תרגיל בית רטוב 3 מבוא למערכות לומדות

מגישים:

נועה דיקמן 315478867

לוי הורביץ 313511602

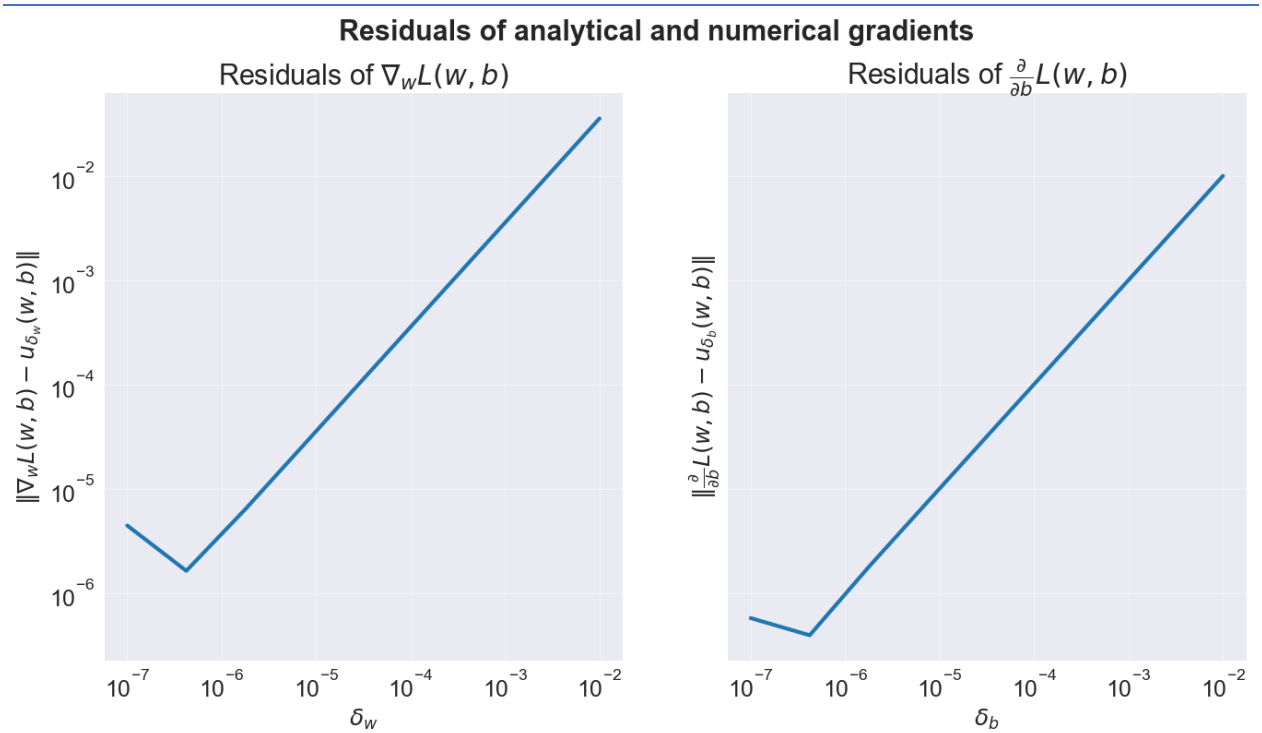
חלק 1

שאלה 1

בשאלה זו התבקשנו לפתח את הגרדיאנט של הloss לפי b. להלן הפיתוח המבוקש:

$$\frac{\partial}{\partial b}L(\underline{w},b)=\frac{\partial}{\partial b}\frac{1}{m}\left|\left|X\underline{w}+\underline{1}_m\cdot b-\underline{y}\right|\right|_2^2=\frac{2}{m}\underline{1}^T{}_m\left(X\underline{w}+\underline{1}_m\cdot b-\underline{y}\right)$$

שאלה 2



שאלה 3

מסקנות כלליות מהגרפים:

- ניתן לראות שקצב התכנסות השגיאות גדל ככל שמגדילים את הlearning rate. עבור ערכי learning rate קטנים מאוד השינוי בMSE לא ניתן להבחנה בגרפים ואילו בקצבים גבוהים (כגון 0.01) השגיאה מתכנסת כבר באיטרציות הראשונות. התנהגות זו תואמת את התיאוריה שלמדנו, לפיה לצעד מאוד קטן ייקח הרבה מאוד זמן להתכנס למינימום. חשוב לציין כי לפי התיאוריה שלמדנו הצעדים הקטנים כן צפויים להביא להתכנסות השגיאה, אך ככל הנראה 1500 אינן מספיקות לשם כך.
- מהצד השני, עבור קצב התכנסות של 0.1 נראה שבסוף הגרף יש מאין קפיצות מחזוריות בערכי השגיאה. התנהגות זו תואמת גם היא את התיאוריה שלמדנו, על פיה, צעד גדול מידי עלול לפספס את המינימום ואף לא להתכנס לעולם.

Learning rate אופטימלי וניתוח מספר צעדים:

- ניתן להסיק כי הlearning rate האופטימלי הוא 0.01 כיוון שהוא מביא להתכנסות מהירה של השגיאה, אך לא גורם לקפיצות בשגיאה לאחר ההתכנסות.
- ניתן לראות שעבור גודל הצעד האופטימלי, השגיאה מתכנסת לאחר כ-250 איטרציות. לכן, אין סיבה להגדיל את כמות האיטרציות מעבר לערך דיפוטיבי של 1500 איטרציות.

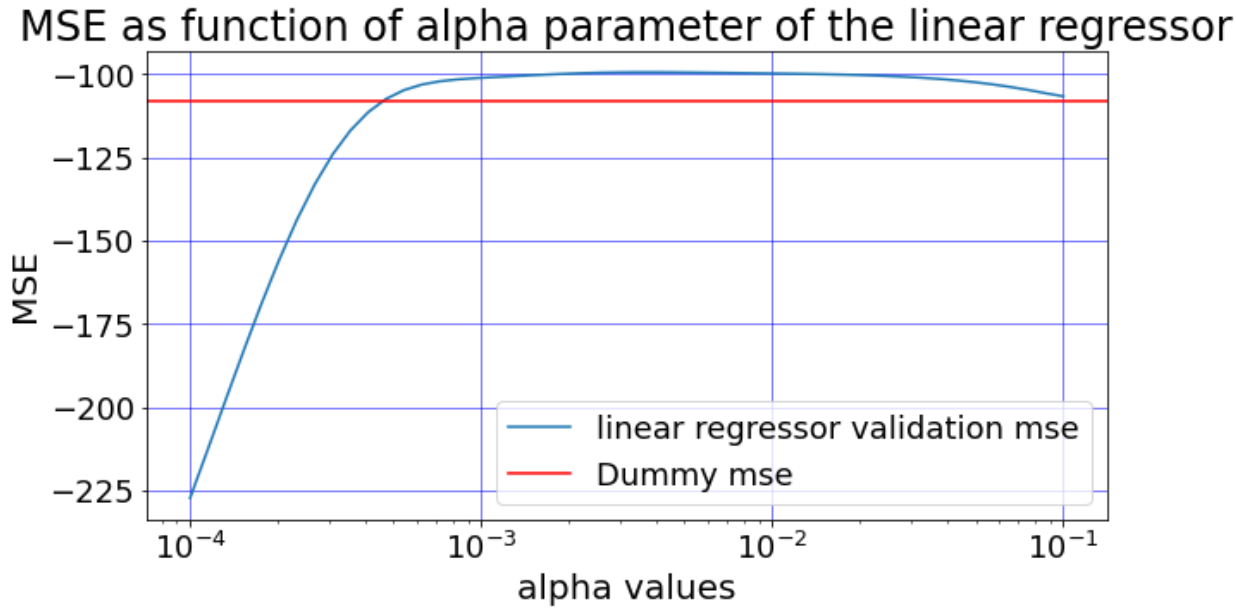
חלק 2

שאלה 4

הערכים נוספו לטבלת השגיאות.

