תרגיל בית רטוב 3

318932365 – 315173344

תוכן

חלק 1

שאלה 1

הוכחה:

$$\frac{\partial l_\delta(w,b;x_i,y_i)}{\partial b} = \begin{cases} \dfrac{\partial \frac{1}{2}(w^T x_i + b - y_i)^2}{\partial b}, & |w^T x_i + b - y_i| \le \delta \\ \dfrac{\partial \delta(|w^T x_i + b - y_i| - \frac{1}{2}\delta)}{\partial b}, & o.w \end{cases}$$

$$= \begin{cases} (w^T x_i + b - y_i), & |w^T x_i + b - y_i| \le \delta \\ \dfrac{\partial \delta(w^T x_i + b - y_i - \frac{1}{2}\delta)}{\partial b}, & w^T x_i + b - y_i > \delta \ge 0 \\ \dfrac{\partial \delta(-w^T x_i - b + y_i - \frac{1}{2}\delta)}{\partial b}, & w^T x_i + b - y_i < -\delta \le 0 \end{cases}$$

$$= \begin{cases} w^T x_i + b - y_i, & |w^T x_i + b - y_i| \le \delta \\ \delta, & w^T x_i + b - y_i > \delta > 0 \\ -\delta, & w^T x_i + b - y_i < -\delta < 0 \end{cases}$$

$$= \begin{cases} w^T x_i + b - y_i, & |w^T x_i + b - y_i| \le \delta \\ \delta \cdot sign(w^T x_i + b - y_i), & o.w \end{cases}$$

שאלה 2

*פונקציית sign על וקטור מבצעת את פונקציית sign על סקלר לכל איבר בווקטור.

* אופרטור $\circ$ מסמל Hadamard product, כלומר כפל וקטורים איבר באיבר (element-wise).

* $\mathbb{1}_{\mathrm{m}\times 1\{|w^T x_i + b - y_i| \le \delta\}}$ מסמן וקטור אינדיקטור שאיברו ה-i הוא 1 אם מתקיים התנאי $|w^T x_i + b - y_i| \le \delta$, אחרת 0.

$$\nabla_w \mathcal{L}_H(w,b) = \nabla_w \cdot \frac{1}{m}\sum_{i=1}^{m} l_\delta(w,b;x_i,y_i) = \frac{1}{m}\sum_{i=1}^{m} \nabla_w l_\delta(w,b;x_i,y_i)$$

$$= \frac{1}{m}\sum_{i=1}^{m}\Big((w^T x_i + b - y_i)x_i \cdot \mathbb{1}_{\{|w^T x_i+b-y_i|\le\delta\}} + \delta \cdot sign(w^T x_i + b - y_i)\cdot x_i$$

$$\cdot \mathbb{1}_{\{|w^T x_i+b-y_i|>\delta\}}\Big)$$

$$= \frac{1}{m}X^T\Big((Xw + b1_m - y)\circ \mathbb{1}_{m\times 1\{|w^T x_i+b-y_i|\le\delta\}} + \delta sign(Xw + b1_m - y)$$

$$\circ\, \mathbb{1}_{m\times 1\{|w^T x_i+b-y_i|>\delta\}}\Big)$$

$$\frac{\partial \mathcal{L}_H(w,b)}{\partial b} = \frac{\partial}{\partial b}\frac{1}{m}\sum_{i=1}^{m} l_\delta(w,b;x_i,y_i) = \frac{1}{m}\sum_{i=1}^{m}\frac{\partial l_\delta(w,b;x_i,y_i)}{\partial b}$$

$$= \frac{1}{m}\sum_{i=1}^{m}(w^T x_i + b - y_i \cdot \mathbb{1}_{\{|w^T x_i+b-y_i|\le\delta\}} + \delta\cdot sign(w^T x_i + b - y_i)$$

$$\cdot \mathbb{1}_{\{|w^T x_i+b-y_i|>\delta\}}$$

$$= \left(\frac{1}{m}1_m\right)^T ((Xw + b1_m - y)\circ \mathbb{1}_{\mathrm{m}\times 1\{|w^T x_i+b-y_i|\le\delta\}} + \delta sign(Xw + b1_m - y)$$

