מבוא למערכות לומדות (236756) | תרגיל בית 3 גדול

ליאל פרבר | 214413437

ראובן טימסיט | 330083858

18 באוגוסט 2024

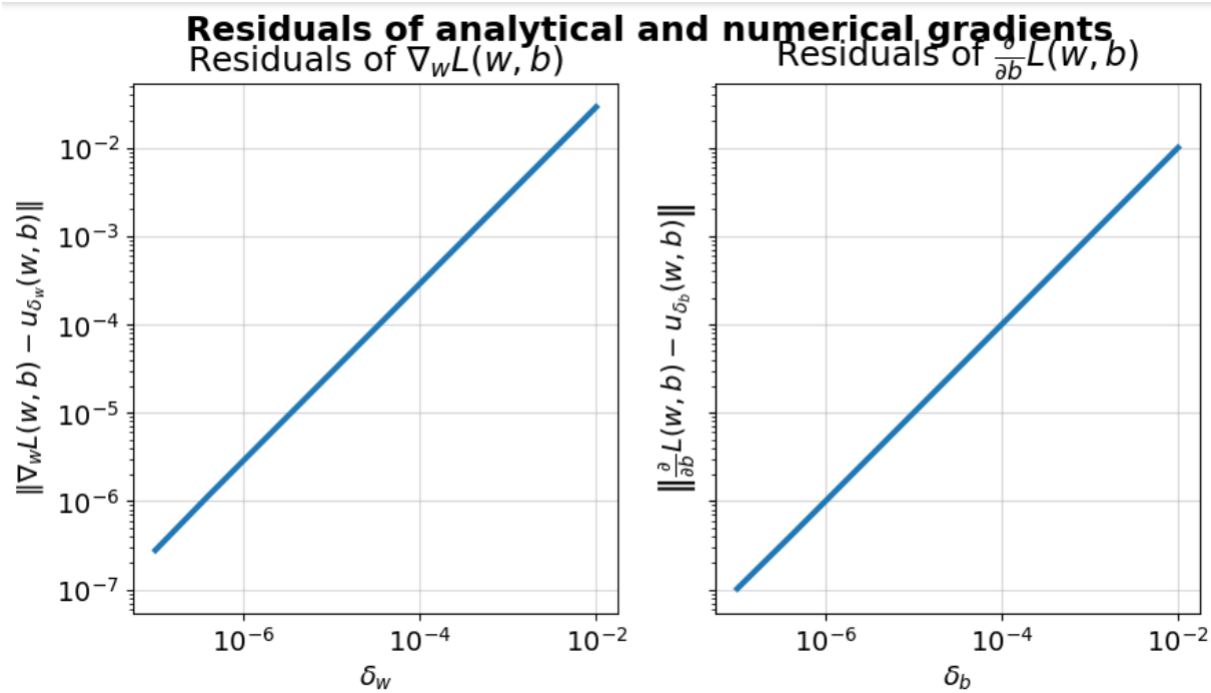
שאלה 1

מתקיים:

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

שאלה 2

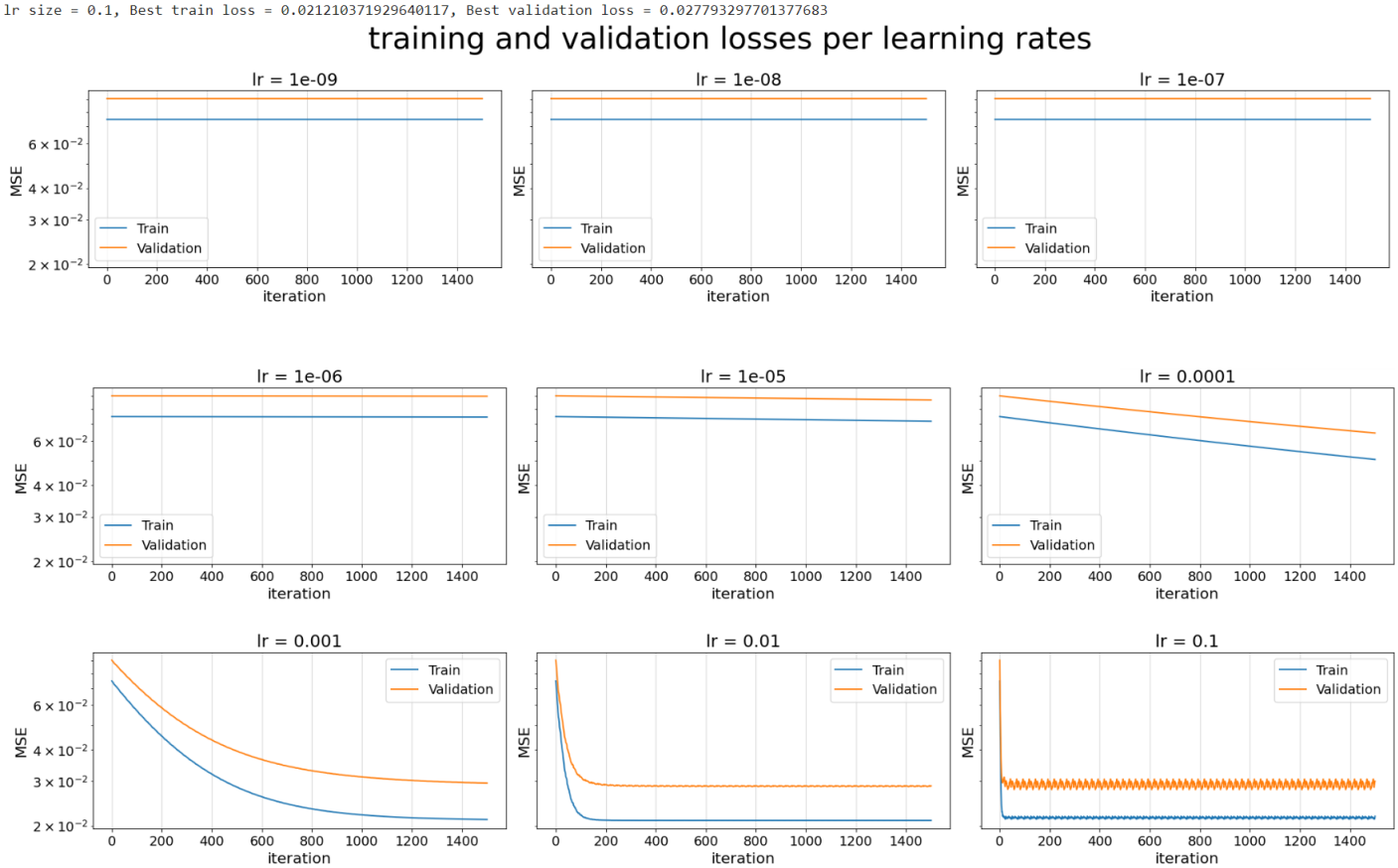
הגרפים שהתקבלו:



שאלה 3

הגרפים שהתקבלו עבור קצבי למידה שונים :

```
lr size = 1e-09, Best train loss = 0.07481921098536524, Best validation loss = 0.09036979523092965
lr size = 1e-08, Best train loss = 0.07481630622574002, Best validation loss = 0.09036671829250281
lr size = 1e-07, Best train loss = 0.07478726734327115, Best validation loss = 0.09033595778120664
lr size = 1e-06, Best train loss = 0.07449774812186245, Best validation loss = 0.09002923819593164
lr size = 1e-05, Best train loss = 0.07168776832574221, Best validation loss = 0.08704885228634808
lr size = 0.0001, Best train loss = 0.05060198230483696, Best validation loss = 0.0644202970611741
lr size = 0.001, Best train loss = 0.02115260049595843, Best validation loss = 0.02947791526898647
lr size = 0.01, Best train loss = 0.02097961181489805, Best validation loss = 0.028549757366151632
lr size = 0.1, Best train loss = 0.021210371929640117, Best validation loss = 0.027793297701377683
```



- נשים לב שעבור קצבי הלמידה $\{e^{-9}, e^{-8}, e^{7}, e^{-6}, e^{-5}\}$ אין התכנסות של ה- MSE עבור $training\ set$ וה- $validation\ test$. כלומר, אלו קצבי למידה איטיים מידי שלא מאפשרים התכנסות של השגיאה הריבועית הממוצעת במסגרת 1500 איטרציות של SGD.

- נשים לב שעבור קצב הלמידה 0.0001 יש התכנסות של ה- MSE עבור $training\ set$ וה- $validation\ test$. אבל קצב ההתכנסות עדיין איטי ולכן תוך 1500 איטרציות של SGD אין התכנסות מספיק חזקה (לערך מספיק נמוך).

- נשים לב שעבור קצב הלמידה 0.001 יש התכנסות, "בינונית בקצב שלה", של ה- MSE עבור $training\ set$ וה- $validation\ test$ במסגרת 1500 איטרציות של SGD. זאת משום שקצב ההתכנסות מהיר ביחס לשאר ה- lr הקטנים יותר אבל איטי ביחס לשאר ה- lr הגדולים יותר. כמו כן, הערכים שאליהם השגיאות מתכנסות זהים עבור $lr \in \{0.001, 0.01, 0.1\}$.

- נשים לב שעבור קצב הלמידה 0.01 יש התכנסות מהירה של ה- MSE עבור $training\ set$ וה- $validation\ test$ במסגרת 1500 איטרציות של SGD. זאת משום שקצב ההתכנסות מהיר ביחס לשאר ה- lr הקטנים יותר אבל איטי ביחס לשאר ה- lr הגדולים יותר. כמו כן, הערכים שאליהם השגיאות מתכנסות זהים עבור $lr \in \{0.001, 0.01\}$.

- נשים לב שעבור קצב הלמידה 0.1 יש התכנסות מהירה של ה- MSE עבור $training\ set$ וה- $validation\ test$ במסגרת 1500 איטרציות של SGD. זאת משום שקצב ההתכנסות הוא המהיר ביותר ביחס לשאר ה- lr (הקטנים יותר). כמו כן, הערכים שאליהם השגיאות מתכנסות זהים עבור $lr \in \{0.001, 0.01\}$. נשים לב שיש קפיצות קטנות מאוד בערכי השגיאה של ה- $training\ set$ וה- $validation\ set$ וזאת עקב קצב למידה מהיר מידי שמונע מהשגיאה להתכנס למינימום. כלומר, ה- MSE של 2 קבוצות אלו מתכנסות לסביבה קטנה של ערכים בניגוד ל- $lr \in \{0.001, 0.01\}$. אבל סביבה זו מוגדרת סביב הערכים המינימליים שאליהם מתכנסים ה- MSE של 2 הקבוצות הנ"ל עבור $lr \in \{0.001, 0.01\}$. לכן, אין הבדל משמעותי בין ערכי ההתכנסות של MSE עבור $lr \in \{0.001, 0.01, 0.1\}$ אלא רק בקצבי ההתכנסות.

