Theoretical part:

1.1: Regularizatio-

a) $X \in \mathbb{R}^{m \times d}$, $y \in \mathbb{R}^m$ $\quad X^T X$ - הפיך

$\hat{w}$ = LS solution, $\hat{w}_\lambda$ = ridge solution $\lambda \geq 0$

$$y = Xw + \varepsilon \quad , \quad \varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2) \quad , \quad Bias(\hat{w}) = 0$$

$$E(\hat{w}) = E\left(w + (X^T X)^{-1} X^T \varepsilon\right) = w + (X^T X)^{-1} X^T E(\varepsilon) = w$$

$$\hat{w}_\lambda = (X^T X + \lambda I_d)^{-1} (X^T X) \cdot \hat{w} \qquad : \text{ל.3}$$

בתרגול מצאנו נוסחה סגורה לridge, נשתמש בה באגף השמאלי:

$$\hat{w}_\lambda = \underbrace{(X^T X + \lambda I_d)^{-1} X^T y} = \arg\min \|Xw - y\|^2 + \lambda \|w\|^2$$

בנוסף לLS יש נוסחה סגורה מהגדרתו ולכן האגף הימני של המשוואה יוצא:

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

$$\Downarrow$$

$$(X^T X + \lambda I_d)^{-1} \overbrace{(X^T X) \cdot (X^T X)^{-1}}^{I_d} X^T y = \underbrace{(X^T X + \lambda I_d)^{-1} X^T y}$$

וקיבלנו שיוויון בין שני האגפים כנדרש $\boxtimes$

בסעיף a הראנו ש:

b) $$\hat{w}_x = (X^TX+\lambda I_d)^{-1}(X^TX)\cdot\hat{w}$$

$$E(\hat{w}_x) = E\left[(X^TX+\lambda I_d)^{-1}(X^TX)\cdot\hat{w}\right]$$

$$\overset{\text{מלינאריות התוחלת}}{=} (X^TX+\lambda I_d)^{-1}(X^TX)\cdot E(\hat{w}) \overset{\text{נתון } E(\hat{w})=w}{=} \underbrace{(X^TX+\lambda I_d)^{-1}X^TX\cdot w}$$

ניתן לראות בנוסחה שלכל $\lambda>0$ שנבחר נקבל מטריצה שאינה המטריצה ההופכית של X^TX וכשנכפיל אותן נקבל מטריצה שאינה מטריצת היחידה ולכן כשנכפיל ב w נקבל וקטור שאינו w כלומר ה $bias\neq 0$ ולכן $E(\hat{w}_x]\neq w$ אם $\lambda>0$ $\boxtimes$

c) $$Var(\hat{w}_x) \overset{\text{סעיף } a}{=} Var(A_x\hat{w}) \overset{\text{מהרצאה}}{=} A_x\, Var(\hat{w})\, A_x^T$$

$$\overset{\text{מהרצאה}}{=} A_x \underbrace{\sigma^2}_{\text{סקלר}} (X^TX)^{-1}A_x^T \overset{\text{חוקי מטריצות}}{=} \sigma^2 A_x (X^TX)^{-1}A_x^T$$ כנדרש $\boxtimes$

d) $$MSE(\hat{x}_x) = E\left[\|\hat{w}_x - w\|^2\right] = E\left[(\hat{w}_x-w)^T(\hat{w}_x-w)\right]$$

$$\overset{\text{שיקולי הסתברות}}{=} \|E(\hat{w}_x-w)\|^2 + E\left(\|\hat{w}_x - E(\hat{w}_x)\|^2\right)$$

$$\underbrace{\qquad\qquad}_{bias^2} \qquad \underbrace{\qquad\qquad}_{Var}$$

$$bias^2(\lambda) = \left(E(\hat{w}_\lambda) - E(\hat{w})\right)^2 = \left((A_\lambda - I)w\right)^2$$

$$Var(\lambda) = \sigma^2 \, Tr\left(A_\lambda (X^T X)^{-1} A_\lambda^T\right) =$$

סכום השונות של כל קורדינטה על האלכסון

$$\Downarrow$$

$$\boxed{MSE(\hat{x}_\lambda) = \|(A_\lambda - I)w\|^2 + \sigma^2 \, Tr\left(A_\lambda (X^T X)^{-1} A_\lambda^T\right)}$$

e) מאחר והנגזרת של ביטוי השגיאה שלילית ב $\lambda=0$ אז הפונקציה נמצאת בשיפוע שלילי והגדלת λ תקטין את ערכה הכללי. זה קורה בגלל שב $\lambda=0$ ה bias אמנם הכי נמוך אבל ה Variance גבוה וכשמגדילים את λ הירידה ב Var תשפיע יותר מהעלייה ב bias ועדיין גדול כך שהשגיאה שלילית ולכן אם המודל הלינארי נכון מעט רגולריזציית ridge תעזור לשפר את השגיאה הכללית.

1.2-

4) $X \in \mathbb{R}^{m\times d}$, $y \in \mathbb{R}^{m}$, $L \in \mathbb{R}^{k\times d}$, X^TX - הפיכה

נרצה למצוא נוסחה סגורה ל $\underset{w\in\mathbb{R}^d}{\operatorname{argmin}}\left(\|Xw-y\|^2+\|Lw\|^2\right)$

נגזיר את הפונקציה ונשווה ל0 וכך נמצא את המינימום בדומה לפרק בו עשינו בתרגול Ridge.

$$\|\cancel{X}w-y\|^2+\|Lw\|^2=(Xw-y)^T(Xw-y)+(Lw)^T(Lw)$$

$$\overset{\text{כפל מקוצר}}{=} w^TX^TXw-2y^TXw+y^Ty+w^TL^TLw$$

$$= w^T(X^TXw+L^TLw)-2y^TXw+y^Ty$$

$$= w^T(X^TX+L^TL)w-2y^TXw+y^Ty$$

כעת, נגזור לפי w: $\quad w^TAw = 2Aw,\ b^Tw = b$

$$\nabla = 2(X^TX+L^TL)w-2X^Ty$$

נשווה ל-0 כדי למצוא מינימום!

$$\cancel{2}(X^TX+L^TL)w=\cancel{2}X^Ty$$

$$\hat{w}_L^{TIK} = \boxed{w=(X^TX+L^TL)^{-1}X^Ty}$$

$\blacksquare$

b)

X^TX לא חייבת להיות הפיכה מאחר והתלות

של y היא ב X^TX+L^TL ולא רק ב X^TX כלומר

כדי שיהיה פתרון אנו צריכים ש X^TX+L^TL תהיה הפיכה.

