+y_i^2} \,\%\, 2)$

⇐ נוסף מימד שמייצג את זוגיות הנורמה
באופן זה נוכל לקבל הפרדה בקלות כאשר
ושלה מימד ג' ששווה 1 או 0 לשם הפרדה.
אחלה!

ב) (a) לא נכון! נראה דוגמה נגדית:

נראה כי היא גרעין אינה PSD.

תהי:

דגימות כמו $S = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$

כאשר: $k = e^{(\|x-y\|^2)}$

$$\Downarrow$$

$$k = \begin{pmatrix} 1 & e^2 \\ e^2 & 1 \end{pmatrix}$$

נראה ש- k לא PSD, יהי $v = \begin{pmatrix} -2 \\ 1 \end{pmatrix}$

$$v^T k v = (-2+e^2 \quad -2e^2+1) \begin{pmatrix} -2 \\ 1 \end{pmatrix} =$$

$$-2(-2+e^2) - 2e^2 + 1 = 4 - 4e^2 + 1 = 5 - 4e^2$$

$\Leftarrow e^2 > 2 \Leftarrow 5 - 4e^2 < 0 \Leftarrow$ לא PSD, כנדרש!

$\boxtimes$

ב) לא נכון! נראה כי $k(x,x)<0$

נבחר שני קרנלים פולינומיאלים כפי שראינו בתרגול

$$k_1(x,z) = (\langle x,z\rangle + -5)^3$$

$$k_2(x,z) = (\langle x,z\rangle + 0)^2$$

$$\Downarrow$$

$$k(0,0) = k_1(0,0) - k_2(0,0) = -5^3 = -125 < 0$$

קרי קיבלנו כי הנ"ל לא PSD ולכן הנ"ל לא קרנל חוקי!

$\boxtimes$

ג) הטענה נכונה!

הוכחה: יהיו k_a, k_b קרנלים ולידים כך ש-

$k = k_a + k_b$ ואת אותן פעולות על

$$x = \begin{bmatrix} x_a \\ \hline x_b \end{bmatrix}$$ ו- $x \in \mathbb{R}^d$

נשים לב כי $k_{i,j} = k_{a_{i,j}} + k_{b_{i,j}}$

$\Rightarrow$ ולכן גם סימטרית!

נראה כי היא PSD.

$$k(x,x) = x k x^T = x(k_a + k_b)x^T$$

נראה כי גם קרנל זה מוגדר חצי כאשר כמות הדגימות של תצפית בגודל.

$$X = X_a^* + X_b^*$$

מה [illegible] השלמות [illegible] ל-a ובכניסות b הלא מוגדרות יש 0.

