תרגיל בית 2

A screenshot of a computer

Description automatically generated

Chart, scatter chart, bubble chart

Description automatically generatedChart, scatter chart

Description automatically generated

ניתן לראות שההבדל העיקרי בין המודלים הוא שכאשר אנו עובדים עם k=1 אנחנו מקבלים שישנה התאמה לכל ערכי המטרה של ה-training\_set אך יש תופעה של overfitting

ניתן לראות זאת על ידי העיגולים הכחולים הנוצרים בתרשים 1-knn אשר נגרמים על ידי זה שדוגמא מסוימת היא קרובה יותר לדוגמא בעלת תיוג כחול כאשר אך מוקפת בדוגמאות בעלות תיוג אדום (אך במרחק גדול יותר). ואנו ממסווגים את הדוגמא הזאת כחול.

ניתן לראות שעבור k=9 רמת הדיוק על סט האימון יורדת מפני שישנם יותר ערכים בעלי תיוג הופכי. אך בעל חסינות גבוהה יותר לoverfitting לעומת המודל עם k=1.

Graphical user interface, text

Description automatically generated

PCR\_10 0.132957

PCR\_07 0.064828

household\_income 0.053053

sport\_activity 0.044553

blood\_type\_O+ 0.028198

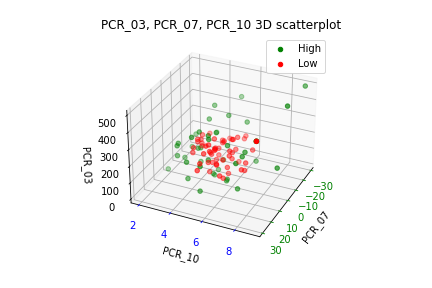
shortness\_of\_breath 0.027824

PCR\_04 0.026156

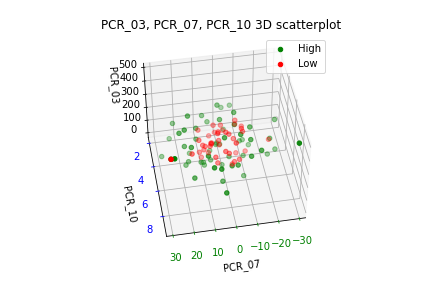
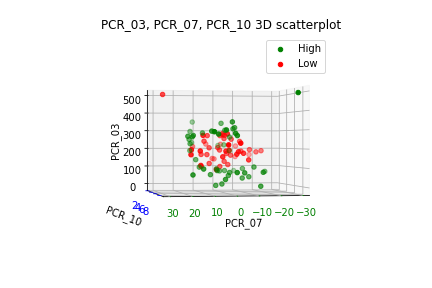
conversations\_per\_day 0.025367

num\_of\_siblings 0.019521

age 0.019385

Graphical user interface, text, application, email

Description automatically generated



Graphical user interface, text

Description automatically generated

ניתן להסיק את הדברים הבאים:

* לפי הגרף שיצאו ניתן לראות שאנשים אשר היו בעלי ערכים גבוהים יותר ב 07\_PCR , 03\_PCR יחדיו ונמוכים ב 10\_PCR קוטלגו כבעלי spread גבוהה
* ניתן לראות שהערכים שקוטלגו כ-spread נמוך נמצאים בתחום סגור (באמצע) ב PCR\_03 בעוד שערכים שקוטלגו כ spread גבוה נמצאים בקווצת ה PCR\_3
* למעשה האנשים שקולטו בעלי סיכון גבוה ,נמצאים יותר בקווצות ואהשים שקולגו בעלי סיכון נמוך, נמצאים במרכז.

Graphical user interface, text, application

Description automatically generated

רמת הדיוק אותה קיבלנו היא 0.82

Graphical user interface, text

Description automatically generated

Z-score יודע להתמודד עם outliers מפני שהכל מוזז בהתאם לממוצע וסטיית תקן  
אך מכיוון שכל תכונה מושפעת רק מהממוצע וסטיית התקן שלה, התחום המתקבל לכל תכונה הוא לא בהכרח זהה, כלומר לא בין 0 ל 1

לדוגמה – לתכונה A עם ו סטיית תקן של 1 נקבל תחום של   
לעומת מקרה של תכונה B עם וסטיית תקן של 20 נקבל שכל הערכים מקיימים

לפי הנוסחה של z-score , לפיכך הערכים יהיו בין   
ותחום זה שונה מהתחום הקודם ולכן כל תכונה תשפיע על המרחק באופן אחר .