ותנאי זה מתקיים כאשר L^TL היא מטריצת PSD מהוה'

מטריצת. ולכן תנאי זה שהוא חלש יותר מספיק לקיום פתרון.

c)

$$X=\begin{bmatrix}1&2\\3&4\end{bmatrix},\quad y=\begin{bmatrix}5\\6\end{bmatrix}\Rightarrow X^T=\begin{bmatrix}1&3\\2&4\end{bmatrix}$$

$$LS\to\hat{w}=(X^TX)^{-1}X^Ty$$

i)

$$\hat{w}=\left(\begin{bmatrix}1&3\\2&4\end{bmatrix}\begin{bmatrix}1&2\\3&4\end{bmatrix}\right)^{-1}\cdot\begin{bmatrix}1&3\\2&4\end{bmatrix}\cdot\begin{bmatrix}5\\6\end{bmatrix}$$

$$=\begin{bmatrix}10&14\\14&20\end{bmatrix}^{-1}\cdot\begin{bmatrix}1&3\\2&4\end{bmatrix}\begin{bmatrix}5\\6\end{bmatrix}=\frac{1}{200-196}\begin{bmatrix}20&-14\\-14&10\end{bmatrix}\cdot\underbrace{\begin{bmatrix}1&3\\2&4\end{bmatrix}\begin{bmatrix}5\\6\end{bmatrix}}^{X^Ty}$$

$$=\frac{1}{4}\begin{bmatrix}20&-14\\-14&10\end{bmatrix}\cdot\begin{bmatrix}23\\34\end{bmatrix}=\begin{bmatrix}5&-\frac{14}{4}\\-\frac{14}{4}&2.5\end{bmatrix}\begin{bmatrix}23\\34\end{bmatrix}=\boxed{\begin{bmatrix}-4\\4.5\end{bmatrix}}$$

ii) $\underline{\lambda=1}$

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

חישבנו בסעיף הקודם:

$$X^T y = \begin{bmatrix} 23 \\ 34 \end{bmatrix}$$

טוב נמשיך מפה:

$$X^T X = \begin{bmatrix} 10 & 14 \\ 14 & 20 \end{bmatrix}$$

$$\hat{w}_X^{Ridge} = \left(\begin{bmatrix} 10 & 14 \\ 14 & 20 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)^{-1} \cdot \begin{bmatrix} 23 \\ 34 \end{bmatrix}$$

$$= \begin{bmatrix} 11 & 14 \\ 14 & 21 \end{bmatrix}^{-1} \begin{bmatrix} 23 \\ 34 \end{bmatrix} = \frac{1}{21 \cdot 11 - 196} \begin{bmatrix} 21 & -14 \\ -14 & 11 \end{bmatrix} \begin{bmatrix} 23 \\ 34 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{21}{35} & -\frac{14}{35} \\ -\frac{14}{35} & \frac{11}{35} \end{bmatrix} \begin{bmatrix} 23 \\ 34 \end{bmatrix} = \boxed{\begin{bmatrix} 0.2 \\ 1.48 \end{bmatrix}}$$

iii) $$\hat{w}^{Tik} = (X^T X + L^T L)^{-1} X^T y \qquad L = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}$$

נחשב את כל החלקים

ונשתמש במה שכבר חישבנו:

$$L^T = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}, \quad X^T X = \begin{bmatrix} 10 & 14 \\ 14 & 20 \end{bmatrix}, \quad X^T y = \begin{bmatrix} 23 \\ 34 \end{bmatrix}$$

$$L^T L = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 4 \end{bmatrix}$$

$$\hat{w}^{Tik} = \left(\begin{bmatrix} 10 & 14 \\ 14 & 20 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 4 \end{bmatrix} \right)^{-1} \cdot \begin{bmatrix} 23 \\ 34 \end{bmatrix} = \begin{pmatrix} 10 & 14 \\ 14 & 24 \end{pmatrix}^{-1} \cdot \begin{bmatrix} 23 \\ 34 \end{bmatrix}$$

$$= \frac{1}{240 - 196} \begin{bmatrix} 24 & -14 \\ -14 & 10 \end{bmatrix} \cdot \begin{bmatrix} 23 \\ 34 \end{bmatrix} = \begin{bmatrix} \frac{24}{44} & -\frac{14}{44} \\ -\frac{14}{44} & \frac{10}{44} \end{bmatrix} \cdot \begin{bmatrix} 23 \\ 34 \end{bmatrix} = \begin{bmatrix} 1.72 \\ 0.41 \end{bmatrix}$$

i) $||\hat{w}|| = \sqrt{4^2+4.5^2} = 6.02$

$$||\hat{w}^{Ridge}_{\lambda=1}|| = \sqrt{0.2^2+1.48^2} = 1.493$$

$$||\hat{w}^{Tik}|| = \sqrt{1.72^2+0.41^2} = 1.765$$

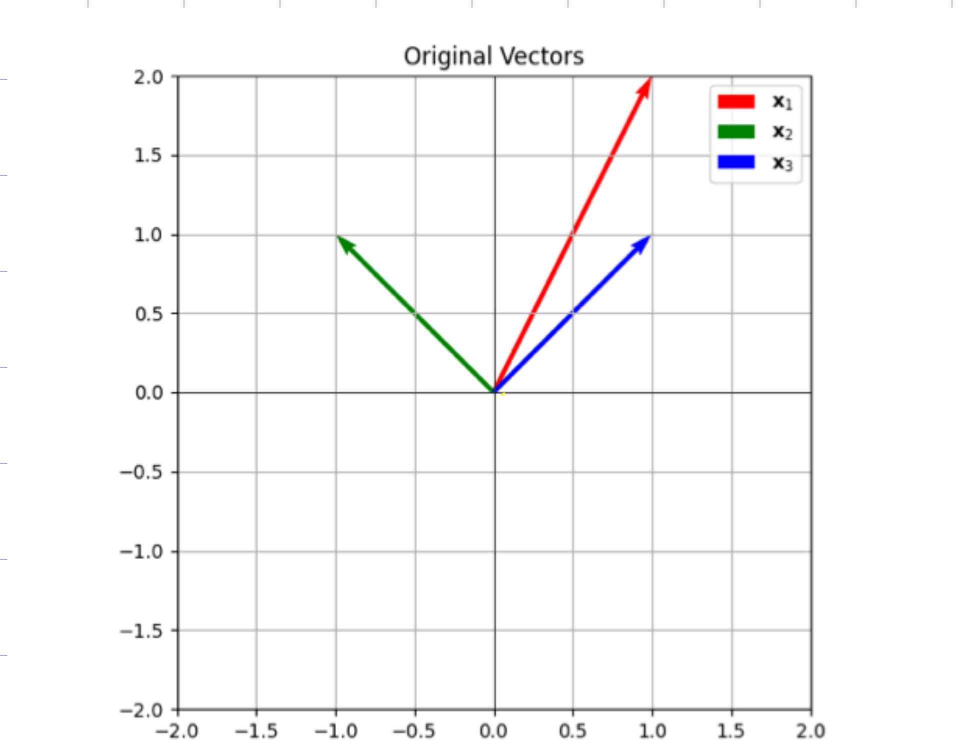
$$\boxed{||\hat{w}^{Ridge}_{\lambda=1}|| < ||\hat{w}^{Tik}|| < ||\hat{w}||}$$

LS ללא רגולריזציה בחר את הנורמה הכי גדולה מאחר והוא אינו מעניש כלל את גודל המשקלים והם יכולים להיות כרצונם. Ridge מתייחס לעונשה שווה לכל הקואורדינטות ולכן החזיר במקרה זה את הנורמה הכי נמוכה מאחר והוא כיווץ את שתי הקואורדינטות, Tikonov מעניש קואורדינטות ספציפיות במקרה שלנו בשונה לפי איך ש L בנויה ולכן כיווץ בעיקר אותה ולכן הנורמה בגודל אמצעי בין שתי הנורמות האחרות.

ii) כפי שפירטתי בסעיף הקודם LS ללא רגולריזציה אינו מעניש את גודל המשקולות ולכן ערך הקואורדינטה השנייה שלו מקסימלי, אחריו Ridge שמעניש באופן שקול את כל הקואורדינטות ומכווץ את כולן באופן שווה ואחרון Tikonov שמענישים בצורה ג׳ישמית לפי המטריצה L שבמקרה שלנו בנויה להעניש רק את הקואורדינטה השנייה ומכווץ אותה הכי הרבה ולכן היא המינימלית.