$$X K X^T = (X_a^* + X_b^*) K (X_a^* + X_b^*)^T =$$

$$= X_a^* K X_a^T + X_b^* K X_b^{*T} =$$

$$= X_a^* (K_a + K_b) X_a^{*T} + X_b^* (K_a + K_b) X_b^{*T}$$

<= מהגדרת הקרנלים החדשים מוסיפים רק את הקורד. נקבל הרכיבים אחרים ויתן ערך 0

$$\Rightarrow X_a^* K_a K_a^* + 0 + 0 + X_b^* K_b K_b^{*T} =$$

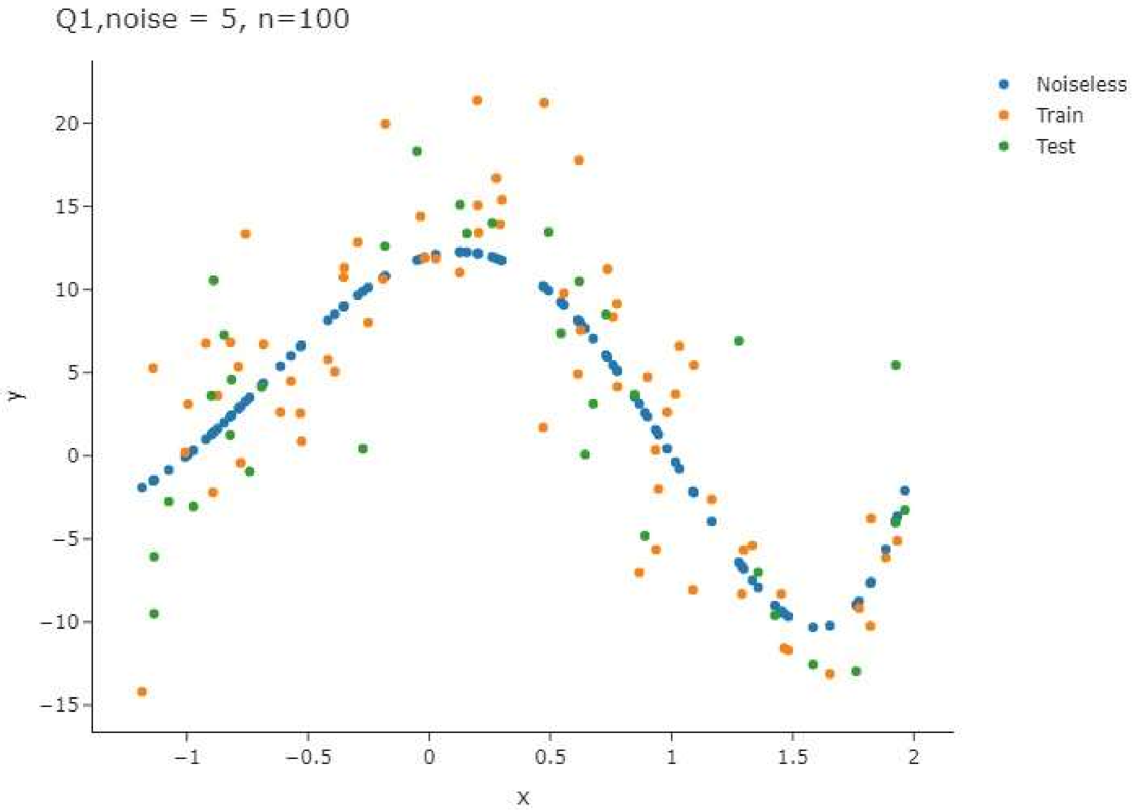
$$= \underbrace{X_a K_a X_a^T}_{\geq 0} + \underbrace{X_b K_b X_b^{*T}}_{\geq 0} \geq 0$$

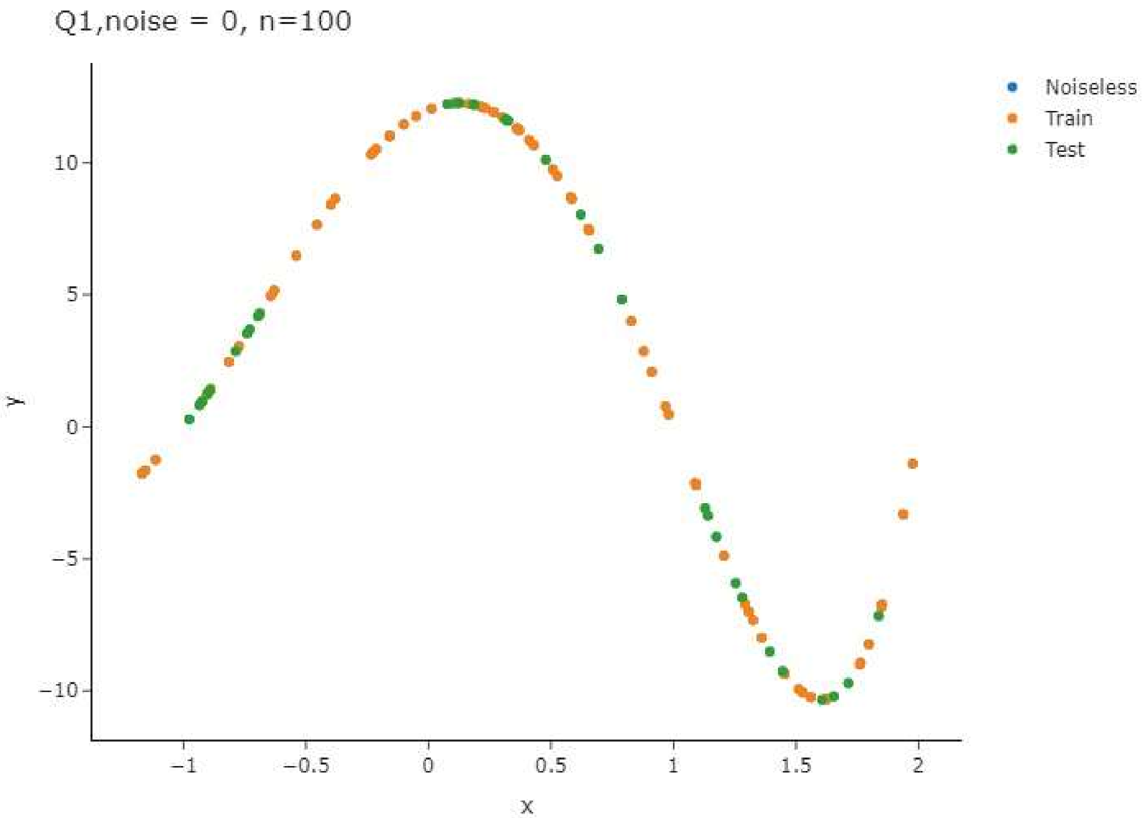
כי הקרנלים של משלים את הוכחת הזוג.

