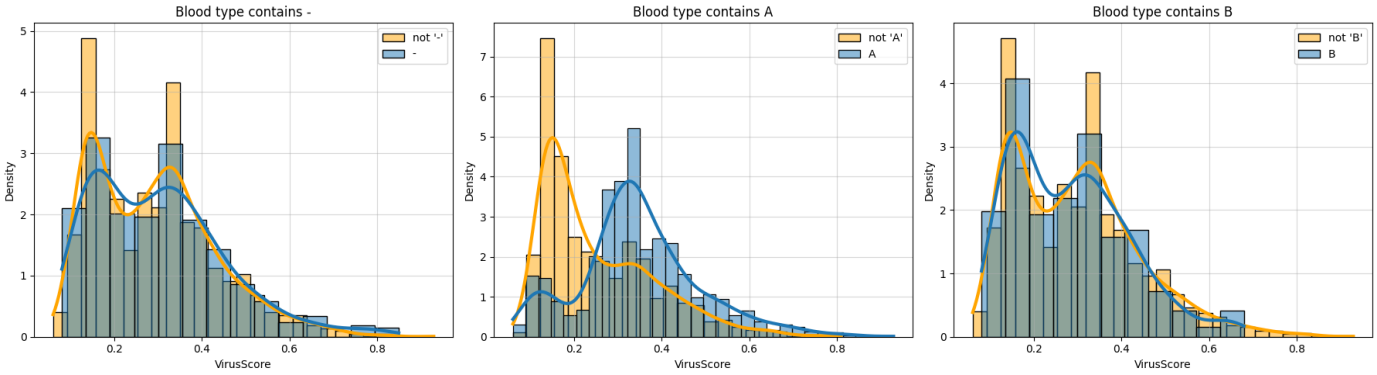
דו"ח

שאלה 1

הגרפים המתקבלים הם :



שאלה 2

ניתן לראות שהקשר הברור ביותר בין VirusScore לאחד מן הגרפים הוא הגרף המתאר את התכונה "A or not A" שזה הגרף האמצעי.

הגרף שמציג באופן ברור שההתפלגות של אנשים בעלי סוג דם A נותה לכיוון ערכים גבוהים יותר של VirusScore מאשר אנשים ללא A .
ולכן קישרנו בין אנשים עם סוג דם A+-,AB ל- VirusScore כאשר התכונה הבינארית שהוספנו בהתאם לתנאי זה היא :

$$blood\_viruse(x) = \begin{cases} 1 & x\ contains\ A \\ 0 & else \end{cases}$$

<u>שאלה 3</u>

להלן חישוב הנגזרת,

$$L(w,b) = \frac{1}{m}\left(X\underline{w} + \underline{1}b - \underline{y}\right)^T \left(X\underline{w} + \underline{1}b - \underline{y}\right) =$$

$$L(w,b) = \frac{1}{m}\left(\underline{w}^T X^T X\underline{w} + \underline{w}^T X^T \underline{1}b - \underline{w}^T X^T y + \underline{1}^T bX\underline{w} + \underline{1}^T \underline{1}b^2 - b\underline{1}^T \underline{y} - \underline{y}^T X\underline{w} - \underline{y}^T \underline{1}b + y^T y\right)$$

$$\frac{d\left(L(w)\right)}{dw} = \frac{1}{m}\left(\underline{w}^T X^T \underline{1}b + \underline{1}^T\ bX\underline{w} + \underline{1}^T \underline{1}b^T b - b\underline{1}^T \underline{y} - \underline{y}^T \underline{1}b\right)$$