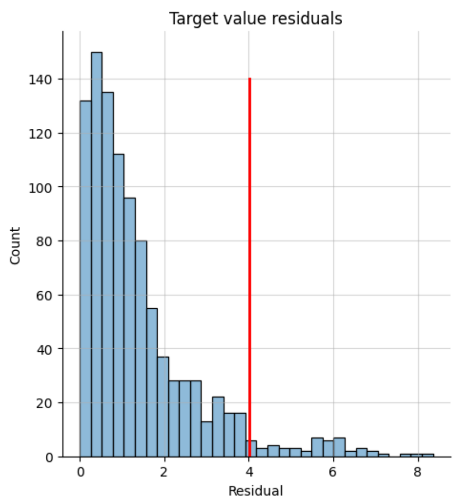
$$\circ\, \mathbb{1}_{\mathrm{m}\times 1\{|w^T x_i+b-y_i|>\delta\}}$$

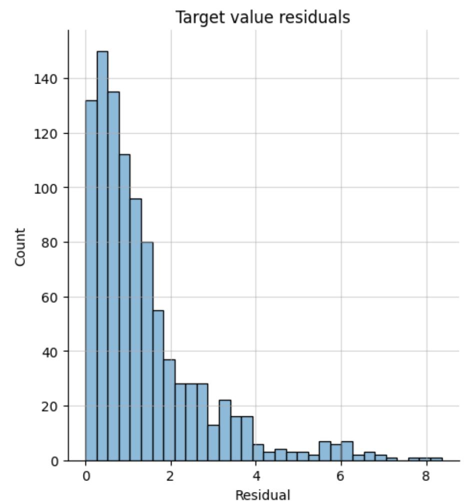
שאלה 3

1. נבצע נרמול של המידע כפי שביצענו בתרגילים הקודמים.
2. בעזרת linear regressor עם squared loss כפונקציית ההפסד, נמצא וקטור משקולות w וסקלר b בעזרתם נחשב את הערך המוחלט של כל ה-residuals (לכל i sample, נקבל את התוצאה של הרגרסור ונחסירה מה-label האמיתי). שימוש ברגרסור לינארי עם SL נותן לנו אפשרות להסתכל על הדאטה וכך לקבל מושג כללי אילו דגימות הן outliers.
3. נשתמש בפונקצייה percentile, בעזרתה, נמצא ערך delta שמגדיר חסם עליון (או threshold) ל-residuals שמחלק את הדאטה כך ש-95% מהמידע לא מוגדר כ outliers. ה-5% העליונים נחשבים כ outliers, שמהווים את ה"זנב" בגרף המצורף.

בחרנו ב-95% כיוון שבחירת ערך זה בתור percentage threshold הינה כלל אצבע מוכר בסטטיסטיקה[1].

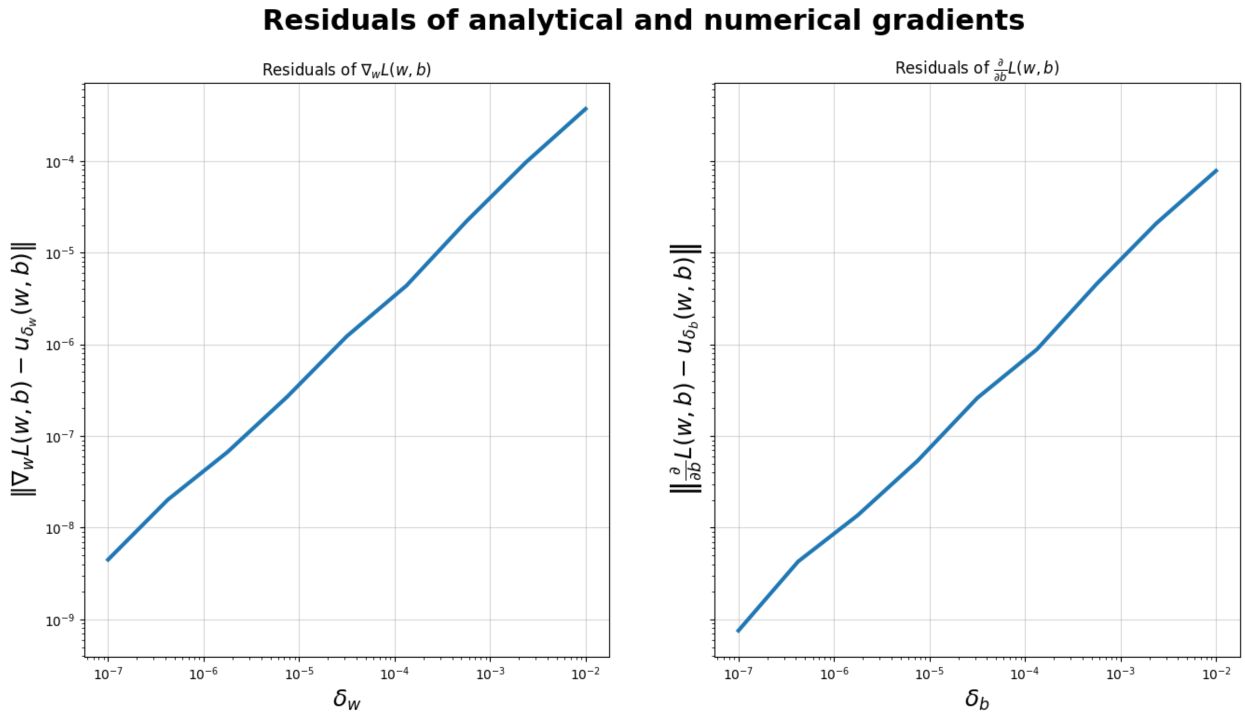
למשל, כאשר מדובר בהתפלגות נורמלית, 95% מהמידע שוכן בסביבת פעמיים סטיית התקן סביב התוחלת ושימוש ב-threshold זה הינו נפוץ בסטטיסטיקה והסתברות.