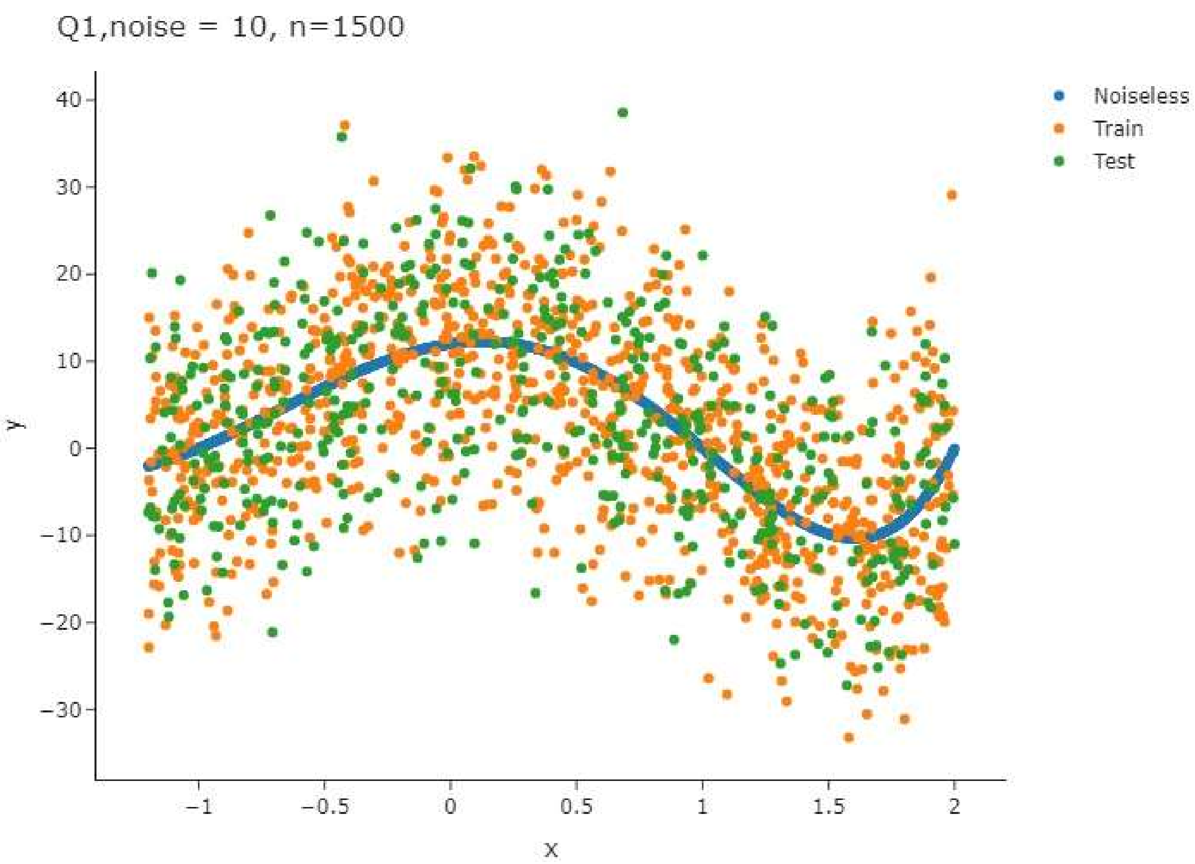
<= הינו PSD וזהו קרנל תקין!