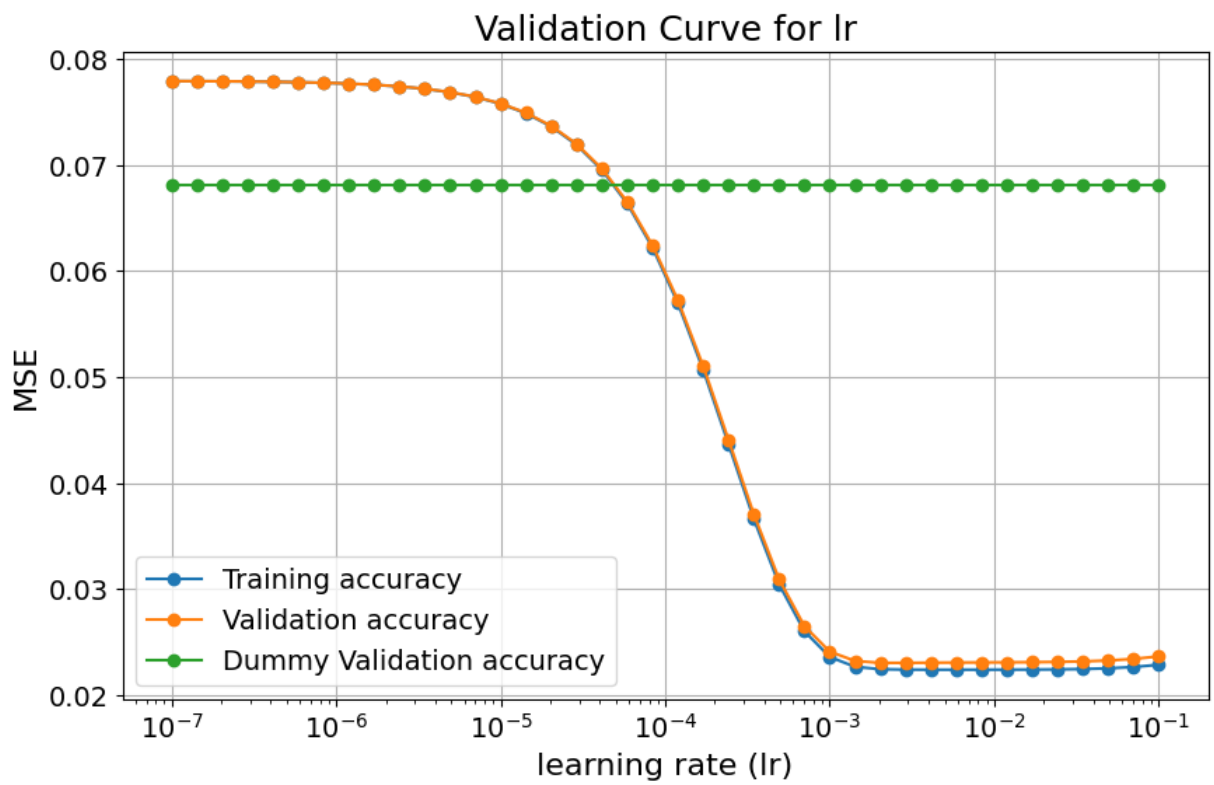
כעת, $lr = 0.1$ מוביל לשגיאת ולידציה מינימלית 0.027 ו- $lr = 0.01$ מוביל לשגיאת ולידציה מינימלית 0.028. אבל $lr = 0.01$ מתייצב <u>על</u> השגיאה 0.028 ו- $lr = 0.1$ מתייצב על <u>סביבה</u> של השגיאה 0.027. מסיבה זאת ומשום שנרצה התכנסות למינימום יותר מאשר קצב התכנסות מהיר ביותר נעדיף את $lr = 0.01$ על $lr = 0.1$. נעיר כי קצב ההתכנסות של $lr = 0.01$ מהיר מספיק (כ- 100 איטרציות). לכן, קצב הלמידה האופטימלי הוא $lr = 0.01$. נשים לב שהגדלת מספר הצעדים שנעשה במסגרת ה- SGD לא תועיל בדבר עם קצב למידה זה. זאת משום שעבורו יש התכנסות מלאה (התייצבות על ערך מינימלי) של ערכי השגיאה של ה- $training\ set$ וה- $validation\ set$ כבר לאחר $100 < 1500$ איטרציות.

שאלה 4

Model	Section	Train_MSE	Valid_MSE
		Cross_validated	
Dummy	2	0.067	0.068

שאלה 5

לאחר $lr\ tuning$ קיבלנו את הגרף הבא:




```
Best lr: 0.0028942661247167516
Average training loss for best lr: 0.0224
Average validation loss for best lr: 0.0230
```

כלומר ה- lr הטוב ביותר הוא $lr = 0.002$ ושגיאת הולידציה הממוצעת שלו היא 0.023. לכן:

Model	Section	Train_MSE	Valid_MSE
		Cross_validated	
Dummy	2	0.067	0.068
Linear	2	0.022	0.023

