**דו"ח**

**שאלה 1**

הגרפים המתקבלים הם :

Chart, histogram

Description automatically generated

**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 :

**שאלה 3**

להלן חישוב הנגזרת,

**שאלה 4**

Chart, line chart

Description automatically generated

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

ולכן אפשר להצדיק התנהגות זאת.

**שאלה 5**

Diagram

Description automatically generated

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 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 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 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 שבמקרה זה הדיוק עלול להיות נמוך יותר אך נבטיח התכנסות.

lr size = 0.01, Best train loss = 0.007996124967117628,

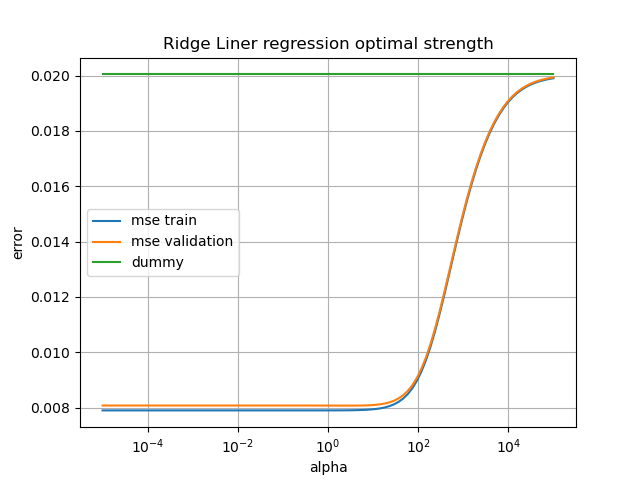
Best validation loss = 0.009689819450089773

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 :

Chart, line chart

Description automatically generated

**שאלה 11**

Chart, line chart

Description automatically generated

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 לLasso 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 LASSO , RIDGE, הדבר היחיד שמשתנה הוא מידת ההשפעה של alpha , אשר ניתן לראות לפי הגרף של LASSO שבשאלה 11 שההשפעה של alpha מתחילה באזור ערכים של לעומת ההשפעה של alpha על RIDGE לפי הגרף בשאלה 7 המשפיעה באזור alpha =

**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**

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

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 כמשתנה בינארי כך ש אם אנחנו עונים על התנאי של שאלה אחת- כלומר מכיל A ו במידה ולא,

ולכן נוכל לומר כי:

כעת נשים לב לדבר הבא,

הינו פולינום ממעלה שנייה וזאת משום שכופלים כל תכונה בתכונה "מכיל את A" ולכן זה פולינום ממעלה שנייה.

הינו פולינום ממעלה ראשונה

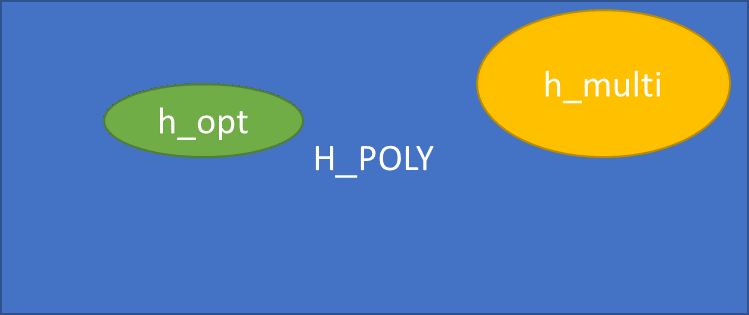
הינו פולינום ממעלה ראשונה

*הינו קבוע*

*ולכן סהכ מקבלים כי 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*

*כמו כן נשים לב כי כל פתרון שנמצא ב גם מוכל ב ,ולכן יכול להניב את אותם תוצאות של ואף יותר טובות על ה train ועל ה validation שזה אותו h\_opt*

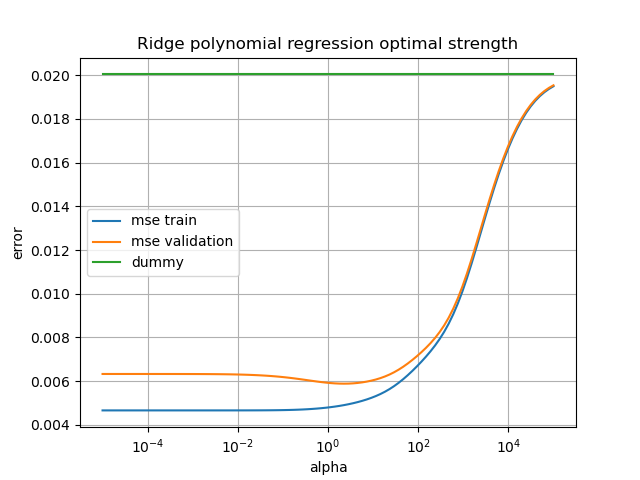
*ובנוסף נשים לב שמבדיקה שעשינו קיבלנו את התוצאה הבאה –*

Text

Description automatically generated

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 ובוחחנים 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.

**שאלה 21**

בשאלה זו החלטנו ללכת על מודל RandomForestRegressor , המכיל 139 עצים, עם עומק של 5 בכל עץ על תכונות פולונומויאליות ממעלה 2 של הדטה שלנו.

הגענו להחלטה הזו לאחר תהליך ארוך של ניסניונות עם מספר רב של מודלים .

המודלים אותם ניסינו הם הבאים:

1. linearSVC אשר הגענו איתו לשגיאה MSE גבוהה יותר ולכן פסלנו אותו

ניסינו אותו עם מספר ארגומנטים אך לא התקרבנו לרמת MSE הרצויה מבחינתנו , יצאנו מההנחה שהבעיה היא convex ושימוש במודל זה יהיה יעיל.

1. Neural network

ביצענו ניסויים על רשת נוירונים בעלות שכבה אחת מוסתרת ועד 3 שכבות מוסתרות , כאשר בכל שכבה יש בין 1000 nodes ל – node יחיד.

בניסויים אלה ראינו שרמת הדיוק אליה הגענו עם רשת הנוירונים היא MSE = 0.0069 עם פונקציית אקטיבציה relu עבור הdata הרגיל

לעומת MSE של 0.0066 עבור הdata בשניה, על התכונות הפולונומיאליות שלו ממעלה 2 . עם פונקציית אקטיבציה identity .

3)CatBoostRegressor עשינו מספר הרצות של CatBoostRegressor

4)RandomForest Regressor

עשינו מספר ניסויים על גודל יער שונה ועומקים שונים.

את התוצאה הטובה ביותר קיבלנו עבור יער עם 139 עצים עם עומק של כל עץ של 5 , אשר נתן MSE 0.0055 על הפולונומיאלי ו MSE של 0.0056 על הדטה הרגיל.