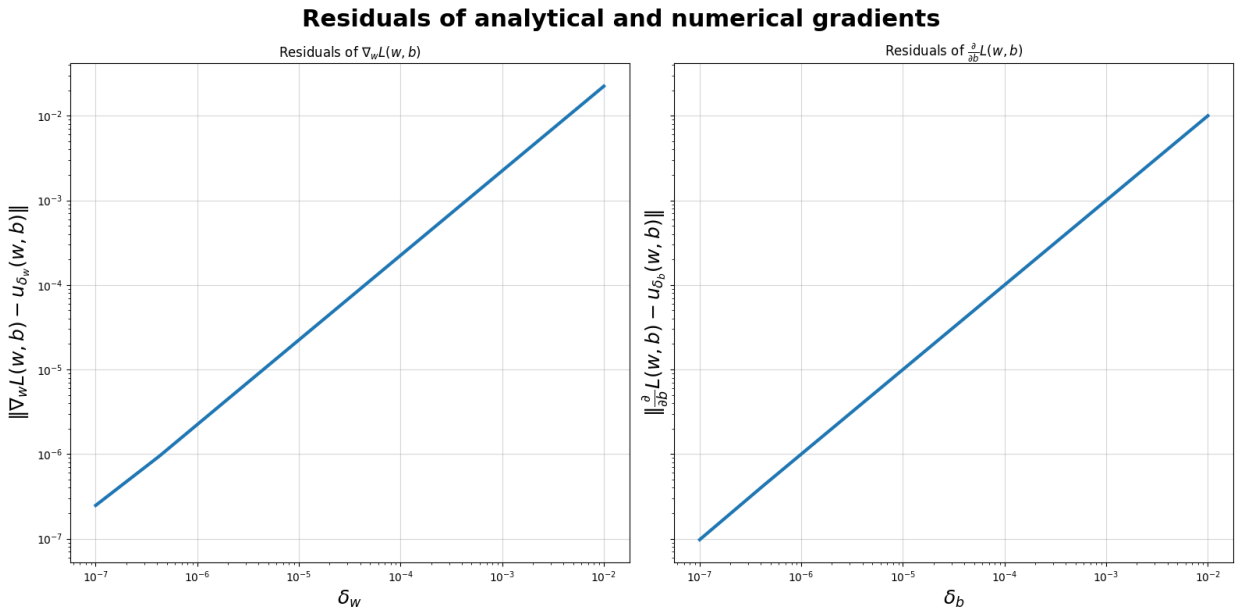
$$\frac{d\left(L(w)\right)}{dw} = \frac{1}{m}\left(\underline{w}^T X^T \underline{1} + \ \underline{1}^T X\underline{w} + 2 * \underline{1}^T \underline{1}b - \underline{1}^T \underline{y} - \underline{y}^T \underline{1}\right)$$

$$\underline{w}^T X^T \underline{1} = \ \underline{1}^T X\underline{w}\ , \underline{1}^T \underline{y} = \ \underline{y}^T \underline{1}$$

$$\frac{d\left(L(w)\right)}{dw} = \frac{1}{m}\left(2 * \ \underline{1}^T X\underline{w} + 2 * \underline{1}^T \underline{1}b - 2 * \underline{1}^T \underline{y}\right)$$

$$\frac{d\left(L(w)\right)}{dw} = \frac{2}{m} * \underline{1}^T (X\underline{w} + \underline{1} * b - \underline{y})$$

שאלה 4



ניתן לראות גם ב W וגם ב b שהגרפים שקיבלנו הינם עולים

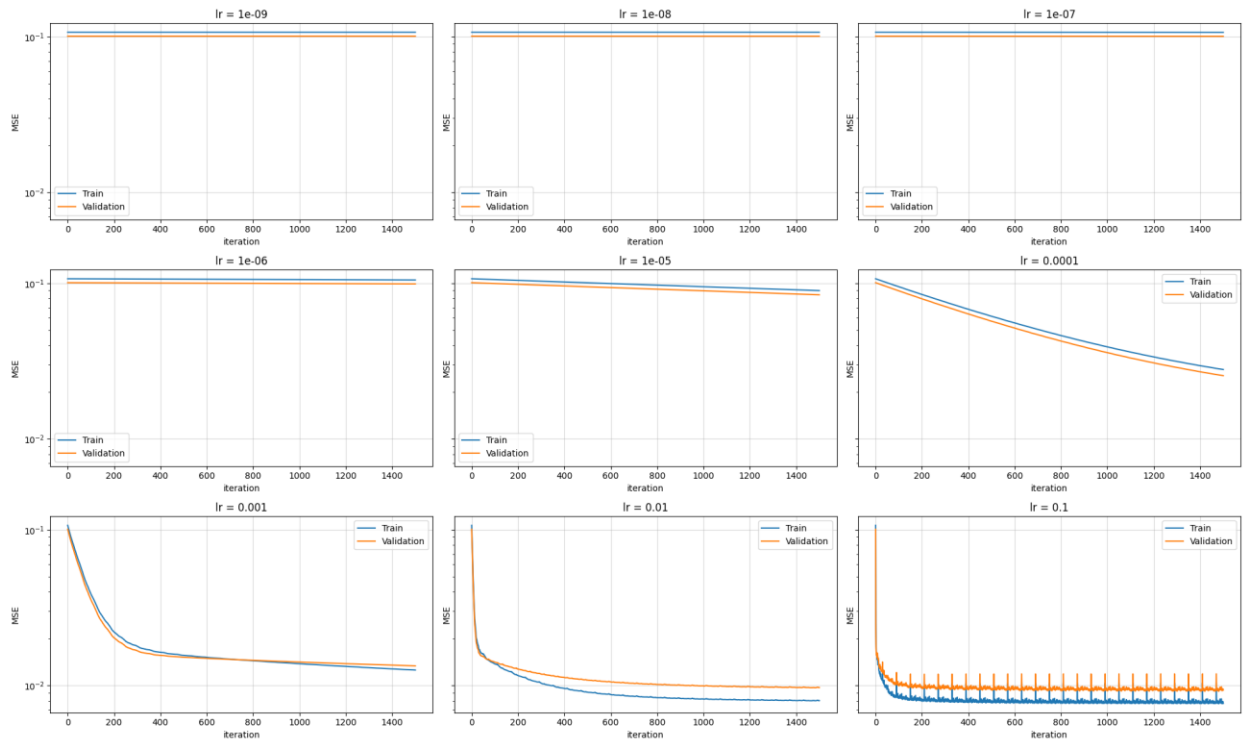
את זה ניתן להצדיק באופן הבא

כך שבכל שמקטינים את delta ההפרש בין הגרדיאנט ההאינלטי לגרדיאנט הנומרי קטן כי לפי הגדרת הנגזרת