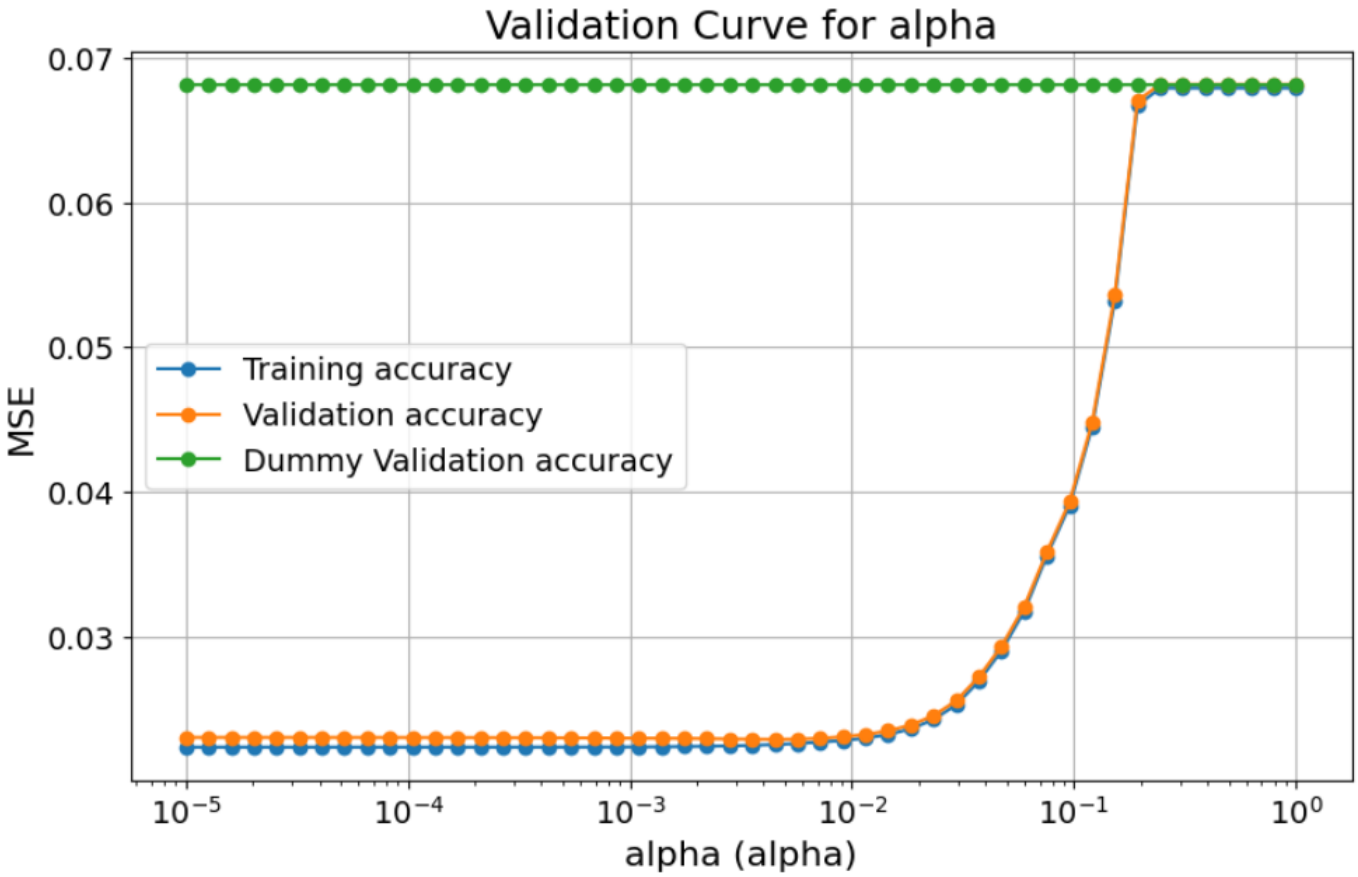
שאלה 6

נשים לב שחישבנו את ביצועי 2 המודלים בטבלה בעזרת מחלקה $LinearRegressor$ שמימשנו ובעזרת מחלקה $DummyRegressor$ שיבאנו למחברת.

- $LinearRegressor$ שמימשנו משתמשת ב- SGD כדי למצוא מסווג עם שגיאה מינימלית על קבוצת האימון שקיבלה. SGD מושפע מנרמול הפיצ'רים של הדוגמאות בקבוצת האימון. כאשר אין נרמול האלגוריתם מתכנס (למינימום לוקאלי) בקצב איטי יותר וכאשר יש נרמול הוא מתכנס בקצב מהיר יותר. לכן, אילו לא היינו מנרמלים את הפיצ'רים ($PCR\_01 - PCR\_10$) המודל $LinearRegressor$ היה משיג מסווג גרוע יותר (במסגרת 1000 האיטרציות ב- SGD שהוא ביצע). כלומר היינו מקבלים $Train\ MSE,\ Valid\ MSE$ גבוהים יותר. בנוסף, אי נרמול הפיצ'רים גורם לכך שמקדמים מסויימים של w יהיו קטנים במיוחד כדי "לאזן" את הפיצ'רים הגדולים. כלומר, המודל ינסה בכוח להתאים את עצמו על בסיס פיצ'רים גדולים בלבד ולכן ביצועיו בתהליך האימון יפחתו.

- אוביקט ה- $Dummy$ שיצרנו משתמש באסטרטגיית "$mean$" ומהווה מסווג שמסווג את כל דוגמאות האימון לפי התווית הממוצעת שלהן. כלומר, הוא מהווה מסווג (טיפש) שלא מושפע מערכי הפיצ'רים של הדוגמאות כלל. לכן, אילו לא היינו מנרמלים את הפיצ'רים ($PCR\_01 - PCR\_10$) המודל $Dummy$ (כמסווג) לא היה מושפע והיינו מקבלים אותם $Train\ MSE,\ Valid\ MSE$.

שאלה 7




```
Best alpha: 0.004498432668969444
Average training loss for best alpha: 0.0226
Average validation loss for best alpha: 0.0229
```

כלומר ה- α הטוב ביותר הוא $\alpha = 0.004$ ושגיאת הולידציה הממוצעת שלו היא 0.022.