מבוא למערכות לומדות 67577
רועי זהבי 208648154

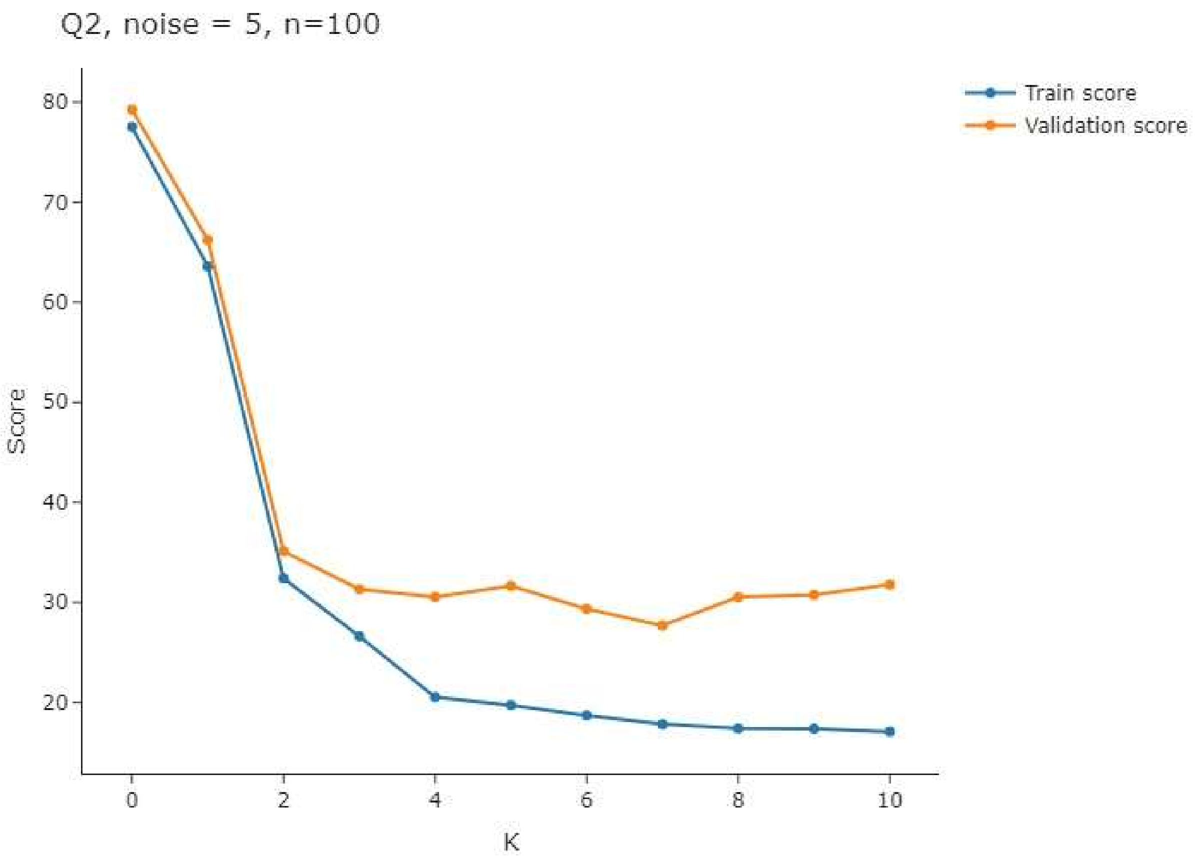
1. להלן הפלטים עבור סט הנתונים בהינתן קלטים שונים