$$f'(x) = \lim_{delta \to 0} \frac{f(x + delta) - f(x)}{x + delta - x} = \frac{f(x + delta) - f(x)}{delta}$$

ולכן אפשר להצדיק התנהגות זאת.

שאלה 5



נשים לב כי הבעיה שלנו היא קמורה ולכן נוכל נוכל להתקרב מאוד לפתרון האמיתי על ידי gd

נשים לב איך שינוי ה lr משנה את התנהגות הגרפים.

כאשר $lr \leq 10^{-5}$ ניתן לשים לב כי ההתכנסות היא מאוד איטית ומכאן קצב הלימוד הוא איטי ולכן על מנת להתכנס לפתרון נדרש להרבה יותר מחזורים. – נשים לב כי במקרה זה מרוב שאנחנו רחוקים מן הפתרון האימתי הגרף של ה valid נמוך מ train.

כאשר $lr = 10^{-4}, 10^{-3}$ רואים התכנסות הרבה יותר טובה ואכן מתחילה להיות החלפה בין ה valid לבין ה train (כלומר ה train נמוך יותר מ valid וזה הגיוני כי אנחנו מחפשים את הפתרון האופטימלי ב train)

ניתן לראות שהדיוק המירבי שהתקבל נמצא ב $\mathrm{lr} = 0.1$ אך במקרה הזה ניתן לראות שישנו רעש ולא ניתן להבטיח את התשובה – למעשה ה GD מבצע קפיצות מסביב לנקודת המינימום ולכן אין התכנסות למינימום במקרה זה.

ולכן הדיוק המריבי שנתעדף הוא עבור $\mathrm{lr} = 0.01$ שבמקרה זה הדיוק עלול להיות נמוך יותר אך נבטיח התכנסות.

lr size = 0.01, Best train loss = 0.007996124967117628,

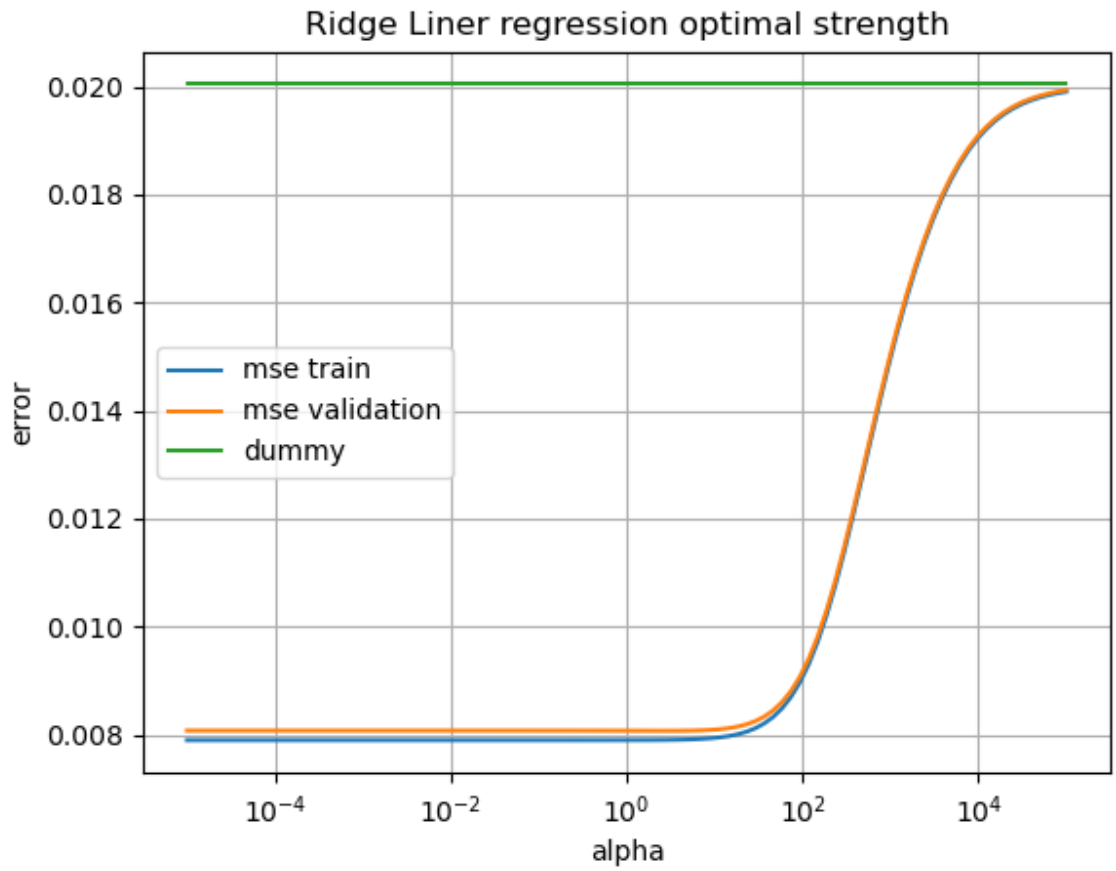
Best validation loss = 0.009689819450089773