שאלה 5

- להלן הגרף המבוקש עבור תהליך הtuning



- ערכי השגיאה נוספו לטבלת השגיאות.

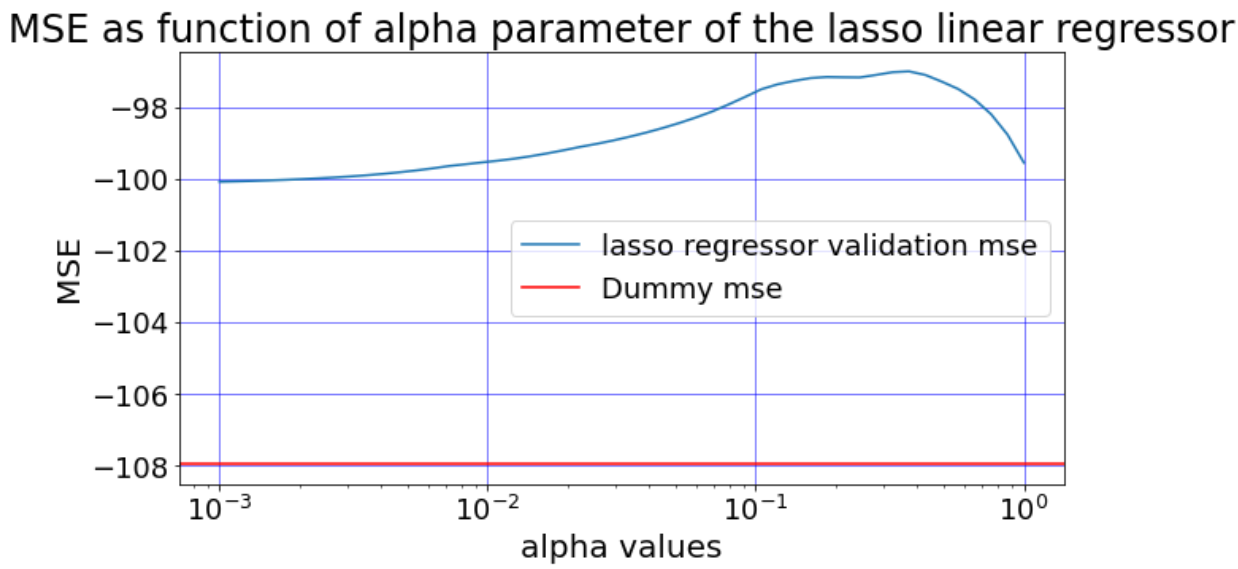
שאלה 6

אם היינו בוחרים שלא לנרמל את התכונות, ובהנחה שאין שגיאות נומריות, ביצועי המודלים הלינארים הנ"ל לא היו משתנים. להלן הנימוקים עבור כל אחד מהמודלים:

- DummyRegressor- המודל חוזה את אותו הערך עבור כל דוגמא, ערך זה תלוי אך ורק במשתנה המטרה, אותו כלל לא נרמלנו.
- LinearRegressor- הנורמליזציה כוללת רק פעולות לינאריות על x. לכן, השגיאה עצמה לא תשתנה, שכן המודל אשר מבצע גם הוא פעולות לינאריות על הדאטה יכול "לתקן" את הנורמליזציה. למרות זאת, עלול להיות שינוי בקצב הלמידה בו עלינו להשתמש.

שאלה 7

- ערך הפרמטר האופטימל שהתקבל הוא: 0.372 עם שגיאת ולידציה ממוצעת של 96.987-. להלן גרף תהליך הtuning:



שאלה 8

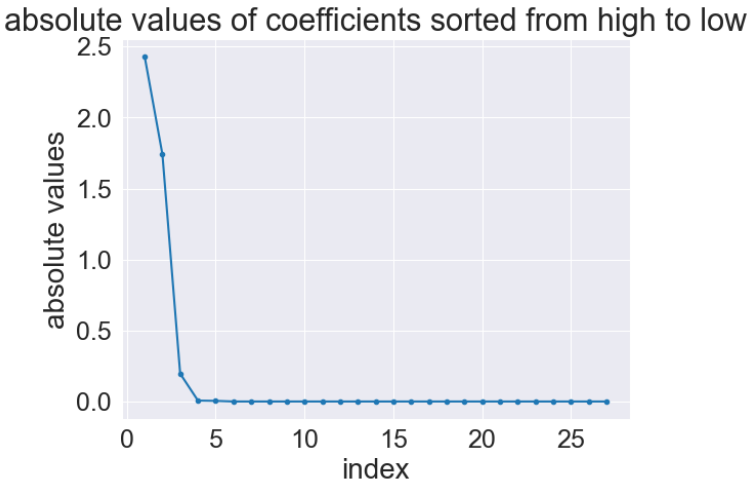
הערכים נוספו לטבלת השגיאות.

שאלה 9

Table 1- top 5 features with largest coefficients

0	age	2.429947
1	low_appetite	1.740277
2	fever	0.196121
3	cough	0.006195
4	blood_type3	0.004216

שאלה 10



שאלה 11

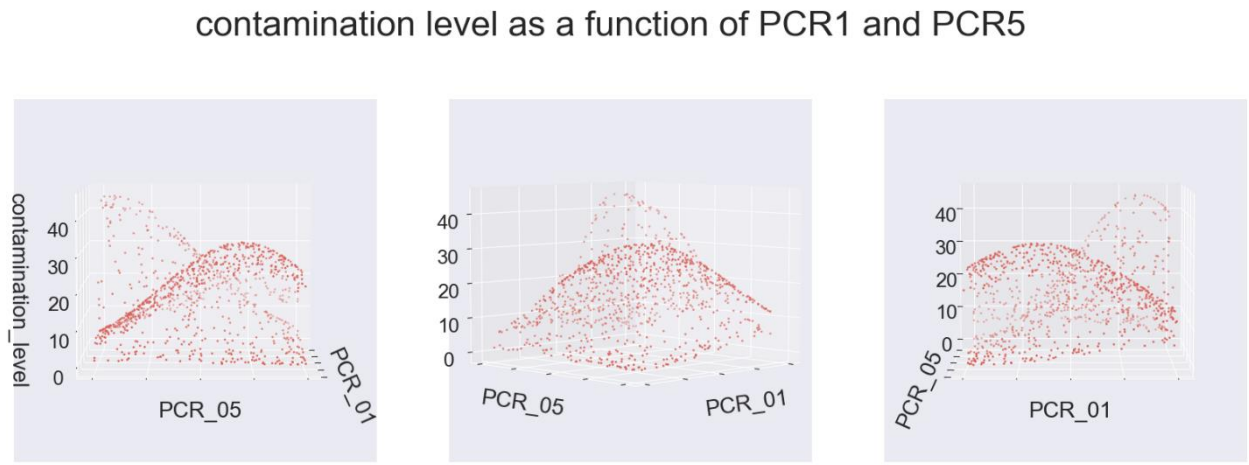
- חשוב לדעת את גודל המקדמים של התכונות כדי לקבל מידע על חשיבות הפיצ'רים למודל. כלומר, פיצ'רים אשר המקדמים שלהם גדולים יהיו משמעותיות לחיזוי המטרה, לעומת זאת, פיצ'רים שהמקדם שלהם הוא 0 לא ישפיעו כלל על החיזוי וניתן להוריד אותם מבלי להשפיע על ביצועי המודל.

שאלה 12

- אם היינו בוחרים שלא לנרמל את ההפיצ'רים ביצועי המודל היו מושפעים מכך.
- במודל lasso אנו נותנים חשיבות לגודל הוקטור w . לכן, אם הפיצ'רים שלנו היו בעלי סדרי גודל שונים, המקדמים של פיצ'רים קטנים היו צריכים להיות מאוד גדולים ביחס למקדמים של פיצ'רים גדולים. לכן, ייתכן שהמודל היה בוחר שלא להשתמש בהם לא בגלל חשיבותם, אלא בגלל ה"עונש" הגדול השגיאה כתוצאה ממקדם גדול מאוד. מהצד השני, ייתכן שפיצ'רים גדולים ייבחרו שלא כתוצאה מהתאמתם למודל אלא מעצם האפרות להשתמש בהם מבלי להגדיל את w משמעותית.