1.2 - PCA -

$$x_1 = \begin{bmatrix}1\\2\end{bmatrix}, \quad x_2 = \begin{bmatrix}-1\\1\end{bmatrix}, \quad x_3 = \begin{bmatrix}1\\1\end{bmatrix}$$

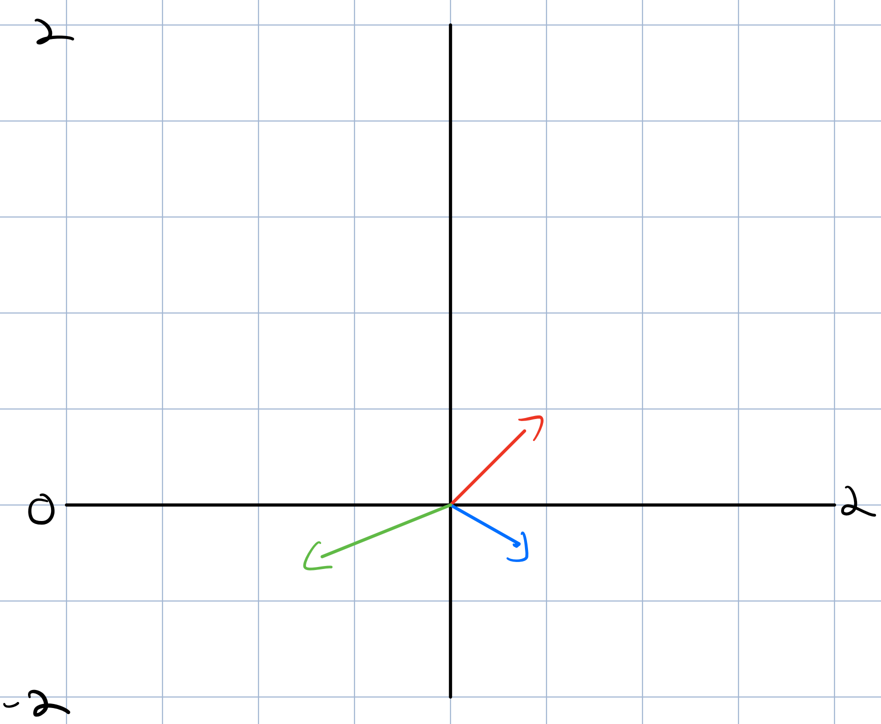


a) $$\bar{x} = \frac{1}{3} \sum_{i=1}^{3} x_i = \frac{1}{3} \begin{pmatrix}1-1+1\\2+1+1\end{pmatrix} = \begin{pmatrix}1\\4\end{pmatrix} = \boxed{\begin{pmatrix}\frac{1}{3}\\\frac{4}{3}\end{pmatrix}}$$

$$z_1 = \begin{pmatrix}1\\2\end{pmatrix} - \begin{pmatrix}\frac{1}{3}\\\frac{4}{3}\end{pmatrix} = \begin{pmatrix}\frac{2}{3}\\\frac{2}{3}\end{pmatrix}$$

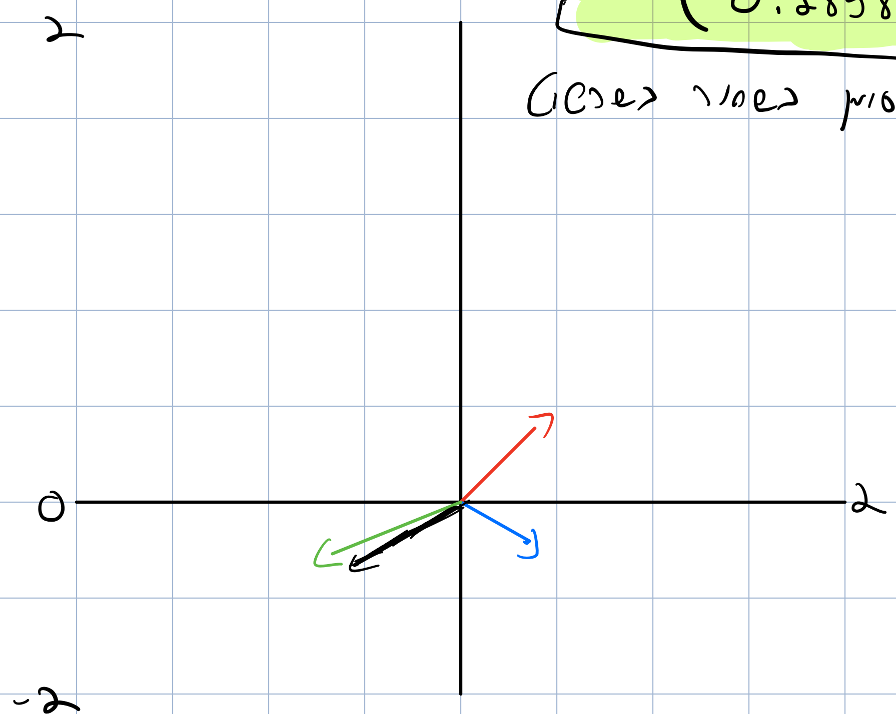
$$z_2 = \begin{pmatrix}-1\\1\end{pmatrix} - \begin{pmatrix}\frac{1}{3}\\\frac{4}{3}\end{pmatrix} = \begin{pmatrix}-\frac{4}{3}\\-\frac{1}{3}\end{pmatrix}$$

$$z_3 = \begin{pmatrix}1\\1\end{pmatrix} - \begin{pmatrix}\frac{1}{3}\\\frac{4}{3}\end{pmatrix} = \begin{pmatrix}\frac{2}{3}\\-\frac{1}{3}\end{pmatrix}$$



b)

$$\Sigma = \begin{bmatrix} \frac{8}{9} & \frac{2}{9} \\ \frac{2}{9} & \frac{2}{9} \end{bmatrix} \Rightarrow \underset{\text{ע"ע מקסימלי}}{\lambda = 0.9562} \Rightarrow \underbrace{\overset{\text{ו"ע מנורמל}}{V = \begin{pmatrix} -0.9571 \\ -0.2898 \end{pmatrix}}}_{\text{מסומן בשחור בגרף}}$$