lr size = 0.1, Best train loss = 0.007654105913138692, Best validation loss = 0.009199048010549963

שאלה 6

Model	Section	Train MSE	Valid MSE
		Cross Validated	
Dummy	3	0.02001108412543403	0.020040073426649308

שאלה 7



השגיאה הקטנה ביותר שהתקבלה היא : 0.008068282291301469

היא התקבלה עבור alpha אשר שווה ל: 2.848035868435799

שאלה 8

Model	Section	Train MSE	Valid MSE
		Cross Validated	
Dummy	3	0.02001108412543403	0.020040073426649308
Ridge	4	0.007899551580020458	0.008068282291301469

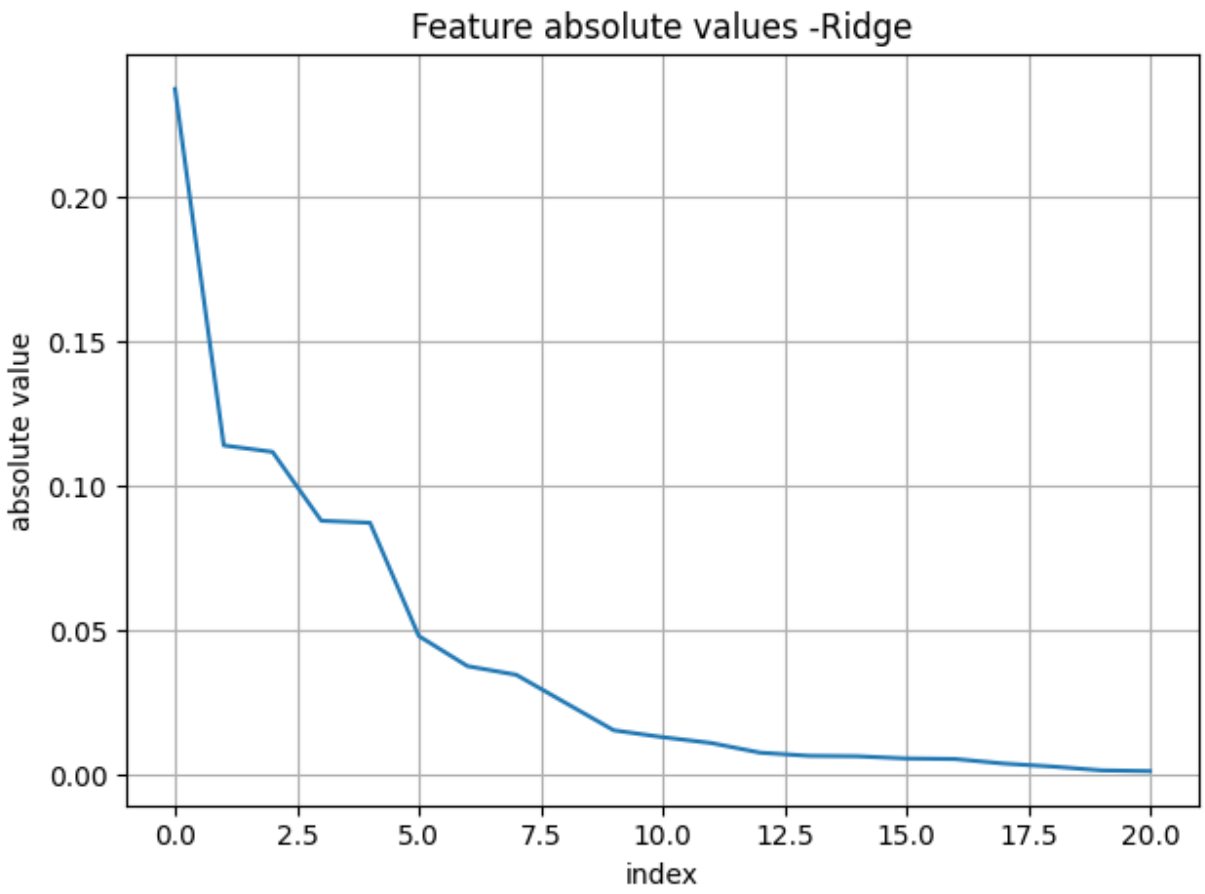