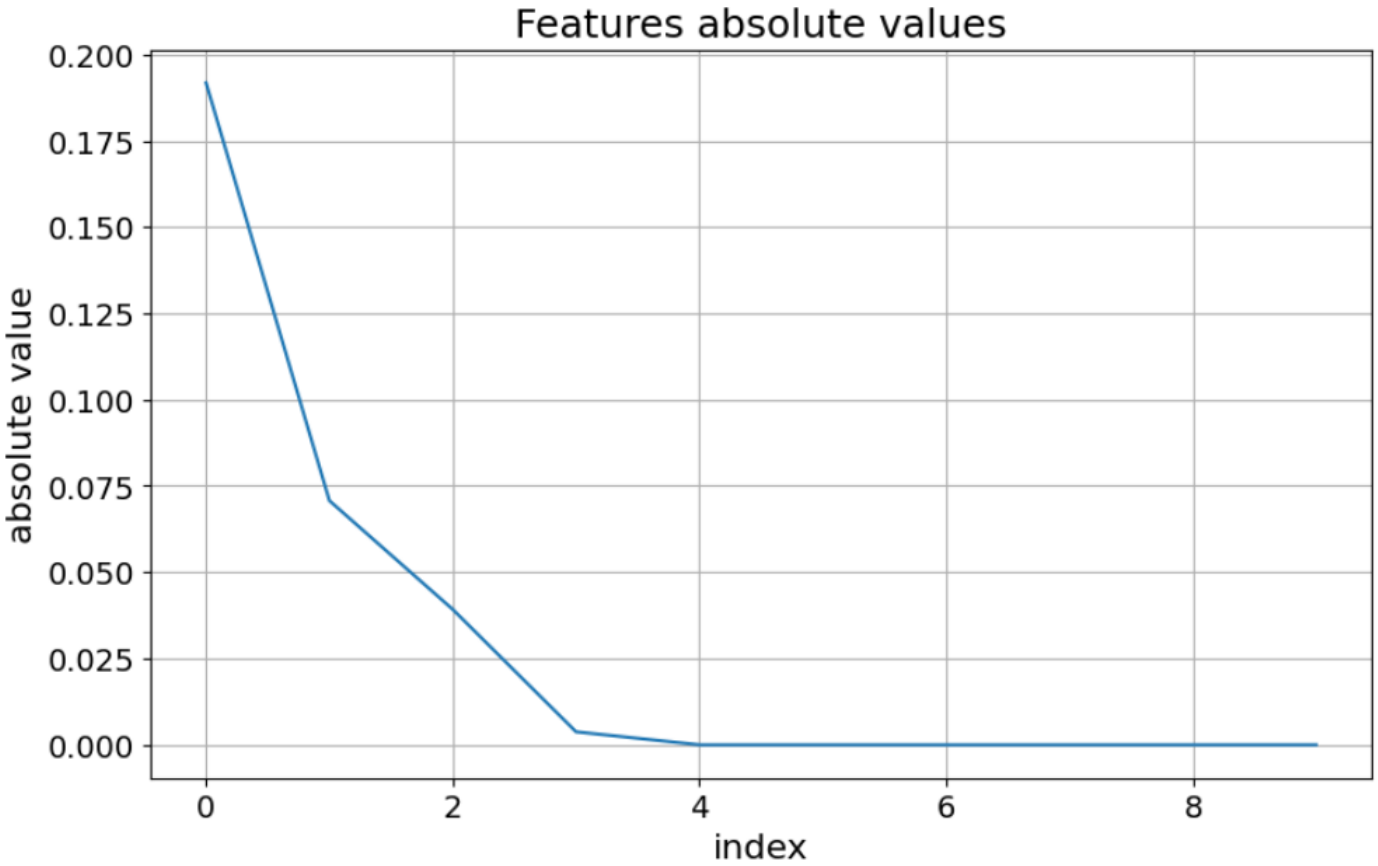
שאלה 8

Model	Section	Train_MSE	Valid_MSE
		Cross_validated	
Dummy	2	0.067	0.068
Linear	2	0.022	0.023
Lasso_Linear	3	0.022	0.022

שאלה 9

הפיצ'רים הם PCR_01, PCR_08, PCR_06, PCR_02, PCR_04 שמקדמיהם בערך מוחלט הם $0.191, 0.07, 0.039, 0.003, 0$ בהתאמה.

שאלה 10


Features absolute values
absolute value
index
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
2
4
6
8

שאלה 11

כאשר אנו מנסים למצוא ריגרסור $Lasso_{w,\alpha}$ אנו מחפשים אחד שמשיג $MSE = \frac{1}{2m}\sum_{i=1}^{m}\left(w^T x_i - y_i\right)^2 + \alpha \left\|w\right\|_1$ מינימלי עבור דוגמאות $(x_i, y_i) \in S$ ופרמטרים $w \in \mathbb{R}^d$ ו- $\alpha \in \mathbb{R}$. המקדמים שבהם אנו עוסקים בשאלה הם רכיבי הוקטור w.

- כאשר למקדם c_j כזה יש סדר גודל גבוה הוא מגדיל יותר את הערך $c_j\left(x_i\right)_j$ מה שמצביע על כך שהמודל מעוניין להתחשב יותר בפיצ'ר ה- j של דוגמאות. כמו כן, מקדמים גדולים מגדילים את $\left\|w\right\|_1 = |c_1| + ... + |c_d|$, מה שמצביע על מקדם רגולריזציה α נמוך. שימוש במקדמים כאלו מגדיל את ה- $overfitting$ של המודל.

- כאשר למקדם c_j כזה יש סדר גודל קטן הוא מקטין יותר את הערך $c_j\left(x_i\right)_j$ מה שמצביע על כך שהמודל מעוניין להתחשב פחות בפיצ'ר ה- j של דוגמאות. כמו כן, מקדמים קטנים מקטינים את $\left\|w\right\|_1 = |c_1| + ... + |c_d|$, מה שמצביע על מקדם רגולריזציה α גדול. שימוש במקדמים כאלו מקטין את ה- $overfitting$ של המודל.