חלק 4

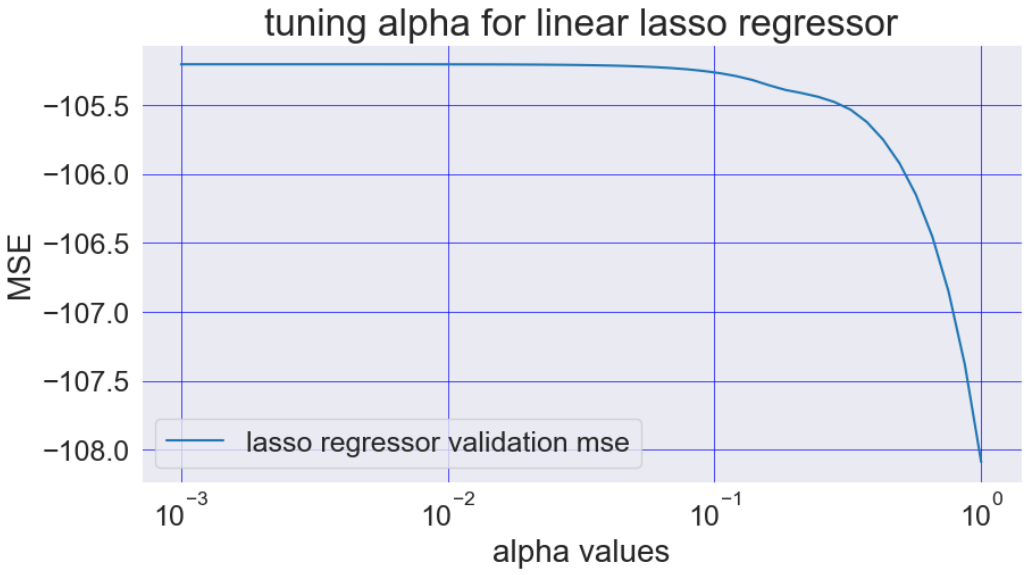
שאלה 13



- בגרפים המצורפים ניתן לראות שיש קשר פולינומיאלי (פרבולי) בין הפיצ'רים לבין contamination_level.
- כדאי לבחור מודל שמחפש קירוב פולינומיאלי, בפרט, ניתן לנסות לקרב לפולינום מסדר 2.

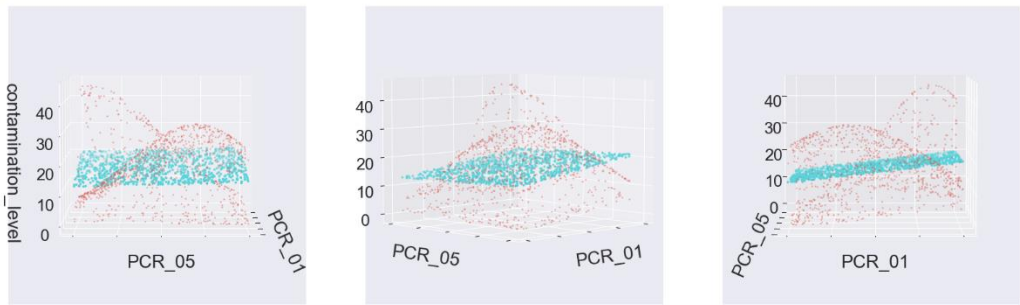
שאלה 14

- האלפא האופטימלי שהתקבל הוא 0.00152 עם שגיאת ולידציה של 105.206- .