שאלה 9

התכונות שנמצאו עם הערכי coefficient הגדולים ביותר בערך מוחלט הם

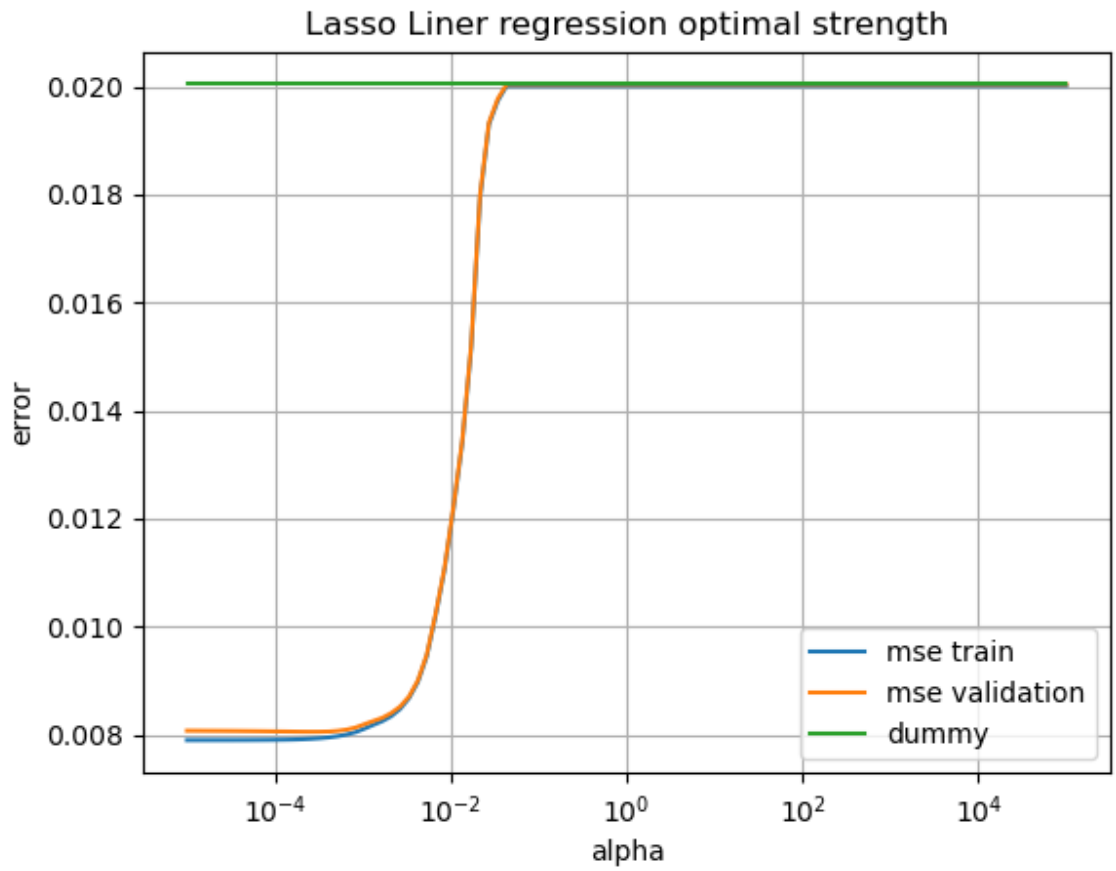
1) Shortness of breath
2) Fever
3) PCR_08
4) Blood_viruse (our added feature according to blood type)
5) Household_income

שאלה 10

הגרף שהתקבל הוא :



שאלה 11



הערך שגיאה הטוב ביותר שהתקבל הוא : 0.008049

עבור alpha : 0.000259

שאלה 12

ישנו שוני בשיפוע ומיקומו, דבר הנגרם כתוצאה מכך שהפונקציה של Ridge משתמשת ב L2 norm אשר שומר על ערכים נמוכים ברכיבים של ווקטור w עם שונות נמוכה בין הערכים.