שאלה 12

נשים לב שחישבנו את ביצועי המודל $Lasso_{w,\alpha}$ ובדקנו מהו ה- $MSE = \frac{1}{2m}\sum_{i=1}^{m}\left(w^T x_i - y_i\right)^2 + \alpha \left\|w\right\|_1$ שהמודל משיג. נשים לב שאם המודל מבצע fit ל- $data$ בעזרת אלגוריתם שמושפע מנרמול הפיצ'רים כמו SGD אז ללא נרמול נקבל התכנסות איטית יותר (אם בכלל) למינימום לוקאלי (של ה- MSE) במסגרת תהליך האימון. בנוסף, אי נרמול הפיצ'רים גורם לכך שמקדמים מסויימים של w יהיו קטנים במיוחד כדי "לאזן" את הפיצ'רים הגדולים. כלומר, המודל ינסה בכוח להתאים את עצמו על בסיס פיצ'רים גדולים בלבד ולכן ביצועיו בתהליך האימון יפחתו.

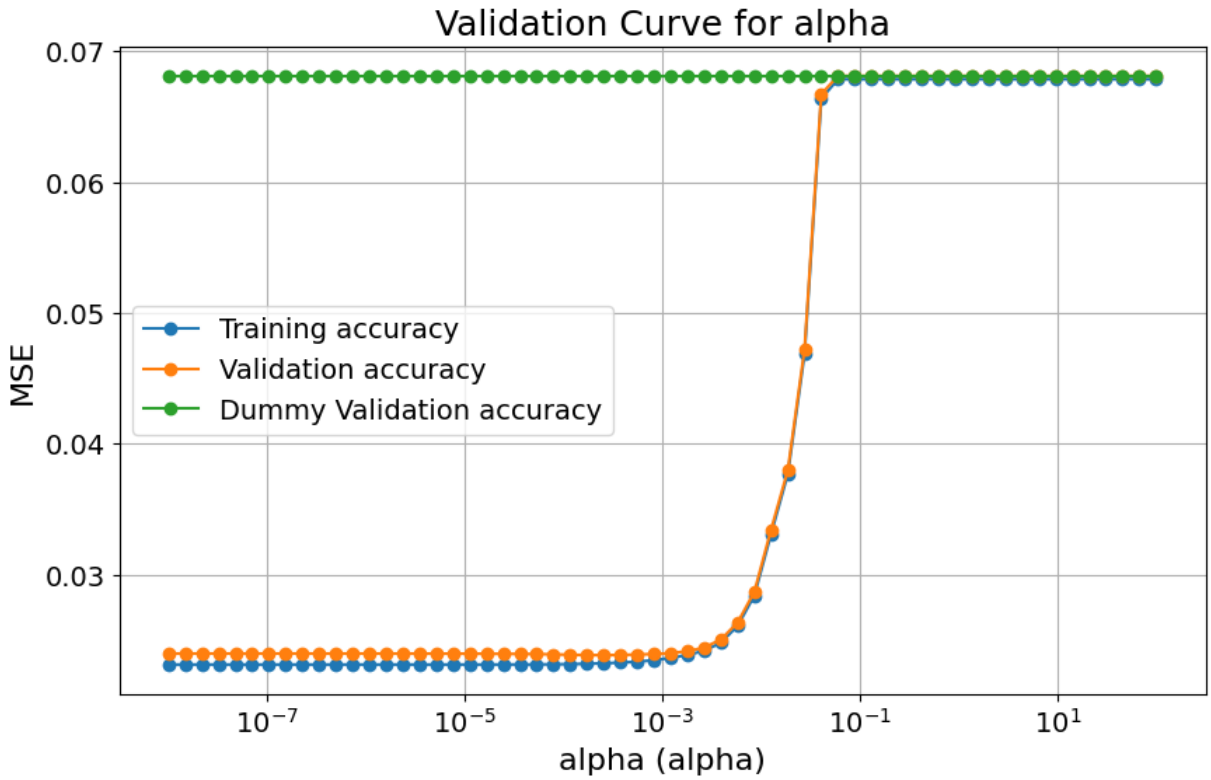
שאלה 13

אילו היינו משתמשים במודל $Ridge\ Regressor$ במקום היינו מצפים (כפי שראינו בתרגול) לקבל פתרונות $w \in \mathbb{R}^d$ לבעיה שהם דלילים פחות. זאת משום שכפי שראינו בתרגול פתרון בעיית ה- $Lasso\ regressor$ מניב מודל עם $w \in \mathbb{R}^d$ דליל (עם מקדמים קטנים יותר שחלקם אף אפסים).

שאלה 14

כאשר אנו מפעילים על הפיצ'רים מיפוי לפולינום מדרגה 4 אנו יוצרים פיצ'רים חדשים גדולים. למשל פי'צר a בדוגמה גורר קיום פיצ'ר a^4 במיפוי החדש. כמו כן, אם נירמלנו את a לטווח $[0,1]$ אז a^4 יהיה מאוד קטן. לכן, אם ננרמל את הפיצ'רים לפני הפעלת המיפוי אנו עלולים לקבל פיצ'רים חדשים <u>לא מנורמלים לטווח הרצוי</u>. כלומר, מיפוי הפיצ'רים עשוי ליצור פיצ'רים חדשים גדולים מהטווח המותר ובכך להרוס את הנרמול שביצענו לפיצ'רים המקוריים. לכן, חשוב לנרמל את הפיצ'רים לאחר המיפוי שלהם ובכך להבטיח שנעשה שימוש בפיצ'רים ממופים מנורמלים לטווחים הרצויים.