שאלה 15

contamination level as a function of PCR1 and PCR5 real data compares to baseline predictions

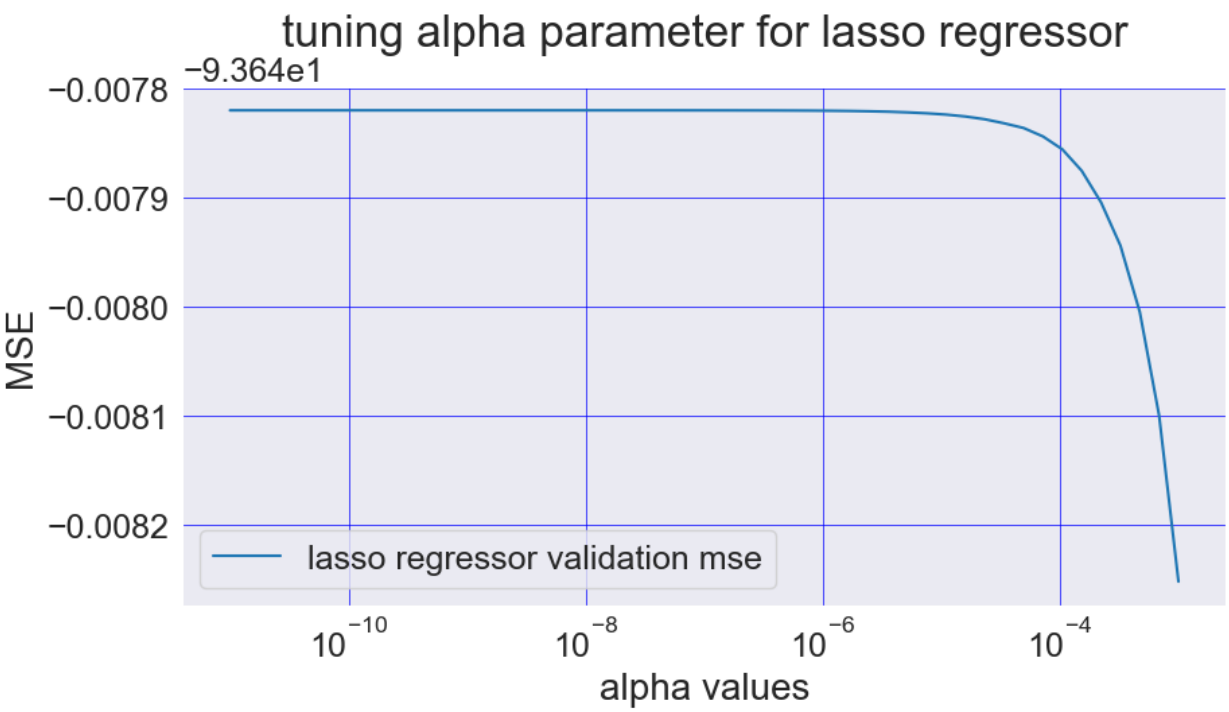


שאלה 16

- לאחר מיפוי פולינומיאלי חשוב לבצע נורמליזציה נוספת לפני שימוש בlasso . הסיבה לכך היא שלאחר המיפוי, נוצרים פיצ'רים חדשים אשר שונים בסדרי הגודל מהפיצ'רים המקוריים ומפיצ'רים אחרים שנוספו, כתוצאה מהעלאה בחזקה והכפלה בפיצ'רים נוספים.
- בשאלה 12 הראינו שהבדלים בסדרי הגודל בין הפיצ'רים משפיעים על ביצועי lasso ולכן יש לנרמל את הפיצ,רים לאחר המיפוי.

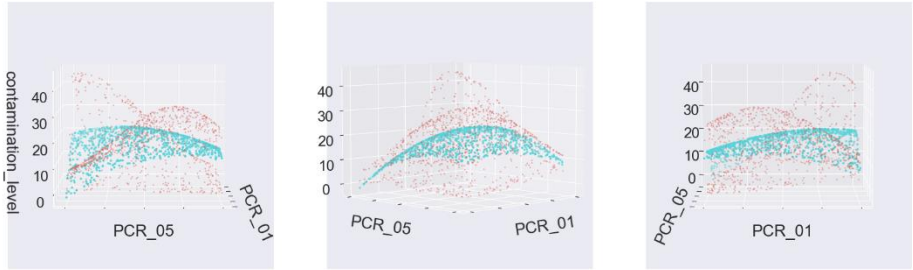
שאלה 17

- האלפא האופטימל שהתקבל הוא 10^{-11} עם שגיאת ולידציה של -93.647 .