נבחר בתכונה זו במקרה שתחומי הערכים של התכונות זהים על מנת לאפשר טיפול ב outliers ושההתפלגות הינה התפלגות גאוסינית.

Min-max scaling הנוסחה הזו מנרמלת את הערכים של כל תכונה לתחום , אך אינה יודעת להתמודד עם outliers , (מפני שהערכי הקיצון יקבלו את הערכים 1 או 0 דבר הישנה את ההתפלגות ויקשה עלינו לזהות את הoutliers )  
אך מכיוון שהנוסחה מעבירה את כל התכונות לתחום כל התכונות משפיעות באותו אופן על המרחק (לדוגמה ב KNN , קירבה תושפעה באופן שווה מכל התכונות ותיהיה מדוייקת יותר לעומת z-score שם תכונה אחת עלולה ליצור קרבה "מזוייפת" )

ההיתרון הבולט של שיטה זו היא שחישוב מרחק של דוגמה מדוגמה אחרת נעשה באופן המתייחס באופן שווה לכל התכונות ובכך יכול למנוע טעויות וכשאין outliers ברורים

Graphical user interface, text

Description automatically generated

לאחר נירמול הערכים באופן הבא :

PCR\_03 בעזרת z-score  
PCR\_07 בעזרת z-score  
PCR\_10 בעזרת z-score

קיבלנו את התוצאה 0.92 על סט האימון

נרמול הערכים מקטין את ההבדלים בין ערכי תכונות ובכך מאזן את ההשפעה של כל תכונה על המרחק בין כניסות שונות ב training set . לכן נוצר שיפור ברמת הדיוק. הקטנו את המקרים בהם ערך של תכונה מסוימת יצר מרחק מאוד גדול (בשל ההבדל בסדרי גודל בין תכונה זו לשאר התכונות) ולכן הובילה לתיוג שגוי.