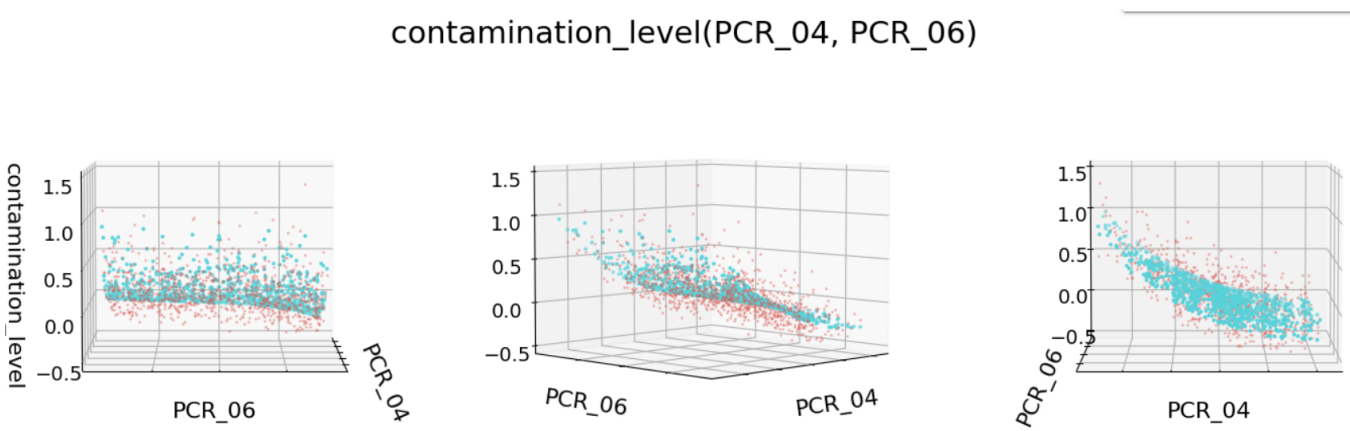
שאלה 15




```
Best alpha: 0.00025514065200312873
Average training loss for best alpha: 0.0232
Average validation loss for best alpha: 0.0239
```

כלומר ה- α הטוב ביותר הוא $\alpha = 0.0002$ ושגיאת הולידציה הממוצעת שלו היא 0.023.