בניגוד ל Lasso המשתמש ב L1 norm אשר מבצע feature selection על ידי הגדרת ערך 0 לחלק מן הרכיבים בוקטור w – ובסופו של דבר לא להתייחס כלל לחלק מן התכונות ולכן הוא רגיש יותר לערכי alpha (כי בlasso למעשה הרכיבים השונים מ 0 בווקטור w מקפיצים מאוד את השגיאה בשונה מ Ridge הנותן לערכים המרכבים את וקטור w ערכים נמוכים הקרובים ל 0)

נשים לב שיש שוני קטן בערכי MSE המיטייבים בין המודלים RIDGE , LASSO, הדבר היחיד שמשתנה הוא מידת ההשפעה של alpha , אשר ניתן לראות לפי הגרף של LASSO שבשאלה 11 שההשפעה של alpha מתחילה באזור ערכים של 10^{-2} לעומת ההשפעה של alpha על RIDGE לפי הגרף בשאלה 7 המשפיעה באזור alpha = 10^{2}

שאלה 13

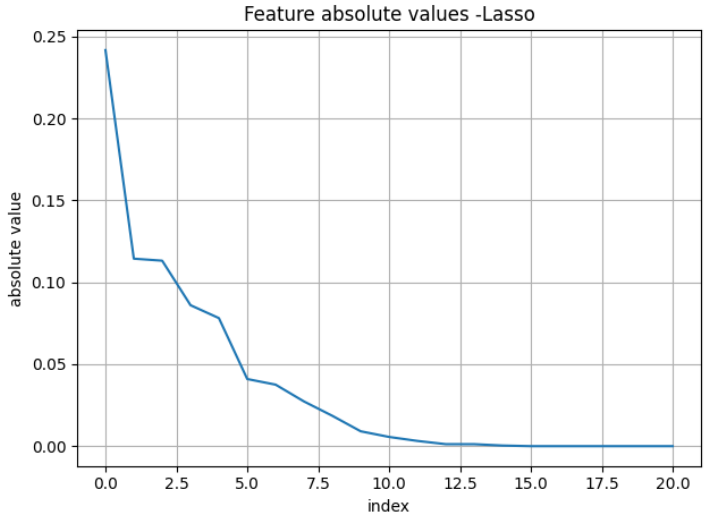
Model	Section	Train MSE	Valid MSE
		Cross Validated	
Dummy	3	0.02001108412543403	0.020040073426649308
Ridge	4	0.007899551580020458	0.008068282291301469
Lasso	5	0.007919	0.008049

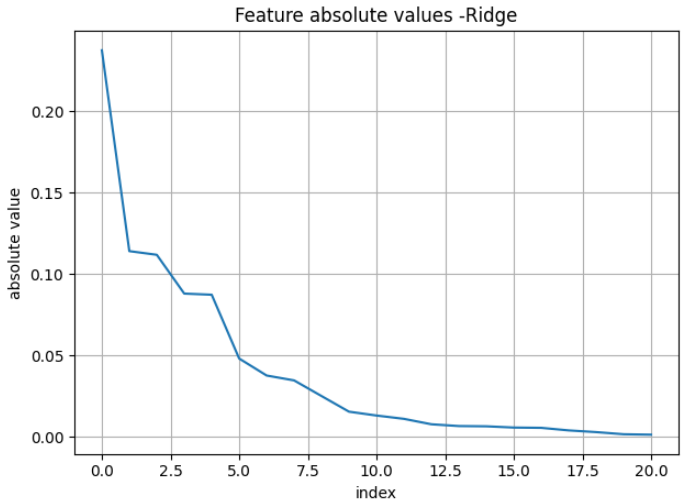
שאלה 14

חמש התכונות המשפיעות ביותר הן :

1) Shortness of breah
2) PCR_08
3) Fever
4) Blood_viruse
5) Household_income

שאלה 15





נשים לב כי התנהגות של 2 הגרפים יחסית דומה, אך נשים לב להבדלים כאשר ה absolute value קטנים,

בנוגע ל Ridge regressor קיבלנו מספר תכונות שקורבות מאוד ל 0 אך לא 0.

ועבור ה- Lasso regressor קיבלנו שישנן תכונות עם ערכי absolute value השווים ל 0 בדיוק.

מפני ש Lasso ממשקל תכונות באופן נמוך יותר עקב נטייה ל- sparse solutions (פתרון המסתמך על פחות תכנות) וזאת בעקבות השימוש של Lasso בנורמת L1 אשר בעלת נטייה לאפס תכונות עם השפעה נמוכה ביחס לתכונות אחרות.

שאלה 16

יתרחש שינוי בשגיאת האימון ושגיאת הוולידאציה עקב הוספת תכונות נוספות, אשר מאפשרות לנו למידה עם מספר רב יותר של תכונות.

במקרה שלנו נשים לב כי התכונות המשפיעות ביותר שקיבלנו הינן:

1)PCR_08 blood_viruse

2)cough shortness_of_breath

3)fever blood_viruse

4)shortness_of_breath^2

5)shortness_of_breath

נשים לב שהתכונות המשפיעות ביותר הינן התכונות החדשות שנוספו לנו על ידי זה שהרחבנו את את המודל הלינארי למודל פולינומי ובכך הורדנו את שגיאת האימון ושגיאת הולידצטיה, נשים לב כי הוספת תכנות גם הייתה יכולה במידה מסוימת להוביל ל over fitting.