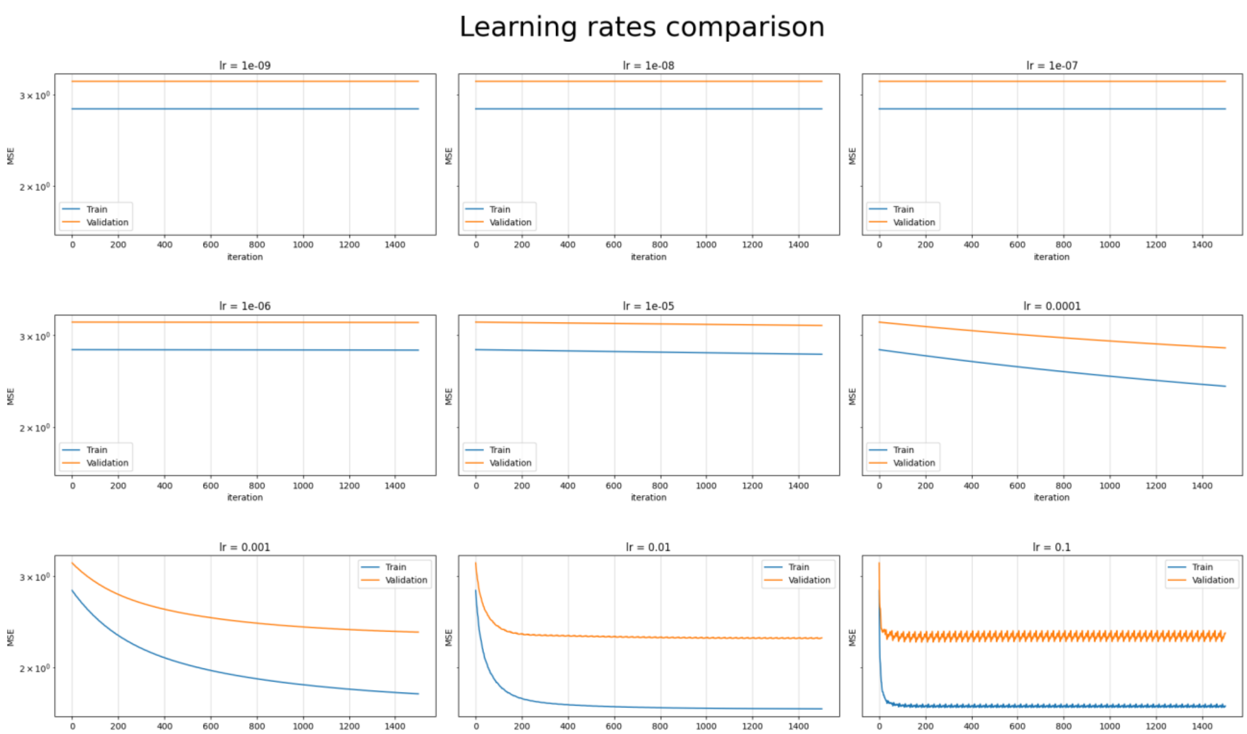


[1] https://tahera-firdose.medium.com/treating-outliers-using-iqr-and-percentile-approach-part-2-9d8c4ec55af7

שאלה 4


Residuals of analytical and numerical gradients
Residuals of $\nabla_w L(w, b)$
Residuals of $\frac{\partial}{\partial b} L(w, b)$
$\|\nabla_w L(w, b) - u_{\delta_w}(w, b)\|$
$\|\frac{\partial}{\partial b} L(w, b) - u_{\delta_b}(w, b)\|$
δ_w
δ_b