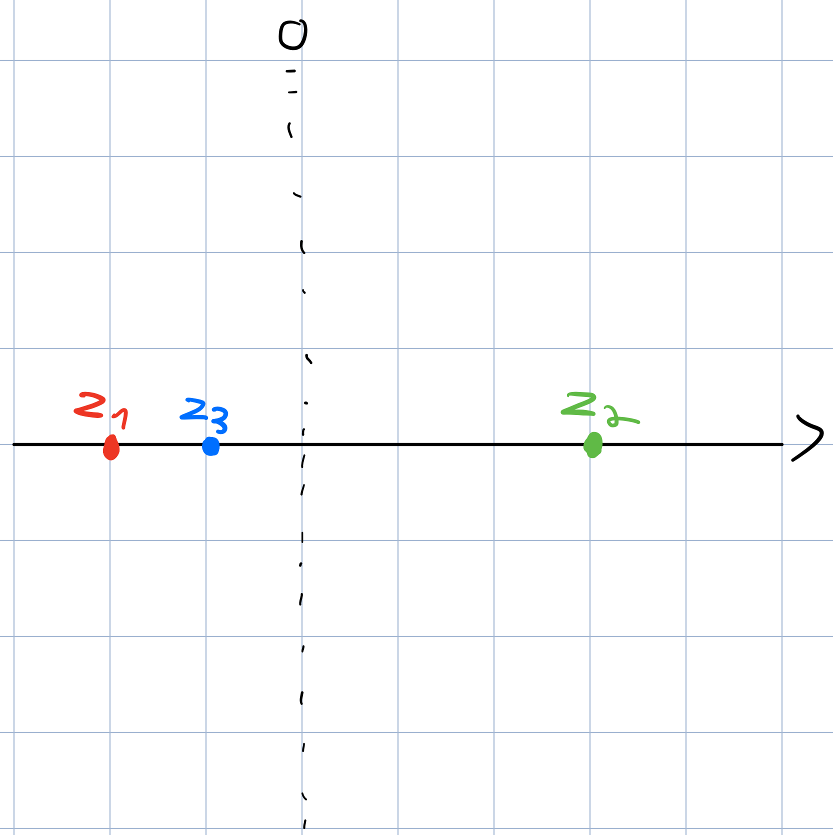


c)

$V^T z_1 = -0.83$

$V^T z_2 = 1.32$

$V^T z_3 = -0.48$

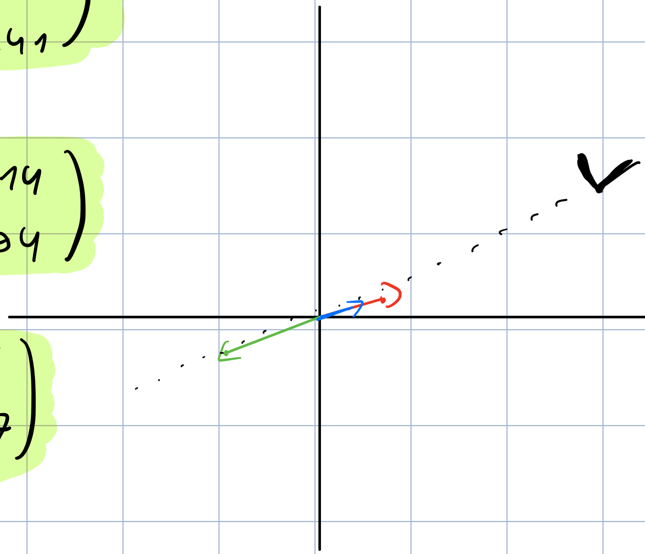


d)

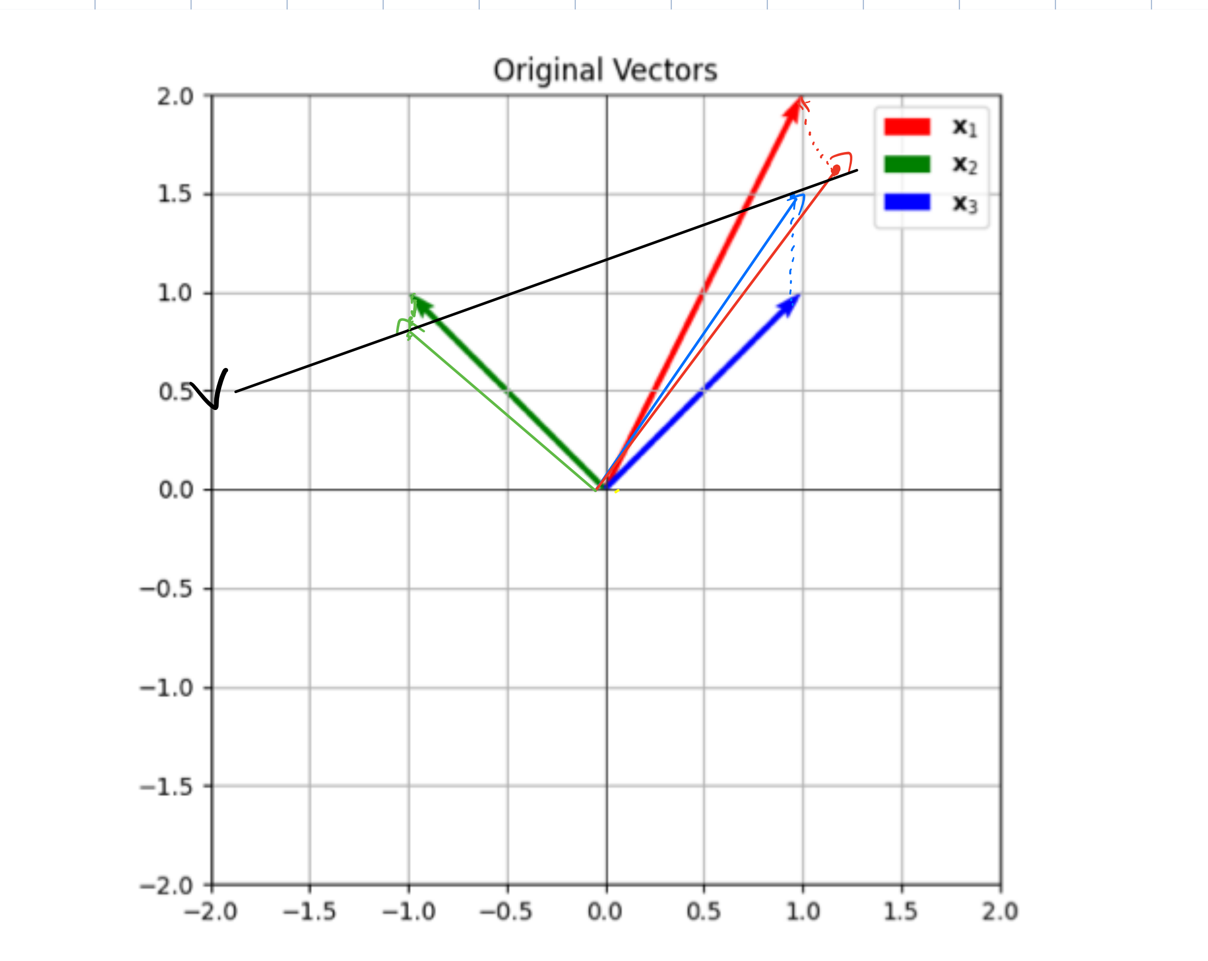
$$v(v^T z_1) = \begin{pmatrix}-0.9571\\-0.2898\end{pmatrix}\cdot -0.83 = \begin{pmatrix}0.796\\0.241\end{pmatrix}$$

$$v(v^T z_2) = \begin{pmatrix}-0.9571\\-0.2898\end{pmatrix}\cdot 1.32 = \begin{pmatrix}-1.314\\-0.394\end{pmatrix}$$

$$v(v^T z_3) = \begin{pmatrix}-0.9571\\-0.2898\end{pmatrix}\cdot -0.48 = \begin{pmatrix}0.518\\0.157\end{pmatrix}$$



e)



$$x_1^R = \begin{pmatrix}1.12\\1.574\end{pmatrix} \qquad x_2^R = \begin{pmatrix}-0.98\\0.93\end{pmatrix} \qquad x_3^R = \begin{pmatrix}0.85\\1.49\end{pmatrix}$$

$$v = \begin{pmatrix}\frac{1}{3}\\\frac{3}{4}\end{pmatrix} + \begin{pmatrix}-0.9571\\-0.2898\end{pmatrix}\alpha$$