שאלה 16

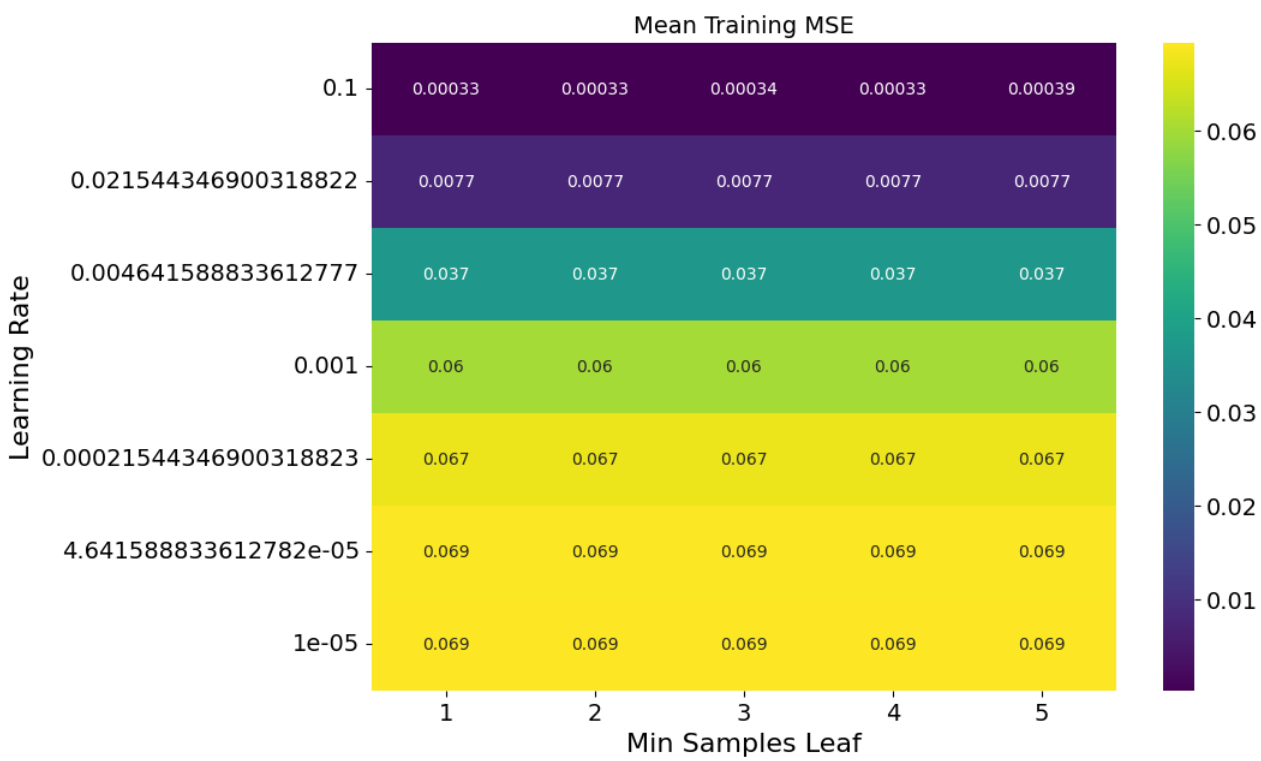


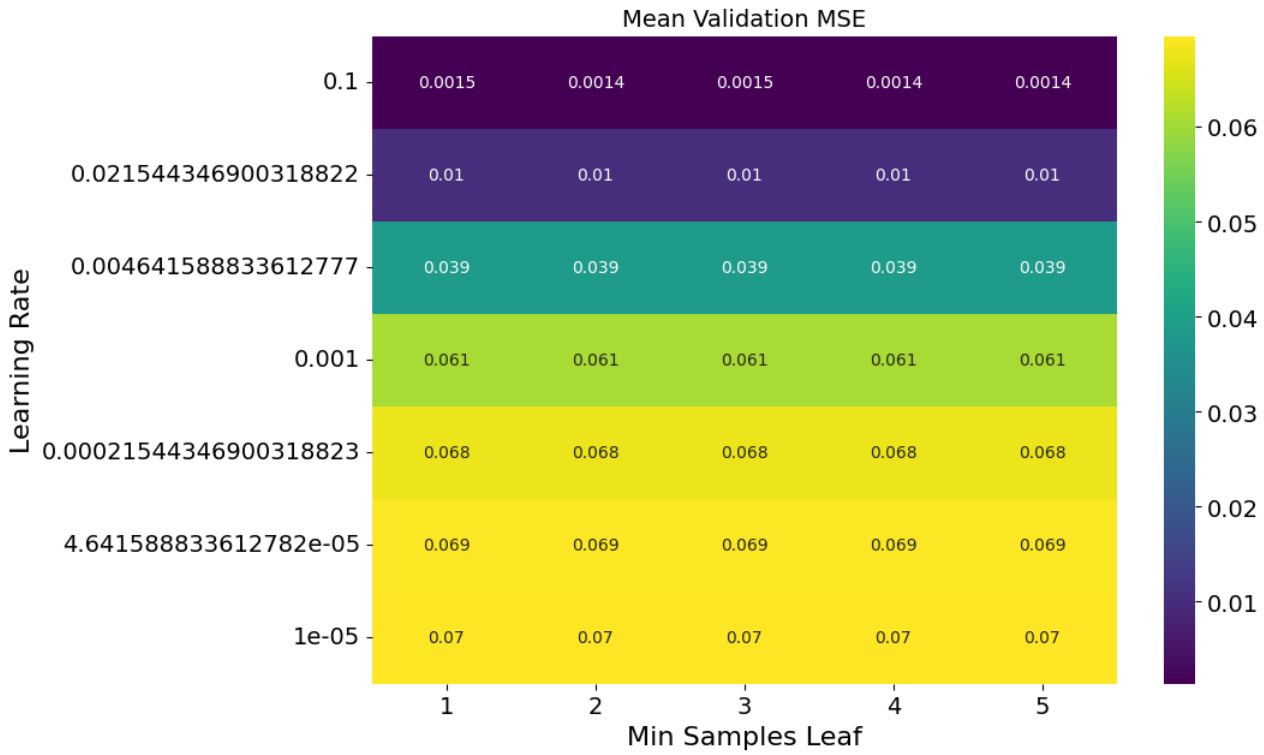
שאלה 17

בהתאם לגרף שקיבלנו בשאלה 15 :

Model	Section	Train_MSE	Valid_MSE
		Cross_validated	
Dummy	2	0.067	0.068
Linear	2	0.022	0.023
Lasso_Linear	3	0.022	0.022
Polynomial_Lasso	4	0.023	0.023

שאלה 18






```
Optimal Parameters after inversion: {'regressor__learning_rate': 0.1, 'regressor__min_samples_leaf': 4}
Best Training MSE:  0.000334647729114145
Best Validation MSE:  0.0013590101986925198
```

כפי שניתן לראות ה- lr האופטימלי הוא $lr = 0.1$ וה- $min\_samples\_leaf$ האופטימלי הוא $min\_samples\_leaf = 4$.

שאלה 19

Model	Section	Train_MSE	Valid_MSE
		Cross_validated	
Dummy	2	0.067	0.068
Linear	2	0.022	0.023
Lasso_Linear	3	0.022	0.022
Polynomial_Lasso	4	0.023	0.023
GBM_Regressor	5	0.0003	0.0013

שאלה 20

Model	Section	Train_MSE	Valid_MSE	Test_MSE
		Cross_validated		Retrained
Dummy	2	0.067	0.068	0.0694
Linear	2	0.022	0.023	0.0302
Lasso_Linear	3	0.022	0.022	0.0301
Polynomial_Lasso	4	0.023	0.023	0.0326
GBM_Regressor	5	0.0003	0.0013	0.0015

כפי שניתן לראות מהטבלה:

- המודל שביצועיו (שגיאת המבחן) היו הטובים ביותר על ה- $test\ set$ הוא $GBM\ Regressor$. מודל זה מהווה $overfitting$ לדוגמאות האימון ביחס לשאר המודלים כי שגיאת האימון שלו היא הקטנה ביותר מבין המודלים.
- שגיאת המבחן של $dummy$ היא הגדולה ביותר שכן מדובר במודל הפשוט והחלש ביותר שמצמיד לכל דוגמה את התווית

הממוצעת של דוגמאות האימון. כמו כן המודל מהווה $under fitting$ לדוגמאות האימון ביחס לשאר המודלים כי שגיאת האימון שלו היא הגדולה ביותר מבין המודלים.

- שגיאות המבחן/הולידציה/האימון של $Linear,\ Lasso\ Linear,\ Polynomial\ Lasso$ כמעט זהות. נסיק מכך שהמודלים אינם עדיפים אחד על השני ביחס ל- $dataset$ שלנו.