שאלה 17

נראה זאת מתמטית,

נגדיר את המשתנה x_{q1} כמשתנה בינארי כך ש $x_{q1} = 1$ אם אנחנו עונים על התנאי של שאלה אחת- כלומר מכיל A ו $x_{q1} = 0$ במידה ולא,

ולכן נוכל לומר כי:

$$h_{multi}(\underline{x}) = \left(\underline{w_1^T}\,\underline{x} + b_1\right) * x_{q1} + \left(\underline{w_2^T}\underline{x} + b_2\right) * \left(1 - x_{q1}\right)$$

$$h_{multi}(\underline{x}) = \underline{w_1^T}\,\underline{x} * x_{q1} + b_1 * x_{q1} + \underline{w_2^T}\underline{x} - \underline{w_2^T}\underline{x} * x_{q1} + b_2 - x_{q_1} * b_2$$

$$h_{multi}(\underline{x}) = \underline{w_1^T}\,\underline{x} * x_{q_1} - \underline{w_2^T}\underline{x} * x_{q1} + (b_1 - b_2) * x_{q_1} + \underline{w_2^T}\underline{x} + b_2$$

$$h_{multi}(\underline{x}) = \left(\underline{w_1^T} - \underline{w_2^T}\right) * \underline{x} * x_{q1} + (b_1 - b_2) * x_{q_1} + \underline{w_2^T}\underline{x} + b_2$$

כעת נשים לב לדבר הבא,

$\left(\underline{w_1^T} - \underline{w_2^T}\right) * \underline{x} * x_{q1}$ הינו פולינום ממעלה שנייה וזאת משום שכופלים כל תכונה בתכונה "מכיל את A" ולכן זה פולינום ממעלה שנייה.

$(b_1 - b_2) * x_{q_1}$ הינו פולינום ממעלה ראשונה

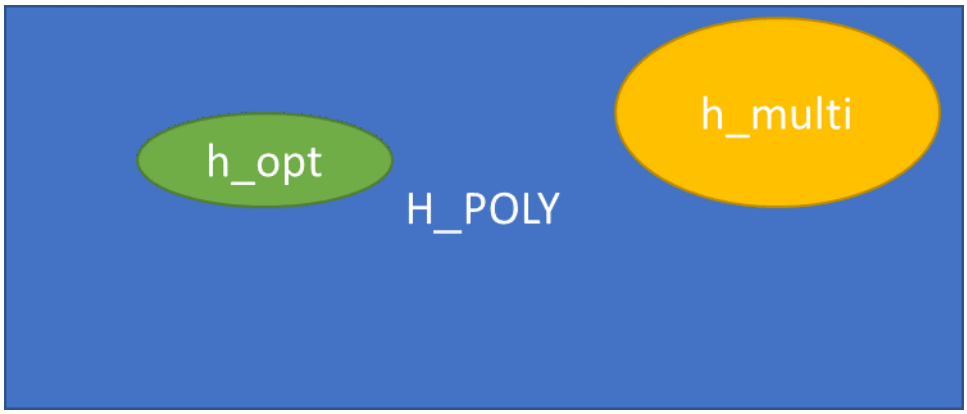
$\underline{w_2^T}\underline{x}$ הינו פולינום ממעלה ראשונה

b_2 הינו קבוע

ולכן סהכ מקבלים כי $h_{multi}(\underline{x}) \in H_{poly}$ ואת מה שנדרש.

לגבי המודל שיביא את התוצאות הטובות ביותר,

נשים לב כי מתקיים הדבר הבא:



נגדיר *h_opt* להיות המודל הפולינומיאלי שמביא את *validation_erorr* ו *training_error* הנמוכים ביותר,