2) $z_i = x_i - \bar{x}$, $W^T W$ אורתוגונלית

a)

צ"ל:

$$\sum_{i=1}^{m} \|x_i - \bar{x} - W^T W (x_i - \bar{x})\|^2 = \sum_{i=1}^{m} \|z_i\|^2 - z_i^T W^T W z_i$$

$z_i = x_i - \bar{x}$

$$\sum_{i=1}^{m} \|x_i - \bar{x} - W^T W (x_i - \bar{x})\|^2 = \sum_{i=1}^{m} \|z_i - W^T W z_i\|^2$$

$$= \sum_{i=1}^{m} \|z_i\|^2 - 2 z^T \cdot W^T W z_i + \|W^T W z_i\|^2$$

נוסחת כפל

$$= \sum_{i=1}^{m} \|z_i\|^2 - 2 z^T W^T W z_i + z_i^T W^T W W^T W z_i$$

תכונות נורמה

$W^T W$ אורתוגונלית

$W^T W W^T W = W^T W$

$$\Longrightarrow \sum_{i=1}^{m} \|z_i\|^2 - 2 z_i^T W^T W z_i + z_i^T W^T W z_i$$

חיבור

$$= \sum_{i=1}^{m} \|z_i\|^2 - z_i^T W^T W z_i \quad \blacksquare$$

b) $\sum_{i=1}^{m} z_i^T W^T W z_i = m \cdot Tr(W \Sigma W^T)$ צ"ל

$$a \in \mathbb{R}$$
$$Tr(a) = a$$

$$\sum_{i=1}^{m} z_i^T W^T W z_i = \sum_{i=1}^{m} Tr(z_i^T W^T W z_i)$$

$$\xRightarrow{Tr(AB)=Tr(BA)} \sum_{i=1}^{m} Tr(W z_i z_i^T W^T) \overset{\text{אלגברה}}{=} Tr\left(W \left(\sum_{i=1}^{m} z_i z_i^T\right) W^T\right)$$

$$\Sigma = \frac{1}{m} \sum_{i=1}^{m} (x_i - \bar{x})(x_i - \bar{x})^T = \frac{1}{m} \sum_{i=1}^{m} z_i z_i^T$$

$$\Longrightarrow Tr(W(m \cdot \Sigma) W^T) \overset{m - \text{סקאלר}}{\Longrightarrow} m \cdot Tr(W \Sigma W^T)$$

$\blacksquare$

c) $W = \begin{bmatrix} v_1^T \\ \vdots \\ v_k^T \end{bmatrix}$ $\quad Tr(W \Sigma W^T) = \sum_{j=1}^{k} v_j^T \Sigma v_j$ צ"ל:

$$Tr(W \Sigma W^T) = Tr\left(\begin{bmatrix} v_1^T \\ \vdots \\ v_k^T \end{bmatrix} \Sigma \cdot \begin{bmatrix} v_1 & \cdots & v_k \end{bmatrix}\right)$$

$$= Tr\begin{pmatrix} v_1^T \cdot \Sigma \cdot v_1 & \cdots & v_1^T \Sigma v_k \\ \vdots & \ddots & \\ v_k^T \Sigma v_1 & \cdots & v_k^T \Sigma \cdot v_k \end{pmatrix} \overset{\text{הגדרת Trace}}{=} \sum_{j=1}^{k} v_j^T \Sigma \cdot v_j$$

$\blacksquare$

d)

מה קיבלנו:

$$\overbrace{\sum_{i=1}^{m} \|x_i-\bar{x}-W^TW(x_i-\bar{x})\|^2}^{\text{שגיאת הבנייה מחדש}} = \sum_{i=1}^{m} \|z_i\|^2 - z_i^T W^T W z_i$$

$$= \underbrace{\sum_{i=1}^{m} \|z_i\|^2}_{\text{קבוע}} - \underbrace{\sum_{i=1}^{m} z_i^T W^T W z_i}$$

כלומר בשביל למזער את שגיאת הבנייה מחדש נרצה למקסם את $\sum_{i=1}^{m} z_i^T W^T W z_i$

מסעיפים ב׳ וג׳ קיבלנו ש: אם $W=\begin{bmatrix} v_1^T \\ \vdots \\ v_k^T \end{bmatrix}$ אז,

$$\sum_{i=1}^{m} z_i^T W^T W z_i = m\cdot Tr(W\Sigma W^T) = m\cdot \sum_{j=1}^{k} v_j^T \Sigma v_j$$

שזאת בדיוק הגדרת בעיית מקסום השונות ולכן הראנו שהבעיות שקולות ■

3)

נתונים: $V\in\mathbb{R}^d$, $\|V\|_2=1$ $\quad \Sigma\in\mathbb{R}^{d\times d}$ $\quad X:\Omega\to\mathbb{R}^d$

$E[X]=0$

נסמן λ_1 הע״ע המקסימלי ב PCA ו v_1 הוקטור העצמי שלו, רק וקטור עצמי רלוונטי לבדיקה כי ה PCA מוריד למימד אחד כלומר למרחב הנפרש על ידי v_1.

בנוסף מהגדרת PCA מתקיים $v_1^T \Sigma v_1 = \lambda_1$

נרצה להראות שמתקיים לכל v שמקיים $\|v\|=1$

$$Var(\langle v,x\rangle) \le v_1^T \Sigma v_1 = \lambda_1$$

$$Var(\langle v,x\rangle) = E\left((v^T x)^2\right) = v^T E[xx^T]\, v \overset{\text{הגדרת } \Sigma}{=} v^T \Sigma v = \lambda_v$$

ומכיוון פעולת PCA בוחרת את הערך העצמי המקסימלי נקבל:

$$Var(\langle v,x\rangle) = \lambda_v \le \lambda_1 = v_1^T \Sigma v_1$$

כלומר השונות היא לכל היותר השונות שניבחרת ע"י PCA כנדרש

$\blacksquare$

1.3-clustering:

a) True

לכל m שנבחר יש אפשרות שKmeans יחזיר קלאסטר עם נקודה אחת שלפי פעולת האלגוריתם יכנס ל C_{final}.

b) False

Kmeans יכול לתת פתרונות שונים לפי דרך פעולתו (בחירו ברנדומליות חלוקה בפוזיציה ההושמה) ולכן כל האלגוריתם יחזיר פתרונות שונים ובפרט לא בהכרח את האופטימלי.

c) True

כשמגדילים את מספר הקלאסטרים בK-means אנחנו מגדילים את מספר הקלאסטרים ובכך מגדלים את מספר האפשרויות לכל דגימה ובתוצאה מכך השונות בתוך הקלאסטר אמנם תרד אך השונות הכללית של המגמות תגדל.