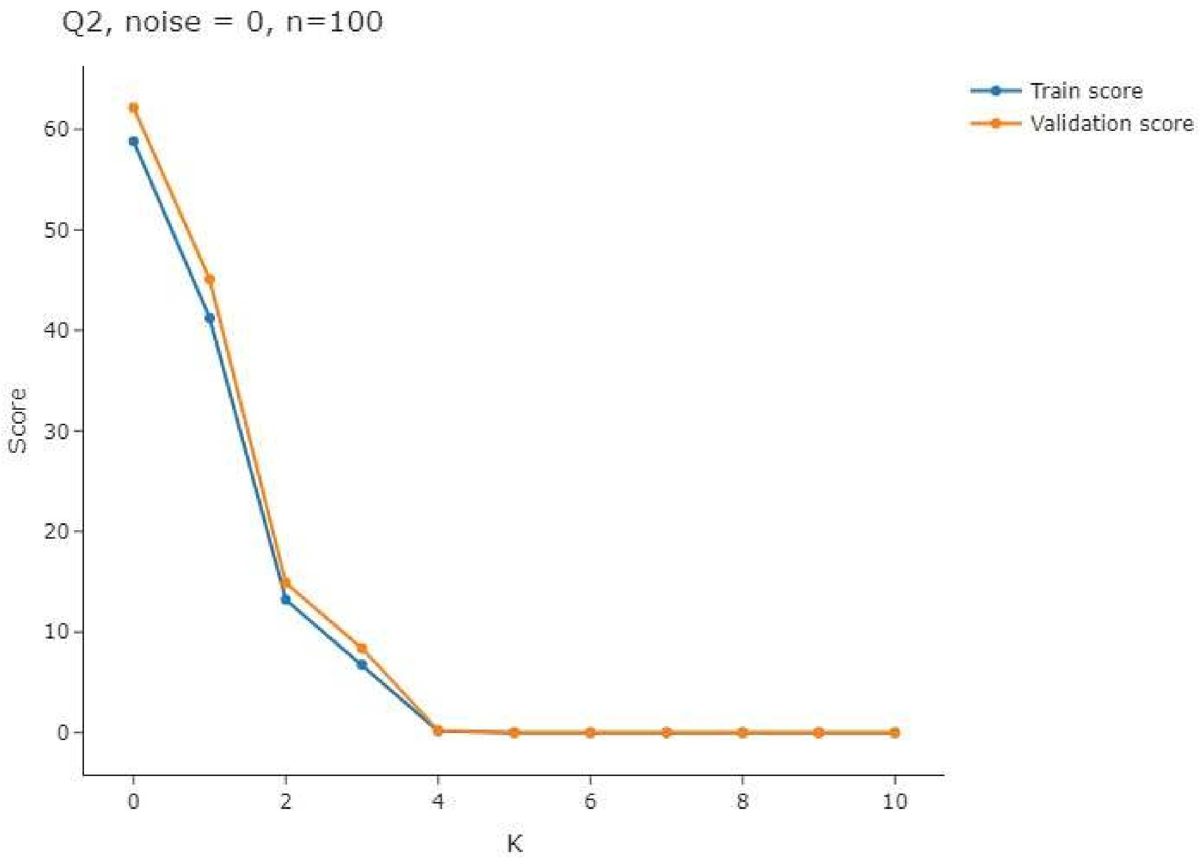


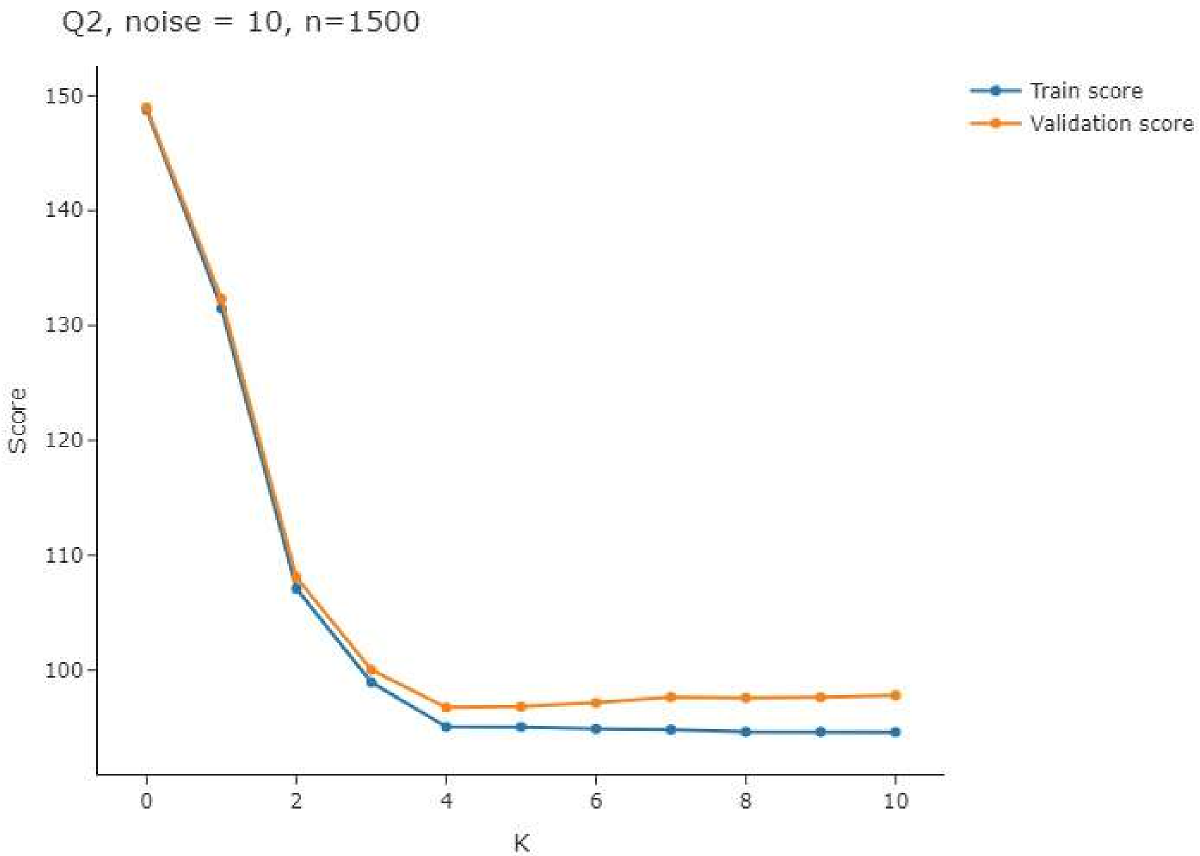




2. להלן הפלטים בהינתן התרחישים השונים (תושבה משולבת עבור סעיפים 4,5):




Q2, noise = 0, n=100
Train score
Validation score
Score
K
0
10
20
30
40
50
60
0
2
4
6
8
10