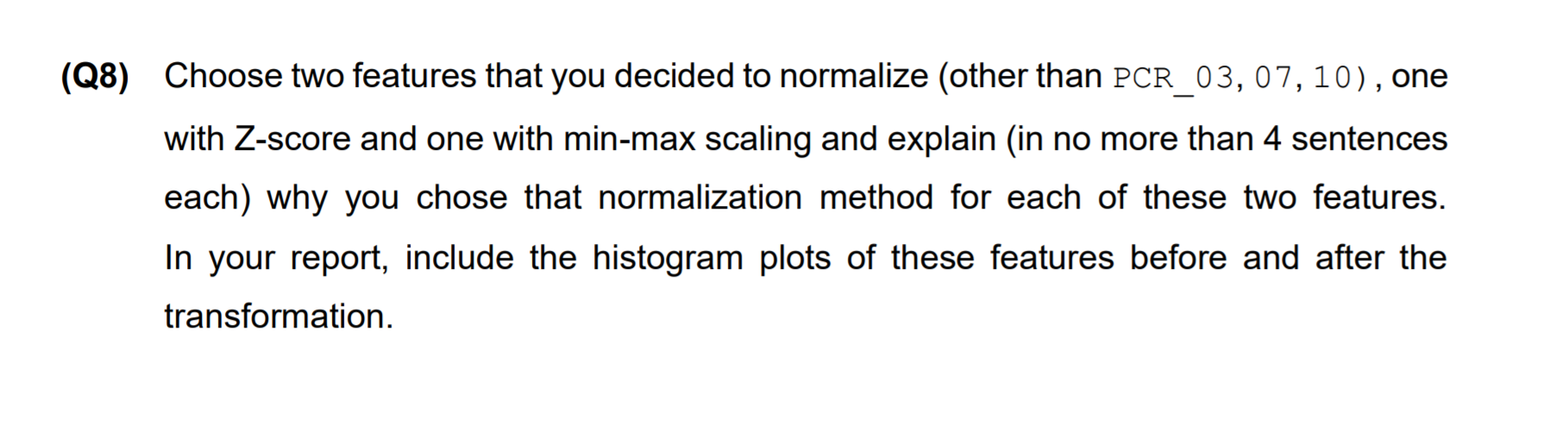
במקרה זה,

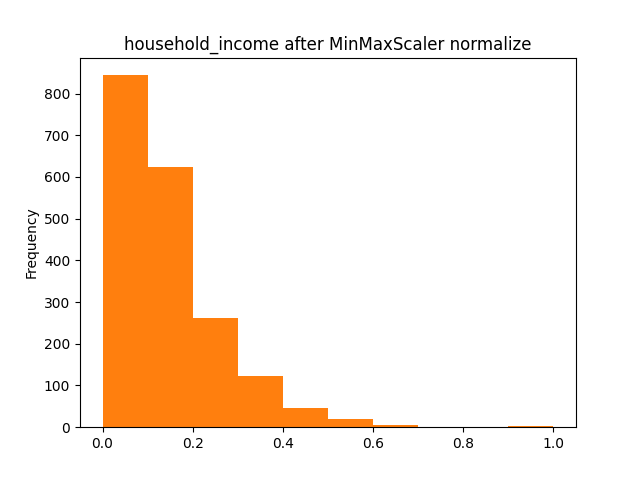
תכונה PCR\_03 עם תחומי ערכים [493,-16]

תכונה PCR\_07 עם תחומי ערכים

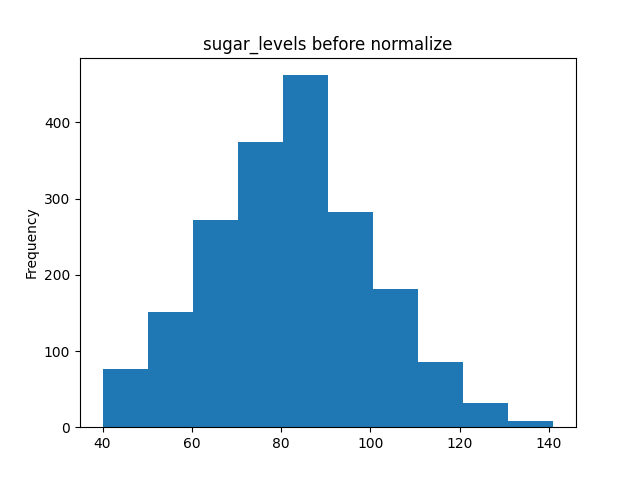
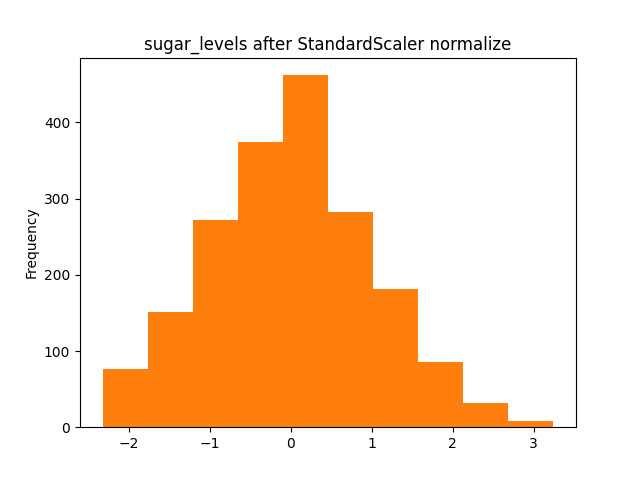
תכונה PCR\_10 בין

למרחק בתכונה PCR\_03 ישנו השפעה גדולה יותר על התוצאה משאר התכונות ועלול להוביל לתיוג שגוי עקב כך



Chart, histogram

Description automatically generatedבחרנו את השדות household\_income ו suger\_levels

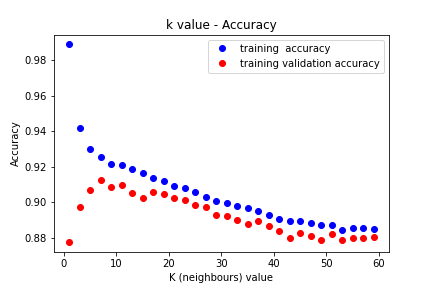


כפי ששמנו לב השדה household\_income אינו דומה להתפלגות גוסיאנית, ולכן בחרנו לנרמל אותו לפי minmax ואכן ניתן לראות שההתפלגות נשמרת בצורה טובה

כמו כן השדה suger\_levels דומה להתפלגות גוסינית ולכן בחרנו לנרמל לפי strandScaler ואכן ניתן לראות שההפתלגות נשמרת בצורה טובה.