d) True

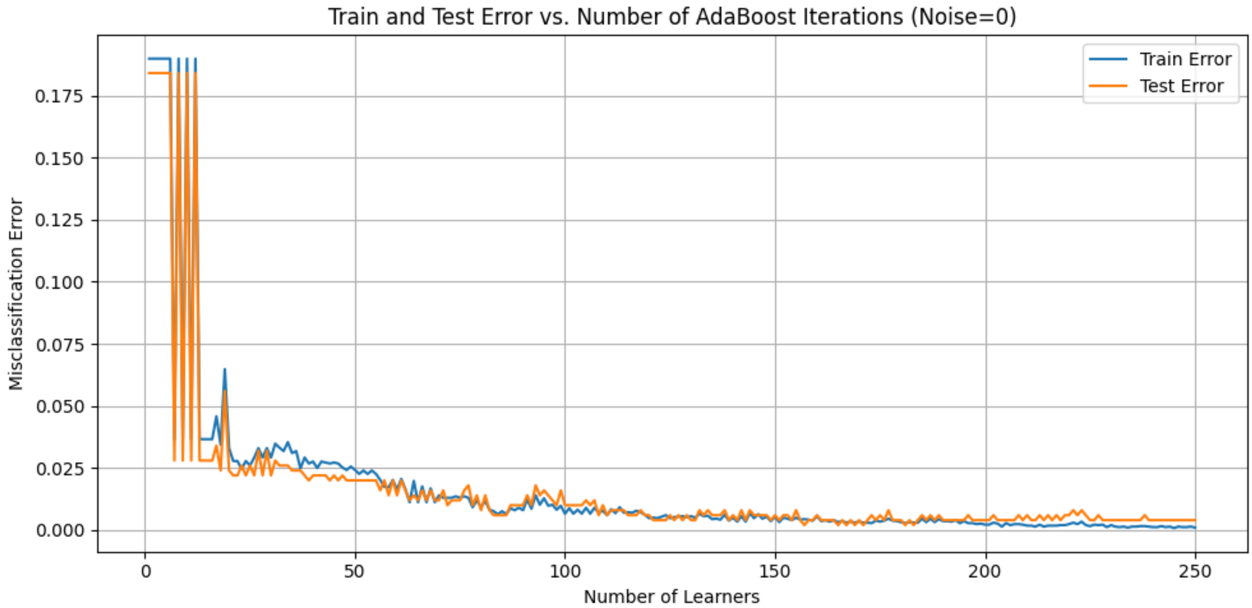
מכך שKmeans עצמו רגיש ל[illegible] האתחול שלו רנדומלי שהאלגוריתם ברנדומליות רגיש ויתן תוצאות שונות לפי דגם האתחול השונה.

Practical part Answers:

Boosting:

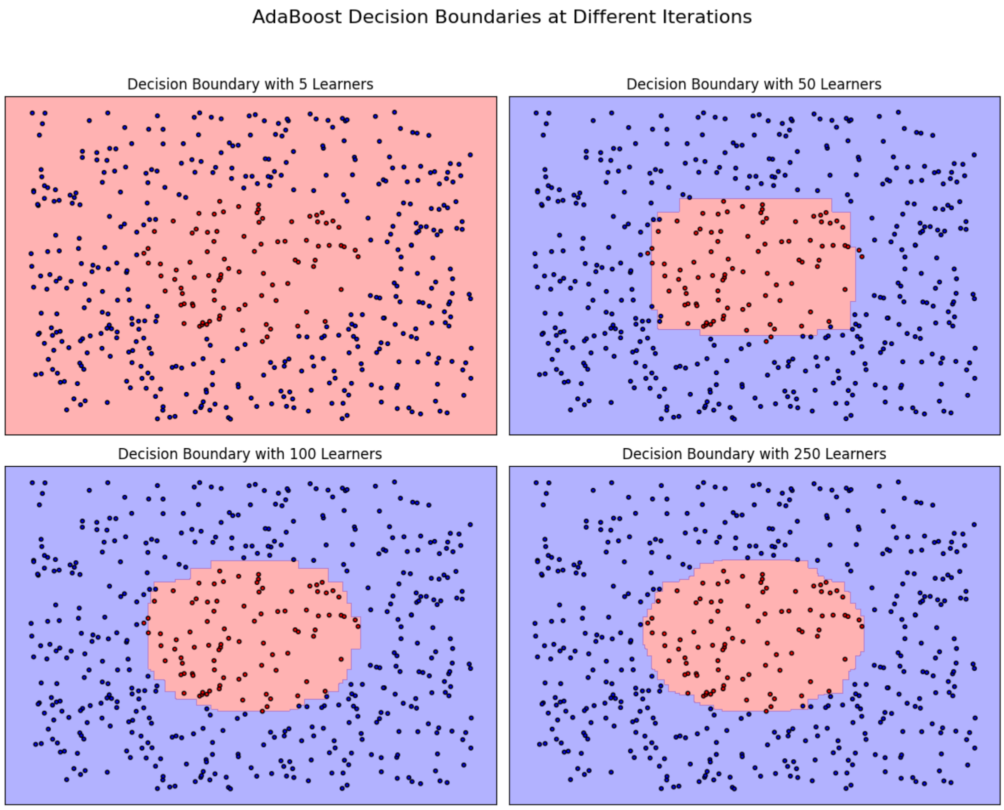
0 noise:

1)



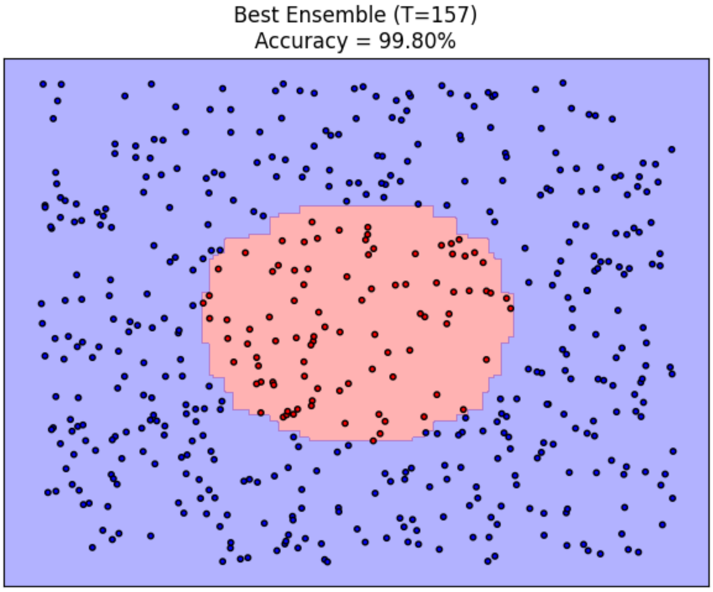
כפי שניתן לראות התוצאות תואמות את הציפיות- ככל שמספר הלומדים עולה שגיאת הקלאסיפקציה יורדת, בערכים הקטנים יחסית השגיאה מאוד תנודתית אך ממספר מסוים באזור ה25 לומדים השגיאה מתייצבת ומשתפרת באופן איטי מאוד כפונקציה של הגדלת מספר הלומדים. מגמה זאת הגיונית מאחר שממספר לומדים מסוים התוצאות כבר די מתמחזרות והלומדים חוזרים על עצמם ולכן מדייקים ברמה מאוד קטנה את השגיאה.