שאלה 5


Learning rates comparison
lr = 1e-09
lr = 1e-08
lr = 1e-07
lr = 1e-06
lr = 1e-05
lr = 0.0001
lr = 0.001
lr = 0.01
lr = 0.1
MSE
iteration
Train
Validation

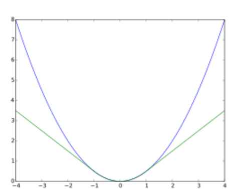
בתרשים המוצג לעיל, ישנן 3 התנהגויות עיקריות –

1. גרף פונקציית ההפסד בעל שיפוע קטן ואין התכנסות למינימום של פונ' ההפסד –
כיוון שה-lr קטן, מספר האיטרציות המקסימלי שהוזן איננו מספיק להצגת הדעיכה לערך ההפסד המינימלי. ה-lr'ים המתאימים להתנהגות זו הם lr=1e-9, 1e-8, 1e-7, 1e-6, 1e-5.
נציין כי עבור lr=1e-4 הגרף עדיין לא דועך לערך ההפסד המינימלי, עם זאת, ערך זה גדול דיו על מנת לראות את דעיכת ערך ההפסד בטווח האיטרציות שצוין.
2. דעיכה למינימום –
עבור lr=1e-3, 1e-2 ניתן לראות כי הגרף דועך לערך מינימלי, בקצבים שונים (כלומר לכל lr ההגעה לערך המינימום מתרחשת באיטרציה אחרת) ושהגרף נשאר על ערך זה לכל האיטרציות שלאחר מכן.
3. אוסילציות סביב ערך ההפסד המינימלי –
עבור ערך lr=0.1, הגרף מתחיל לבצע אוסילציות סביב הערך המינימלי (לפונקציה יש אותו ערך מינימום לכל lr אז בעזרת הגרף של lr=0.01 ניתן ללמוד מהו ערך זה) ולעולם לא תהיה התכנסות לערך המינימום בשל מחזוריות האוסילציות הללו.

ערך ה-lr האופטימלי לדעתנו הוא lr=0.01.
זהו הערך האופטימלי לדעתנו כיוון שעבור ערך זה מתרחשת התכנסות לערך המינימלי של פונקציית ההפסד, והתכנסות זו מתרחשת בקצב מהיר – פחות מ-200 איטרציות.

אין סיבה להוסיף איטרציות נוספות, נבחין כי פונקציית ה-huber loss היא קמורה (כפי שניתן לראות בתרשים למטה) לכל דלתא אי שלילית, ופונקציית המטרה שאנו מנסים להביא למינימום היא סכום של huber loss'ים והכפלת התוצאה בסקלר חיובי, לכן גם המטרה היא פונקציה קמורה כמו כן $\mathbb{R}^d, \mathbb{R}$ הם תחומים קמורים.
כיוון שהגרף מגיע למינימום והפונקציה קמורה, מינימום זה הוא גלובאלי והוספת איטרציות נוספות לא תועיל למציאת מינימום עדיף.

Huber loss



שאלה 6

Huber loss יעזור לנו במקרים בהם יש דאטה מורעש או דאטה שמגיע מהתפלגות עם "זנבות", כלומר עם הרבה outliers (או לחלופין, מעט outliers שהינם בעלי התרחקות קיצונית מהשאר). רגרסור לפי Huber loss הוא רובסטי בכך שהוא מתמודד עם מצבים כאלו ע"י מתן חשיבות מופחתת (ביחס ל-Squared loss) ל-outliers וזאת בשל העובדה שלערכים שנמצאים מחוץ לסביבת δ החישוב מתבצע בדומה לפונקציית ערך מוחלט שמקבלת ערכים קטנים יותר משל הפונקציה בריבוע. בניגוד לכך, לרגסור ע"פ OLS פונקציית loss אשר גדלה בצורה ריבועית ונותנת ל-outliers חשיבות רבה (בכך, שגיאה שגדולה פי שניים משגיאה אחרת כלשהי מוענשת פי 4 לעומת אותו ערך). על כן, במקרים כאלו רגרסור לפי Huber-loss יספק מודל טוב יותר שפחות מושפע מאותם רעשים או outliers.