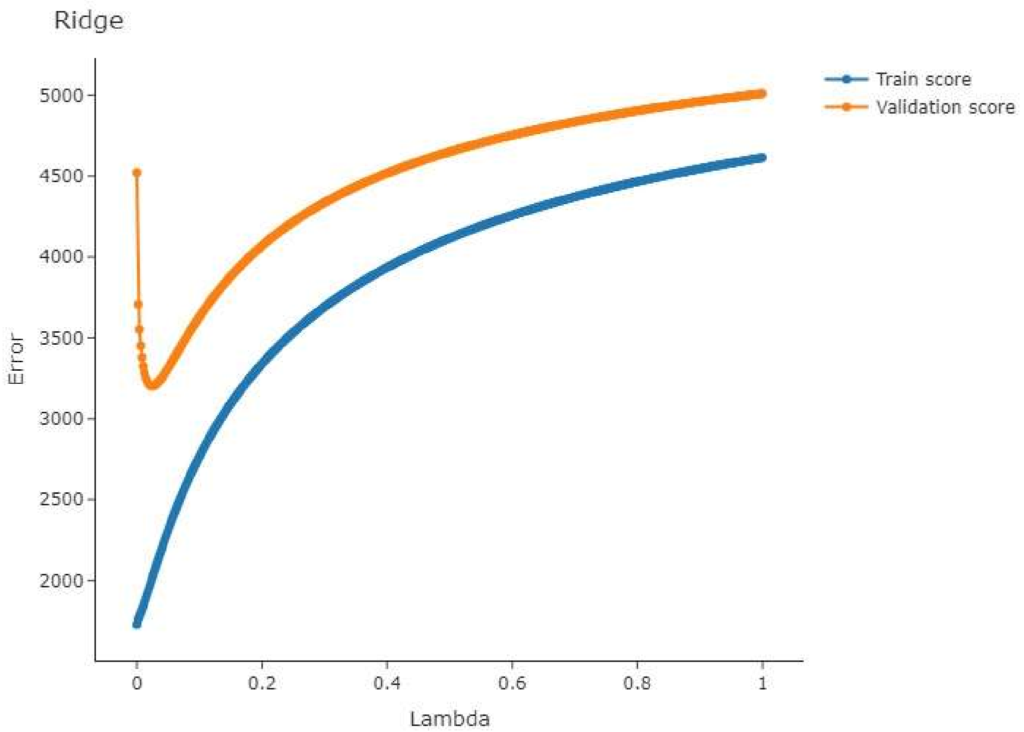

Q2, noise = 10, n=1500
Train score
Validation score
Score
K
100
110
120
130
140
150
0
2
4
6
8
10

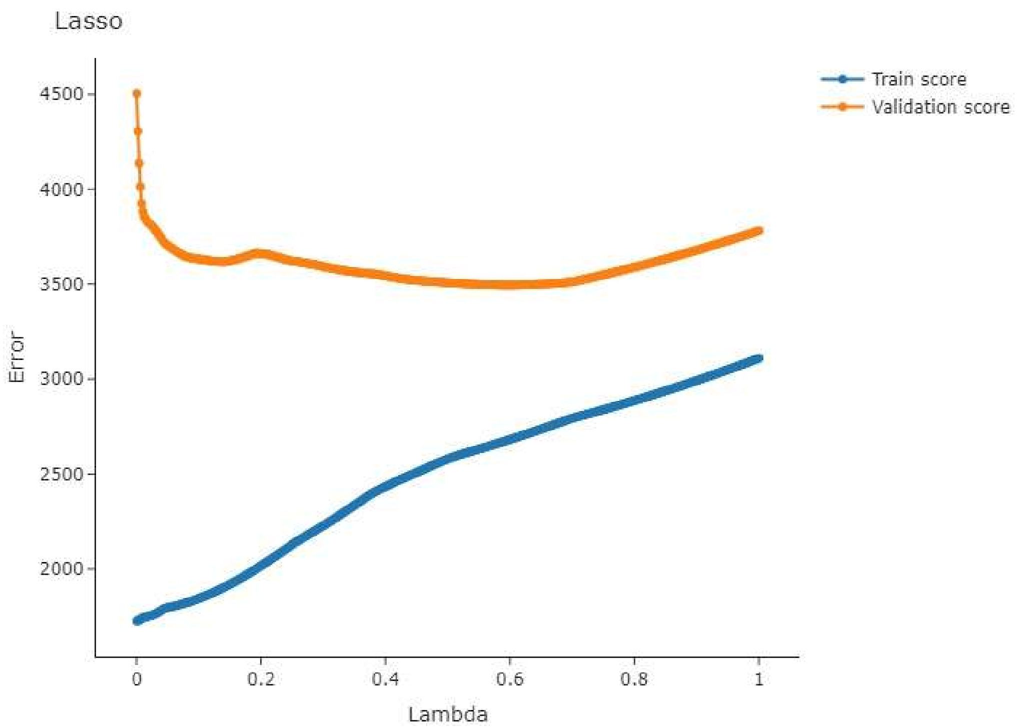
נתבונן בתוצאות שקיבלנו, נשים לב למספר אבחנות:

- ראשית כאשר כמעט ואין רעש שגיאת האימון מתלכדת עם שגיאת הולידאטציה – אבחנה שנובעת ישירות מן העבודה כי כל הנקודות מייצגות את המציאות, כאשר אין משמעות אמיתית לאימון המודל על חלקים שונים (רועשים יותר או פחות) – כאשר הרעש עולה ההפך מתקבל, שוב אבחנה לה אנו מצפים.
- כמו כן, מנגנון הבקרה שלנו עובד כאשר ניתן לראות כי שגיאת הולידאציה גדולה לרוב משגיאת הסט האימון – כמו כן דבר שנרצה לראות.
- נשים לב כי תחת שינוי דרגות הפולינום השגיאות גם מתנהגות בצורה קוהרנטית, כאשר שדרגת הפולינום מתקרבת לדרגה האמיתית השגיאות הן מזעריות ביחס לקצוות האחרים – נשים לב כי שתי המדידות מקיימות גם את עיקרון $bias$ – $variance$ שראינו מספר פעמים עד כה בקורס ($under\ fit$ מול $overfit$).
- לבסוף נראה כי גם כאשר אנו מעלים את מספר המדידות שלנו למרות שהרעש גדול, המודל מצליח להתמודד ולספק תחזית טובה. זאת אנו מסוגלים להפיק תחת שימוש ב-$cross\ validation$, בכל שלב מבצעים אימון על חלק אחר בסט האימון שלנו ובודקים שוב ושוב כך שהמודל נחשף לכמה שיותר מן המידע. באופן זה אנו מקטינים את ה-$overfit$ שיכול להתקיים בהתאמת פולינומים תחת רעש גבוהה ומחזקים את השימוש בחוק המספרים הגדולים – שגם תחת הרעש מספר רב של דגימות ישאף לתצורה האמיתית (נשאף לתוחלת, אשר 0) ולכן נוכל עדין לקבל פרדיקציה טובה.

3. להלן דרגות הפולינום אשר המודל חזה עם הטעויות שקיבלנו עבור שימוש בדרגה זו:
   - במקרה עבור 100 דגימות ורעש 5 - $k = 7, MSE = 22.62$
   - במקרה עבור 100 דגימות ורעש 0 - $k = 5, MSE = 1.35$
   - במקרה עבור 1500 דגימות ורעש 10 - $k = 4, MSE = 98.89$

7. להלן הפלטים עבור $\lambda \in [0,1]$ אשר נלקחו 500 נקודות במרחק שווה. אציין כי ע"מ להגיע לטווח זה, ראשית התחלתי עם טווח גדול $[0,10]$ ניתן היה לראות כי השגיאה מתקבעת ב-$4\sim \geq \lambda$ ולכן צמצמתי את הטווח עד לקבלת מדידה שאפשרה לקבל ערך מינימלי.





שוב ניתן לראות את הבדל הניכר בין השגיאות על סט האימון וסט הואלידציה, כאשר הפער היה מעט גדול יותר עבור השימוש ב-$Ridge$. כמו כן, ניתן לראות את מגמת השגיאה כמגמה שראינו

בכיתה – לפיה השימוש בנורה $l2$ מייצרת פונקציה גזירה ב-0 ולכן עקומת רידג' מעט מזכירה יותר פרבולה מאשר עקומת $Lasso$ אשר דומה יותר לפונקצית הערך המוחלט.

כמו כן, ניתן לראות כי הוספת ערך הרגולריזציה השפיע יותר בהתחלה על $Lasso$ מאשר על $Ridge$ ביכולת להוריד את השגיאה תחת סט הולידאציה.

8. להלן תוצאות השגיאה שהתקבלו עבור שלושת המודלים:

$$Ridge\ achieved\ 3249.69\ with\ lambda\ of\ 0.024$$

$$Lasso\ achieved\ 3641.16\ with\ lambda\ of\ 0.597$$

$$Basic\ linear\ regression\ achieved\ 3612.25$$

ניתן לראות כי במקרה זה $Ridge$ הוא זה שהשיג את השגיאה הקטנה ביותר, תחת שימוש בפרמטר רגולריזציה מאוד קטן, כ-0.02 אך על סט האימון הקטן הנ"ל גרר שיפור משמעותי (מעל 10%) בשגיאה.