2)

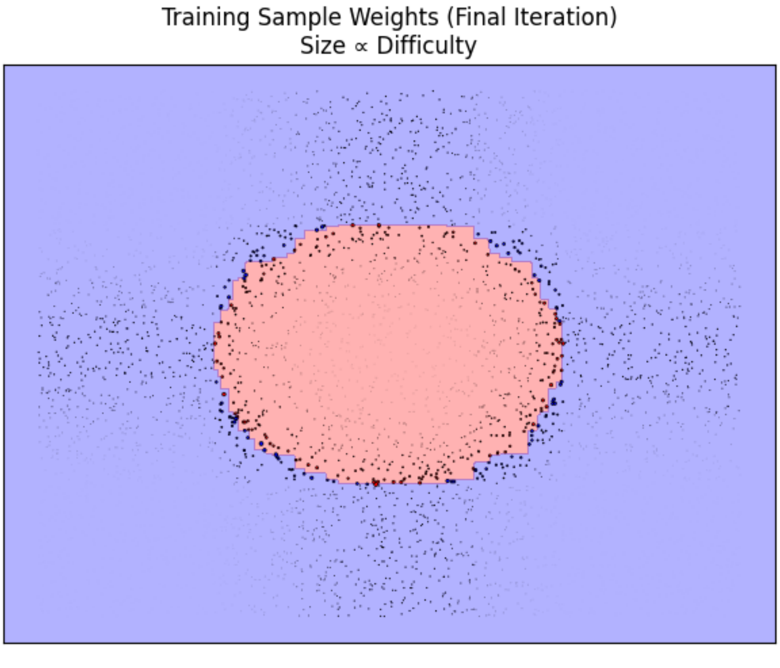


ניתן לראות שבמספר לומדים נמוך גבול ההחלטה היה כל הדאטה ולכן נתן שגיאה גדולה ולא מדויקת אך ככל שמספר הלומדים עולה הגבול מתהדק על הדאטה ונותן תוצאות איכותיות יותר. בנוסף ניתן לראות שהגבול עצמו נהיה יותר ויותר מעגלי בהתאם למספר הלומדים ופחות מרובע וזה הגיוני מאחר ויש למודל יותר גדמי החלטה לעבוד לפיהם כלומר יותר מלבנים לסמן על גרף ההחלטה ולכן הוא יכול לדייק יותר ויותר את הגבול האמיתי על הדאטה עצמו שנמצא בצורה מעגלית.

3)

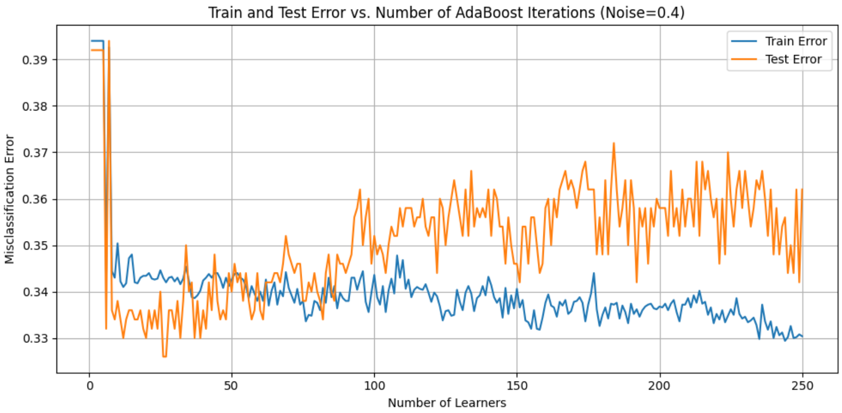


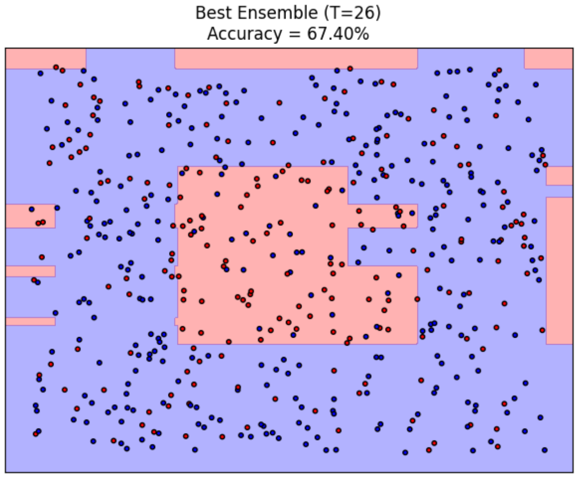
4)

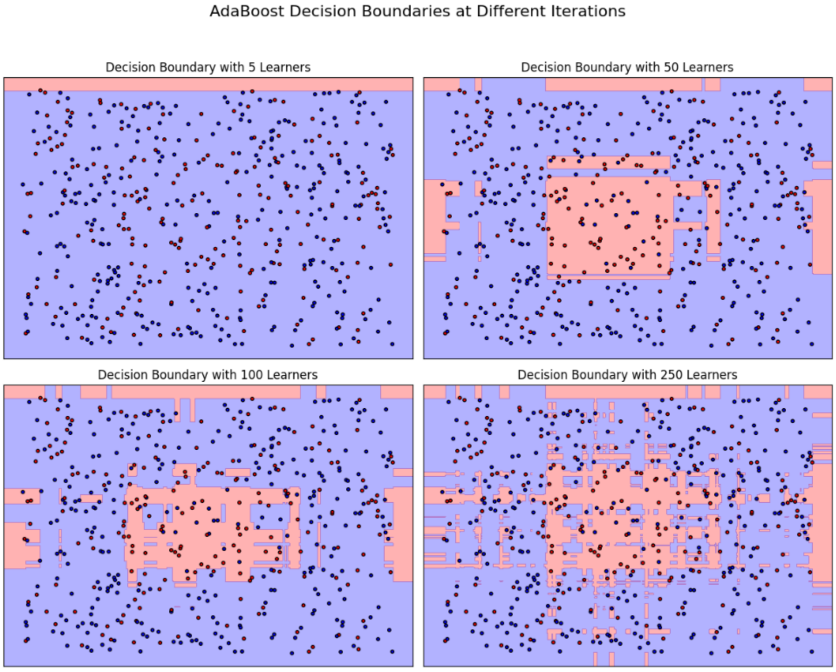


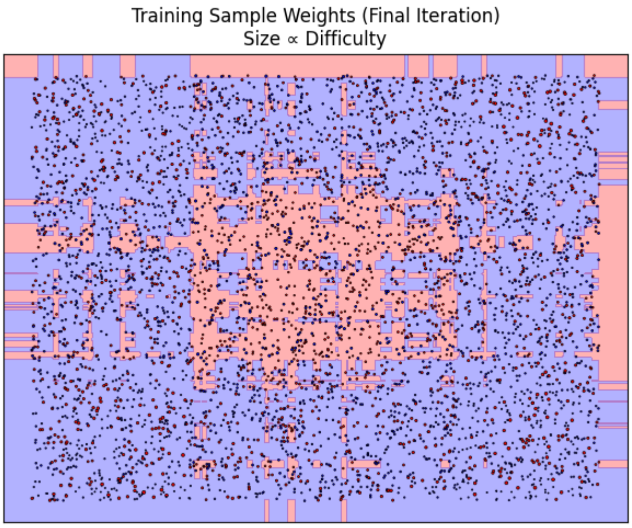
ניתן לראות בגרף הממושקל שלנקודות על התפר של גבול ההחלטה היו משקלים יותר גדולים בהרבה מהנקודות הרחוקות מהגבול, זאת תוצאה הגיונית מאחר ונקודות אלה הרבה יותר משמעותיות לגבול ההחלטה מאחר והדאטה בהן פחות חד משמעי ויותר חשוב לדייק בהן בקבלת הסיווג הסופי.