שאלה 7

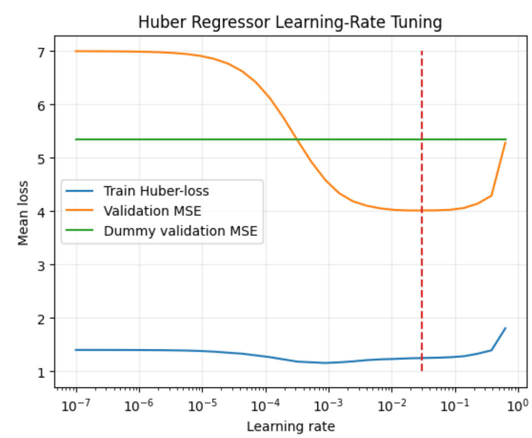
Model	Section	Train Huber Loss	Valid MSE
		Cross validation	
Dummy	2	Huber=1.4793 MSE=5.3272	5.3533

שאלה 8

i. ניקח כהיפר-פרמטר את learning-rates בטווח ערכים לוגריתמי בין 1e-7 ובין 1e-0.2.

ii. נבצע cross-validation על רגרסורים ליניאריים בעלי learning rate בטווח הנ"ל

iii. נקבל train loss ו- validation loss כפי שמוצגים בגרף הבא:



iv. נעדכן את הטבלה עם הערכים המתאימים:

Model	Section	Train Huber Loss	Valid MSE
		Cross validation	
Dummy	2	Huber=1.4793 MSE=5.3272	5.3533
Linear	2	1.2494	4.0138

הערך האופטימלי הוא LR=0.0305.

שאלה 9

אם לא ננרמל לפני ביצוע ה-cross validation, עבור ה-Dummy נצפה לשגיאת אימון (ולידציה) זהה, עבור המודל הלינארי נצפה לשגיאות גדולות יותר.

שגיאת ה-Dummy הינה זהה בשל אופי פעולתו. ה-Dummy מבצע "רגרסיה" על סמך ממוצע של פיצ'ר המטרה והנרמול לא מתבצע על פיצ'ר זה, על כן, אותה רגרסיה לא תושפע מהנרמול וחישוב השגיאה יהיה זהה בשני המקרים.

מאידך, המודל הלינארי מסתמך על SGD לשם מציאת המודל המנבא, אי ביצוע נרמול עשוי להשפיע לרעה על ביצועי ה-SGD, למשל, אי הגעה לערך המינימלי בתום אלף איטרציות SGD (כמות האיטרציות הדיפולטית). ללא הנרמול יש לדאטה מימדים שונים בעלי סדר גודל שונה, אבל משתמשים באותו learning-rate עבור כל המימדים. בשל קיום פיצ'ר (ממד) מסוים בעל סקאלת ערכים גדולה ביחס לאחרות, הוא יטה את הצעד ב-SGD לכיוון הממד הזה. זאת בעוד שלאחר נרמול, הנגזרות המכוונות לממד לעומת של הממדים האחרים עשויות להיות דומות בערכיהן. לכן, נרמול הדאטה יסייע לאופטימיזציה ע"י "איזון" בין הממדים השונים (נרמול) ושיפור קצב ההתכנסות של האלגוריתם.

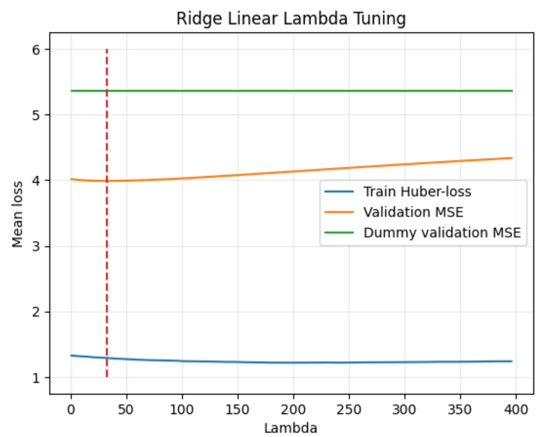
חלק 3

שאלה 10

i. ניקח כהיפר-פרמטר את lambda בטווח ערכים בין 1 ל-400.

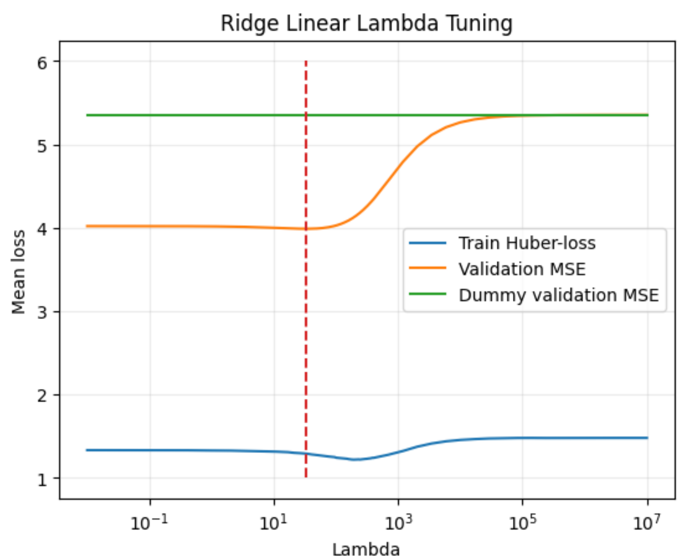
ii. נבצע cross-validation על Huber Regressors בעלי lambda בטווח הנ"ל

iii. נקבל train loss ו- validation loss כפי שמוצגים בגרף הבא:



הערך האופטימלי הוא lambda=33.