שאלה 18

contamination level as a function of PCR1 and PCR5 compare data and prediction of polynomial fitted lasso



שאלה 19

- ניתן לראות שלמודל היתה יכולת גבוהה יותר להתאים לדאטה כאשר השתמשנו במיפוי פולינומי. קיימים מספר אינדיקציות לכך:
  - אם נסתכל על הגרף משאלה 15 ביחס לגרף משאלה 18 נשים לב שבגרף השני יש התאמה גבוהה יותר בין הפרדיקציות לבין הדאטה המקורי.
  - מבחינת שגיאת הולידציה- עבור המסווג הראשון קיבלנו שגיאה של 105.206-, לעומת שיגאה של 93.647- עבור המסווג השני.
- לסיכום, באמצעות מיפוי הפיצ'רים הצלחנו ליצור מסווג לא לינארי מרגרסור לינארי, מסווג זה התאים יותר לדאטה שלנו כפי שצפינו.

חלק 5

שאלה 20

- על הפיצ'רים הבאים ביצענו מיפוי RBF: "PCR_06","weight","sugar_levels","PCR_10"
- על הפיצ'רים הבאים היצענו מיפוי פולינומיאלי: "PCR_05","PCR_01"
- את שאר הפיצ'רים השארנו כפי שהם.
- **בחירת הפיצ'רים**: לצורך בחירת הפיצ'רים למיפוי, ראשית יצרנו גרפים של "contamination_level" כתלות בפיצ'ר כדי לראות באופן כללי איך נראית התלות. עבור פיצ'רים שנראו כמתאימים לאחד המיפויים ניסינו לייצר את הרגרסור עם המיפוי ובלעדיו וראינו כיצד המיפוי משפיע על השגיאה, אם המיפוי שיפר את השגיאה בחרנו פיצ'ר זה לעבור מיפוי, אחרת, השארנו אותו כפי שהוא.

שאלה 21

התבקשנו להסביר איך ומדוע שגיאות המודל יושפעו משימוש במיפוי RBF.

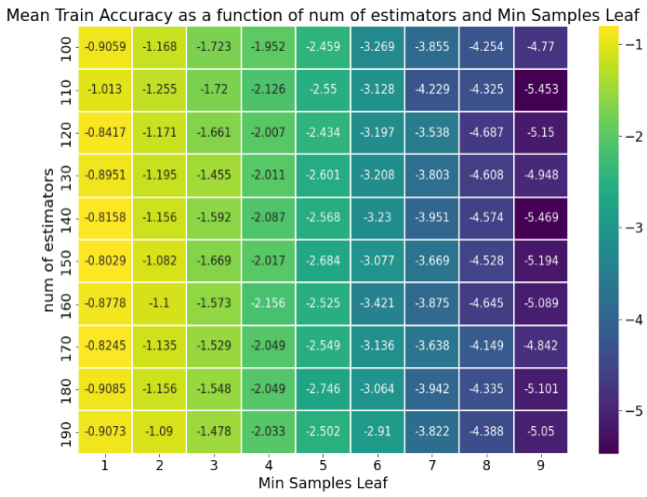
- באופן כללי, שימוש בRBF מעלה את המימד של X ולכן יאפשר לRFR יותר גמישות בהפרדת הדאטה.
- **השפעה על שגיאת האימון:** שימוש בRBF ישפר את שגיאת האימון. זאת מכיוון שהRFR יוכל ללמוד גם כללי החלטה שאינם לינארים ובכך להתאים את עצמו לסוגי דאטה שונים.
- **השפעה על שגיאת הולידציה:** שימוש ב-RBF יכול לשפר או לגרוע משגיאת הולידציה. מצד אחד ייתכן שללא המיפוי, הRFR יהיה underfitting והשגיאות הן של האימון והן של הולידציה יהיו גבוהות. מצד שני ייתכן שהמיפוי יגרום לoverfitting כך ששגיאת האימון תרד אבל שגיאת הולידציה תעלה.