5)









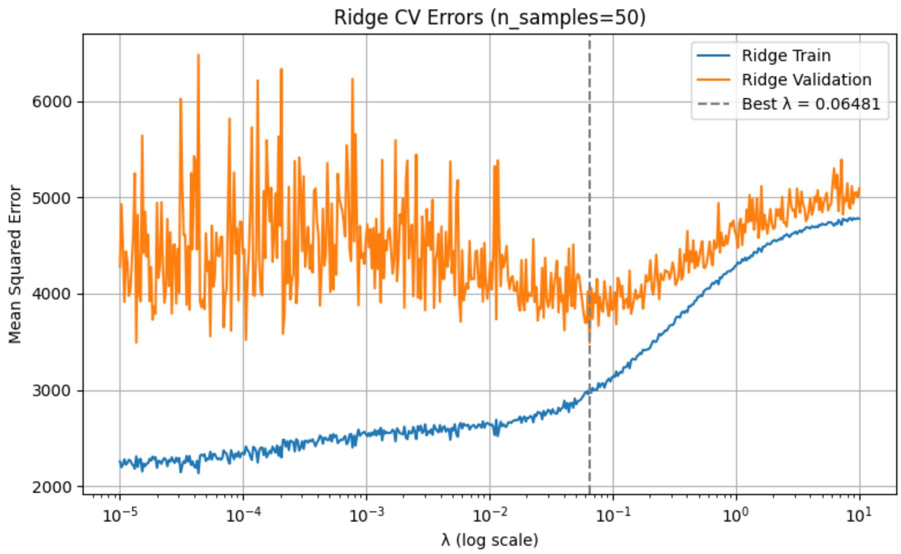
ניתן לראות מספר שינויים בגרפים כאשר הרעש הוא 0.4. ראשית המודל הרבה פחות מדויק באופן כללי כפי שניתן לראות בגרף הראשון והשלישי, שנית ככל שיש יותר לומדים מאחר והדאטה אינו מעגלי מושלם נוצרות יותר הפרדות בין הדגימות וגבול ההחלטה נהיה יותר ויותר מפוזר ומבולגן. שלישית בגרף הרביעי ניתן לראות שמשקל הדגימות מפוזר מאוד ולא בהכרח קרוב לגבולות ההחלטה כנראה בגלל הרעש שגורם לדגימות רבות לקבל משקל גבוה מהנדרש.

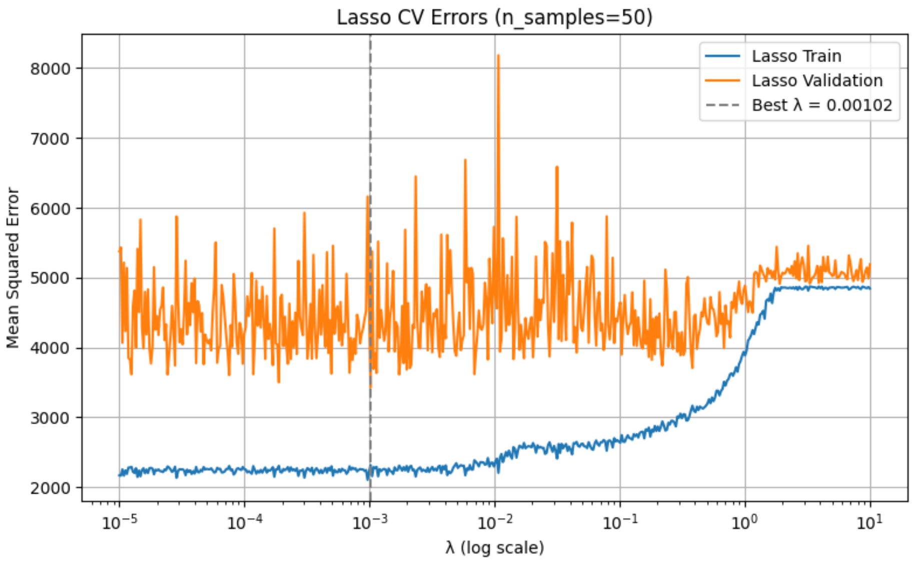
ניתן לראות בבירור בגרף הראשון את הביאס ואריאנס טרייד אוף, ככל שיש יותר לומדים השגיאה על דאטה האימון יורדת, כלומר הביאס יורד אך השגיאה על דאטה הטסט עולה- כלומר האוריאנס עולה. דבר המייצג באופן מדויק את הטרייד אוף ואת הסכנה באובר פיטינג.

Cross Validation:

2) תחילה ניסיתי למבדות בטווח של 0 עד 3 בקפיצות די גדולות וקיבלתי גרפים שלא כל כך הראו את המגמה כמו שצריך בנוסף ללמבדות אופטימליות מאוד קטנות ולכן החלטי ללכת על טווח למבדות בקפיצות מאוד קטנות ולייצג אותן בגרף בסולם לוגריתמי כדי להדגיש את ההבדל הקפיצות ולהראות את המגמה הנדרשת.

(3






```
Best Ridge λ = 0.06481, Test Error = 3423.47
Best Lasso λ = 0.00102, Test Error = 7305.27
Least Squares Test Error = 7991.27
```

ניתן לראות מספר דברים מהגרפים והתוצאות: ראשית בשני מודלי הרגולריזציה קיבלנו למבדה אופטימלי מאוד קטן שהביא לשיפור בשגיאה לעומת רגרסיה לינארית ללא רגולריזציה. שנית ניתן לראות שלדאטה המסוים שלנו רידג' הביא לתוצאה יותר טובה כנראה בגלל שיש יותר פיצ'רים בעלי קישוריות לינארית לתוצאות (מאופן פעולת רידג') . בנוסף נשים לב למגמתיות בטווחי הלמבדה של שני מודלי הרגולריזציה, בשניהם בתחילת העלייה בערכי הלמבדה הקטנים ניכר שיפור בשגיאה בדאטה הואלידציה כלומר המודל כולו השתפר אך החל מנקודת המינימום ניכרה עלייה כלומר המודל כבר לא משתפר ומתחיל להיות פחות טוב- זה מתכתב בדיוק עם החומר הנלמד בכיתה ומעיד על מציאת הלמבדה האופטימלי למודל. בנוסף נראה שבדאטה האימון ניכרת עלייה תמידית בשגיאה שזה הגיוני מאחר וככל שמשקל הרגולריזציה עולה הביאס עולה גם הוא ואיתו השגיאה על סט האימון. נשים לב שבמודל הלאסו החל מלמדה מסוים השגיאה מתייצבת שזה הגיוני מאחר ומודל זה מאפס פיצ'רים ולכן יכול לגרום להתייצבות השגיאה החל ממשקל מסוים.