על מנת לקבל תמונה רחבה יותר ניסינו גם טווח ערכים רחב יותר עבור lambda (שלא שינה את תשובתנו, שכן האופטימלי נמצא בטווח 0-400):

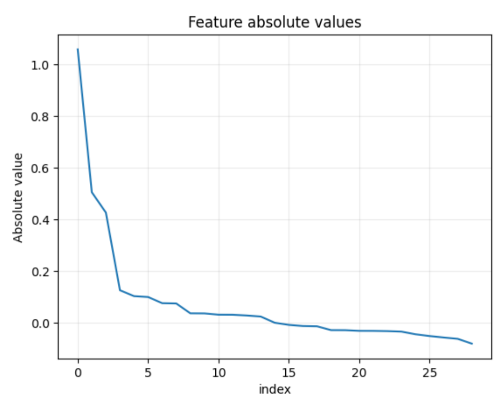


שאלה 11

.iv נעדכן את הטבלה עם הערכים המתאימים:

Model	Section	Train Huber Loss	Valid MSE
		Cross validation	
Dummy	2	Huber=1.4793 MSE=5.3272	5.3533
Linear	2	1.2494	4.0138
Ridge Linear	3	1.2903	3.9873

שאלה 12



שאלה 13

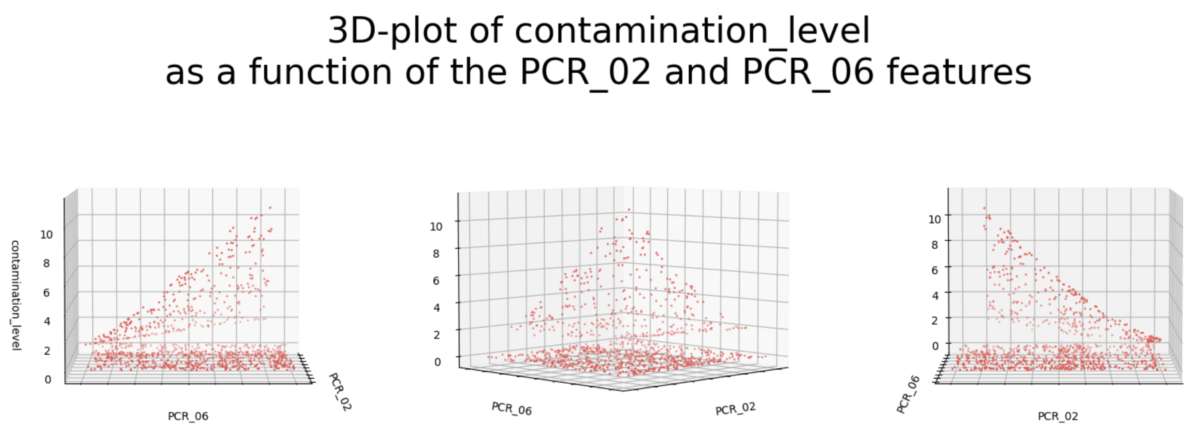
בהנחה שרצוננו לאפשר את יכולת הפרשנות המירבית לפיצ'רים השונים, כלומר לקבוע אילו פיצ'רים הם משמעותיים יותר/פחות לחיזוי משתנה מטרה כלשהו, ridge regularization לא יהווה את האפשרות המיטבית. הסיבה לכך היא שridge regularization משתמש כגורם רגולריזציה בנורמת L2, לכן האלגוריתם מנסה למצוא w שיביא למינימום את הגודל האוקלידי של w על פני כלל ערכיו, לכן הוא ישאף ל-w בעל ערכים קטנים, ללא הבדל בהכרח בין הפיצ'רים השונים. אפשרות עדיפה למיטוב הפרשנות היא שימוש ב-LASSO (l1 regularization), זאת משום שלרגולריזציה זו יש נטייה ל-sparsity, בכך, פיצ'רים שלא יתאפסו בווקטור המשקולות w יהיו פיצ'רים שנרצה לייחס להם חשיבות מוגברת בפרשנותנו לחשיבותם.