נשים לב כי $h_{opt} \in H\_poly$

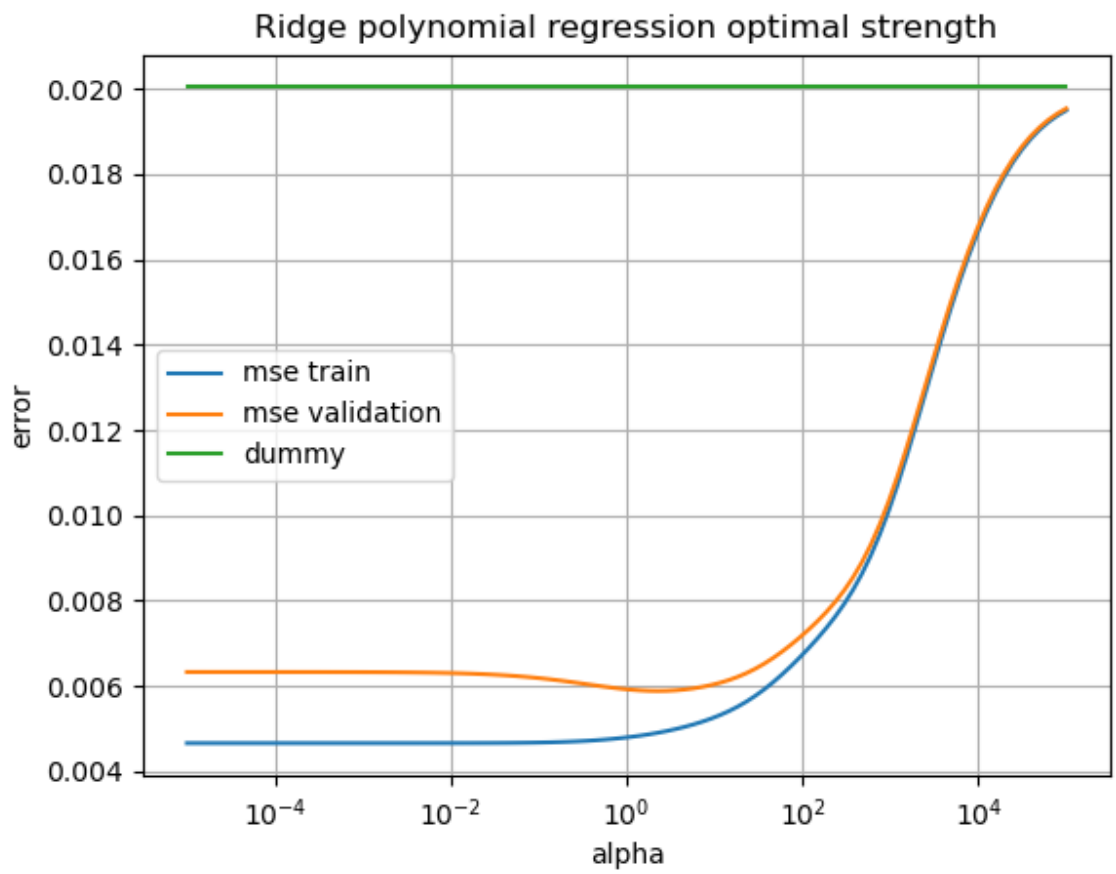
כמו כן נשים לב כי כל פתרון שנמצא ב h_{multi} גם מוכל ב H_{poly}, ולכן H_{poly} יכול להניב את אותם תוצאות של h_{multi} ואף יותר טובות על ה *train* ועל ה *validation* שזה אותו *h_opt*

ובנוסף נשים לב שמבדיקה שעשינו קיבלנו את התוצאה הבאה –

```
model_a (contain A) mse  0.013409289048133 81
model_b (Not contain A) mse  0.010861466037288724
polynomial ridge mse  0.006680487030642845
```

וזה מאשש את המסקנה שלנו.

שאלה 18



ערך השגיאה הטוב ביותר שהתקבל הוא : 0.0058869

עבור alpha : 2.2519719633

שאלה 19

Model	Section	Train MSE	Valid MSE
		Cross Validated	
Dummy	3	0.02001108412543403	0.020040073426649308
Ridge	4	0.007899551580020458	0.008068282291301469
Lasso	5	0.007919	0.008049
Ridge polynomial	6	0.004905504397315521	0.005886920556338286

שאלה 20

Model	Section	Train MSE	Valid MSE	Test MSE
		Cross Validated		Retrained
Dummy	3	0.02001108412543403	0.020040073426649308	0.021396513611111112
Ridge	4	0.007899551580020458	0.008068282291301469	0.007651409804232747
Lasso	5	0.007919	0.008049	0.00762251432378922
Ridge Polynomial	6	0.004905504397315521	0.005886920556338286	0.0066804870306428656

נעבור כל מודל בנפרד

Dummy הינו מודל בעל הדיוק הנמוך ביותר (train,valid,test) בגלל אופן האימון שלו.

Ridge , בעל ערכי valid ו train שהם בערך אותו דבר (הפרש הקטן מ 0.0002) ולכן זה מעיד שהמודל לא מבצע התאמת יתר על קבוצת האימון ולכן שהגדלנו את קבוצת האימון ובחנו אותו על test קיבלנו שגיאה נמוכה יותר ממה שהיה ב train וב valid.

באותו אופן גם לגבי ה lasso שהוא בעל התנהגות דומה.

אך ב ridge polynomial קיבלנו ש valid נמוך מ train ביותר מ 0.0009 , דבר המעיד על זה שהמודל מבצע התאמת יתר על קבוצת האימון, ואכן ניתן לראות שכאשר מגדילים את קבוצת האימון ובוחנים על ה test , מקבלים שגיאה הרבה יותר גדולה ממה שהיה ב valid ו ב train

אך נשים לב כי למרות שהמודל ridge polynomial מבצע התאמת יתר, הוא בעל הביצועים הטובים ביותר על קבוצת המבחן מכל ה 4 מודלים, הוספת התכונות גם גרמה ליותר התאמת יתר וגם גרמה לביצועים טובים יותר.