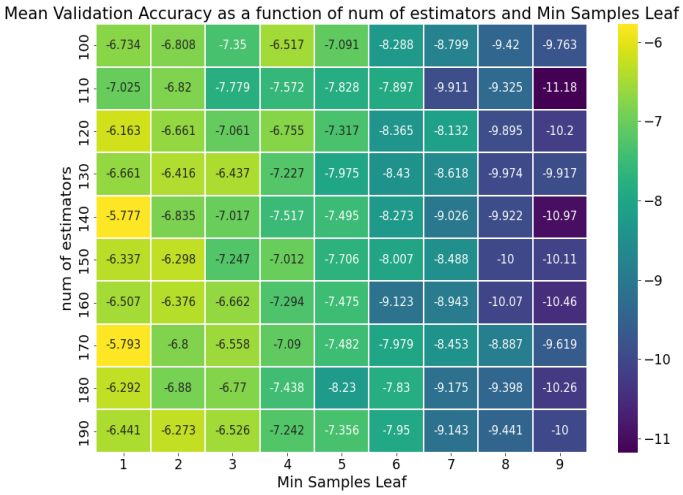
שאלה 22

- RFR הוא מודל לא לינארי בעוד המודלים הקודמים שהשתמשנו בהם היו מודלים לינארים. כלומר, במודל הקודם ביצענו מיפוי פולינומיאלי של הקלט וסיפקנו למודל את הקלט הממופה כך שנוצר מסווג לא לינארי, אך המודל עצמו כן לינארי ומ. מצד שני, RFR הוא מודל לא לינארי מבוסס עצי החלטה.

שאלה 23

- הפרמטרים האופטימלים שהתקבלו: {'RFR__min_samples_leaf': 1, 'RFR__n_estimators': 140}
- שגיאת אימון: -2.754
- שגיאת ולידציה: -7.961





שאלה 24

הערכים נוספו לטבלת השגיאות.

חלק 6

טבלת השגיאות הכוללת

Model	Section	Train MSE	Valid MSE	Test MSE
		Cross validated		Retrained
Dummy	2	-108.866	-108.944	103.298
Linear	2	-97.915	-103.379	95.654
Lasso Linear	3	-100.349	-101.852	92.891
Random Forest	5	-0.903	-6.533	1.851

שאלה 25

- ניתן לראות שהמודל בעל הביצועים הטובים ביותר (משמעותית) הוא Random Forest.
- נשים לב שכל המודלים היו בעלי ביצועים טובים יותר מה-Dummy ולכן ניתן להסיק שאכן היתה למידה כלשהי בכולם, אם כי במודלים הלינארים ההבדלים אינם משמעותיים.
- בשני המודלים הלינארים, הביצועים ביחס לDummy טובים אך לא משמעותית,בנוסף, השגיאה עבור דוגמאות המבחן קטנה מהשגיאה על דוגמאות האימון, כלומר קיימת תופעה של underfitting. מכך ניתן להסיק שמחלקת היפוטזות לינארית אינה רחבה מספיק ללמידת הדאטה. ייתכן שאם היינו מבצעים mapping על הפיצ'רים היינו מגיעים לתוצאות טובות יותר גם עבור המודלים הלינארים.
- מבחינת ההבדלים בין lasso לבין הlinear הרגיל, ניתן לראות ששגיאת האימון של lasso גבוהה משל המודל הלינארי הרגיל, כלומר תופעת הunderfitting גבוהה יותר במודל הlasso. מסקנה זו תואמת את התיאוריה, שכן, בlasso אנחנו מקטינים את סיבוכיות המודל.
- במודל Random Forest אנו רואים תופעת overfitting, ניתן להסיק זאת מכך ששגיאת האימון קטנה פי 2 משגיאת המבחן ופי 6 משגיאת הולידציה. עם זאת, שגיאת המבחן במודל זה טובה בשני סדרי גודל מהשגיאה בשאר המודלים ולכן מודל זה עדיף על אף הoverfitting.