כפי שהוסבר במועד א' חורף 20-21 בסעיף ג' של שאלה 3, רגולריזציה עשויה לעזור לצמצם את מספר

המקדמים, זאת תוך שימור על הפיצ'רים החשובים, בשאלה ניתן לראות כי גודל המקדם לא בהכרח מעיד על חשיבותו של פיצ'ר. LASSO מתגברת על בעיה זו בכך שהיא מגבילה גדלים של מקדמים בהתאם לאילוץ שנקבע בהתאם ל-λ, זאת תוך שהיא מוצאת וקטור w דליל שמקרב את הבעיה למינימום.

חלק 4

שאלה 14

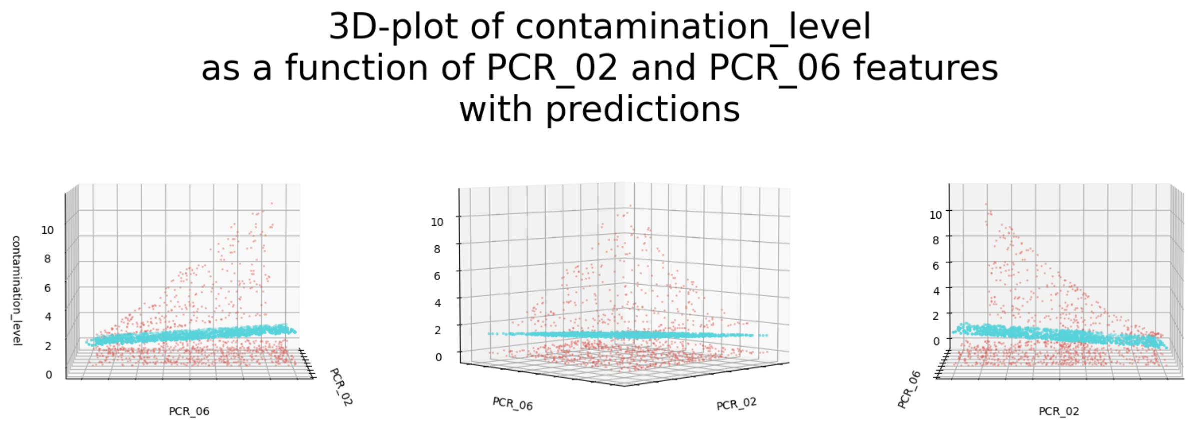


מהוויזואליזציה הנ"ל ניצן לראות כי ערכי משתני המטרה משתייכים לשני משטחים. עם זאת, נדמה כי נקודות הנמצאות בקאורדינטות (pcr_2, pcr_6) קרובות זו לזו עשויות לקבל ערכי contamination_level שונים (כלומר, להשתייך למשטחים שונים), ולכן לא ניתן למצוא regression של נקודה לערך contamination על סמך שני פיצ'רים אלו בלבד.

המודל הלינארי (במקרה של תחום דו מימדי) חוזה מישור בו כל נקודה ניתנת לייצוג על ידי צירוף לינארי של שני וקטורים בצירים pcr_2, pcr_6. כיוון שבוויזואליזציה לעיל ניתן לראות שני מישורים, לא ייתכן כי המודל הלינארי (אשר נותן מישור יחיד) יוכל למצוא קשר לינארי מוצלח בין שני הפיצ'רים הללו.

לאור כך, נרצה למצוא קשר יותר מורכב בין הפיצ'רים, לשם כך נפנה למודל פולינומיאלי, ייתכן כי לשם כך נצטרך לבצע feature mapping (כפי שנרמז בשם של חלק זה במטלה).

שאלה 16

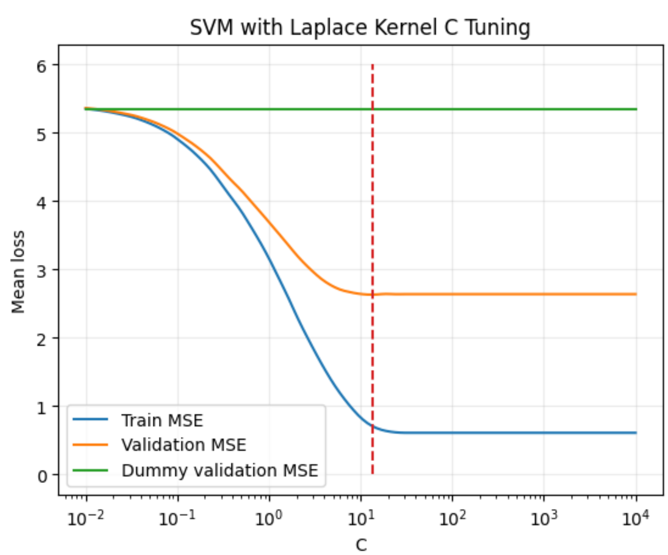


שאלה 17

i. ניקח כהיפר-פרמטר את C בטווח ערכים לוגריתמי טווח [1e-2,1e+4].

ii. נבצע cross-validation על Regressors עם Laplacian mapping בעלי C בטווח הנ"ל

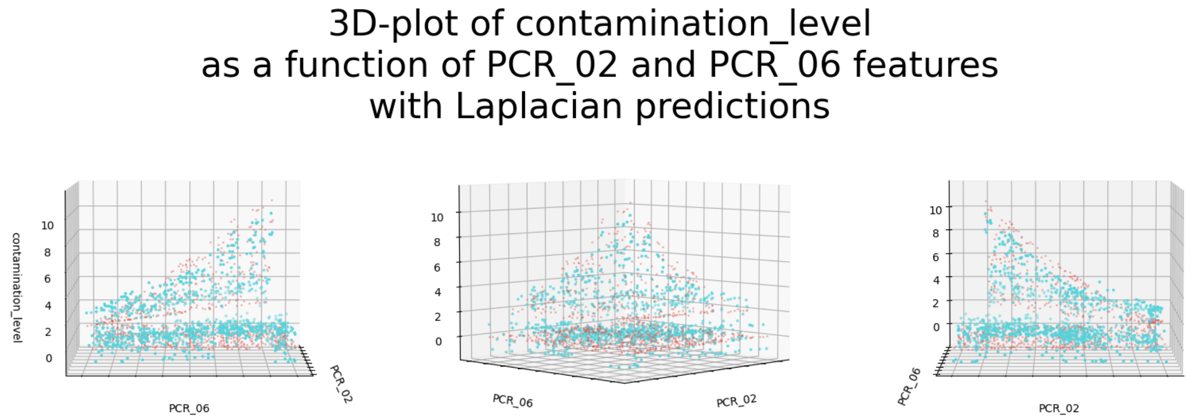
iii. נקבל train loss ו- validation loss כפי שמוצגים בגרף הבא:



הערך האופטימלי שנמצא הוא C=13.3675. ערכי השגיאה עבור רגרסור מאומן על כל הפיצ'רים עם C אופטימלי:

Validation MSE: 2.6315, Train MSE: 0.7078

שאלה 18



שאלה 19

לאחר אימון הרגרסון הפולינומיאלי המשתמש ב-SVR עם קרנל לפלסיאן על כל הפיצ'רים, נקבל ערך שגיאת אימון MSE=0.7078 וערך שגיאת ולידציה MSE=2.6315. בהשוואה בין הרגרסור הלינארי אשר משתמש ב-Huber Loss לבין הרגרסור הפולינומיאלי אשר משתמש ב-SVR בעל קרנל לפלסיאן, ניתן לראות כי תוצאות האחרון טובות יותר מאשר של הראשון. ניתן לראות שהרגרסיה השתפרה ושהערכים שהתקבלו קרובים יותר לערכי האמת, דבר זה בא לידי ביטוי הן בערכי השגיאה (על קבוצת האימון) אשר קטנו והן בוויזואליזציה בה ניתן לראות את התקרבות הנקודות הכחולות (הרגרסיה) לאדומות (ערכי האמת). זאת בהתאם לכך שכפי שראינו בשאלה 14, בה עלה כי לא קיים קשר לינארי יחיד בין קואורדינטות במישור (pcr_2, pcr_6) לבין ערך יחיד בציר z. השימוש של של הרגרסור ב-SVR בעל קרנל לפלסיאן מאפשר לו לתאר קשר מורכב יותר בין שני הפיצ'רים לבין ה-contamination_level שאינו לינארי.

חלק 6

שאלה 20

Model	Section	Train Huber Loss	Valid MSE	Test MSE
		Cross validation		Retrained
Dummy	2	Huber=1.4793 MSE=5.3272	5.3533	4.9403
Linear	2	1.2494	4.0138	3.7332
Ridge Linear	3	1.2903	3.9873	3.6728
SVR + Laplace	4	MSE= 0.7078	2.6315	2.4375

כפי שניתן לראות מהטבלה, המודל עם הביצועים הטובים ביותר על ה-test set הוא מודל ה-**SVR + Laplace**.