A screenshot of a computer

Description automatically generated



ה K שנבחר הינו 9, משום בערך זה ,נקבל את הלוידציה הגבוהה ביותר שהיא 0.9127

ממוצע דיוק האימון הינו 0.9056

ממוצע דיוק הולידציה הינו 0.8924

A screenshot of a computer

Description automatically generated

Chart, treemap chart

Description automatically generated

ניתן לראות כי יש יותר false positive מאשר false negative

Graphical user interface, text

Description automatically generated

שאלה 11,

כפי שראינו בדוח הקודם שמנו לב שיש קורולציה בין risk לבין sugar\_level, ניתן לראות זאת על סמך הגרף הבא

*Chart, histogram

Description automatically generated*

ועל סמך חישוב הקורולוציה שיוצא **0.2876366**

וכמו כן נשים לב כי –

Chart, scatter chart

Description automatically generated

כלומר , PCR\_01 וגם PCR\_02 יכולים לשמש לסיווג טוב של risk

לפיכך נבחר את הנתונים הבאים:

PCR\_1,PCR\_2, sugar\_level

Text

Description automatically generated with medium confidence

חישובי הקורולציה, ונקבל:

sugar\_levels 0.287637

PCR\_06 0.184797

sport\_activity 0.105596

age 0.101714

PCR\_03 0.053002

PCR\_07 0.043751

PCR\_02 0.036989

PCR\_01 0.032614

blood\_type\_B- 0.028734

fever 0.023476

Graphical user interface, text

Description automatically generated

להלן העץ המתקבל

Diagram

Description automatically generated

ורמת הדיוק המתקבלת הינה 0.772

Graphical user interface, text, application, email

Description automatically generated

כן, נשים לב למשל כי sugar\_level מופיע גם בשאלה 11 וגם בשאלה 12.

כעת נבצע ניתוח קטן של כל השדות ואיפה הם נמצאים בעץ ההחלטה:

sugar\_levels 0.287637

PCR\_06 0.184797

sport\_activity 0.105596

age 0.101714

PCR\_03 0.053002

PCR\_07 0.043751

PCR\_02 0.036989

PCR\_01 0.032614

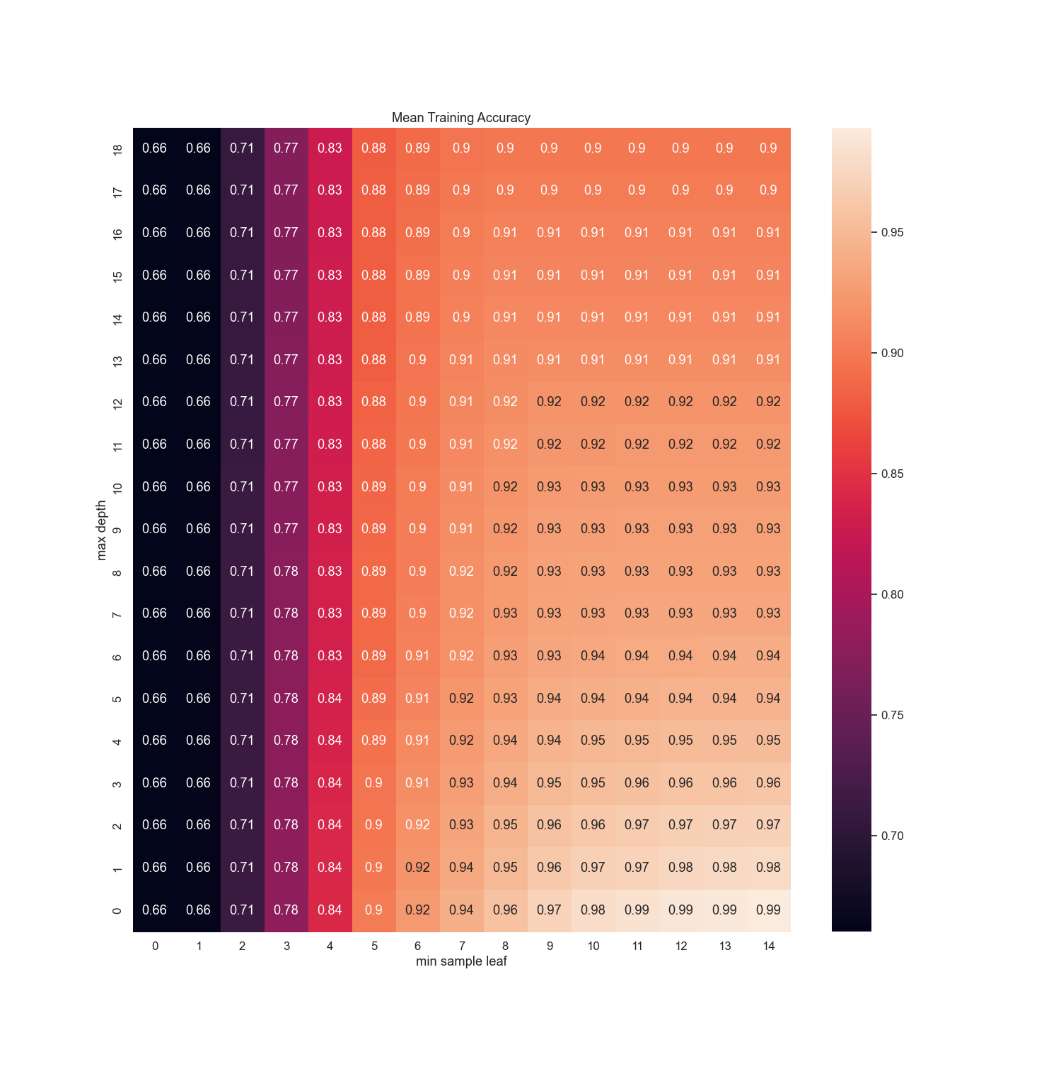
blood\_type\_B- 0.028734

fever 0.023476

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| באיזה רמה | מופיע בעץ החלטה | מופיע ב שאלה 13 | מופיע ב שאלה 12 | תכונה |
| רמה 1 רמה 2 רמה 3 | כן | כן | כן | sugar\_levels |
| רמה 0 רמה 2 רמה 3 | כן | כן | לא | PCR\_06 |
|  | לא | כן | לא | sport\_activity |
|  | לא | כן | לא | age |
| רמה 3 | כן | כן | לא | PCR\_03 |
|  | לא | כן | לא | PCR\_07 |
| רמה 2 רמה 3 | כן | כן | כן | PCR\_02 |
| רמה 1 רמה 2 | כן | כן | כן | PCR\_01 |
|  | לא | כן | לא | blood\_type\_B- |
| רמה 3 | כן | לא | לא | Household\_income |
|  | לא | כן | לא | fever |

להלן מספר ממצאים:

* נשים לב כי מרבית התכונות שרשמנו בשאלה 12, מופיעות גם בעץ ההחלטה
* נשים לב ש sugar\_leveles וגם sport יש קורולציה מאוד גבוה יחד עם risk וכמו כן מפיעים ברמות העלינות של העץ החלטה.
* כמו כן נשים לב שחלק מן התכונות שמופיעות בשאלה 13,לא מופיעות בעץ ההחלטה:
  + דבר זה ייתכן משום שיש קורלציה בין sugar\_level לבין sport\_activety ולכן לסנן עם שניהם לא יוסיף יותר מיידי מידע,
  + כמו כן יתכן שבין age לבין חלק מן התכנונות המפויעות בעץ גם יש קורולציה המונעת ממנו להופיע גם בעץ ההחלטה
* נשים לב כי אמנם hosehold\_income יש קורולציה נמוכה יחד עם risk אך מצד שני כן מופיע בעץ החלטה ולכן יתכן כי שילוב תכונה זו עם שאר התכונות עוזרות לבנות את העץ החלטה.

Calendar

Description automatically generatedGraphical user interface, text, application, Word

Description automatically generated with medium confidence

* דוגמא ל overfitting ניתן לראות כאשר מספר העלים הוא גדל ועומק העץ קטן, למשל ניתן לראות זאת כאשר יש לפחות 13 דוגמאות בעלים ויש רק פיצול יחיד , אז ה training נותן דיוק של 99 אחוז ולעומת זאת ה validation נותן דיוק של 86 אחוז.
* דוגמא ל underffiting ניתן לראות כאשר מספר הדוגמאות המינימלי בעלים הוא 0, או 1 , ניתן לראות שלא משנה כמה פיצלולים נבצע בעץ גם ה validation וגם ה training טועים (66 אחוזי דיוק)

Graphical user interface, text, application

Description automatically generated

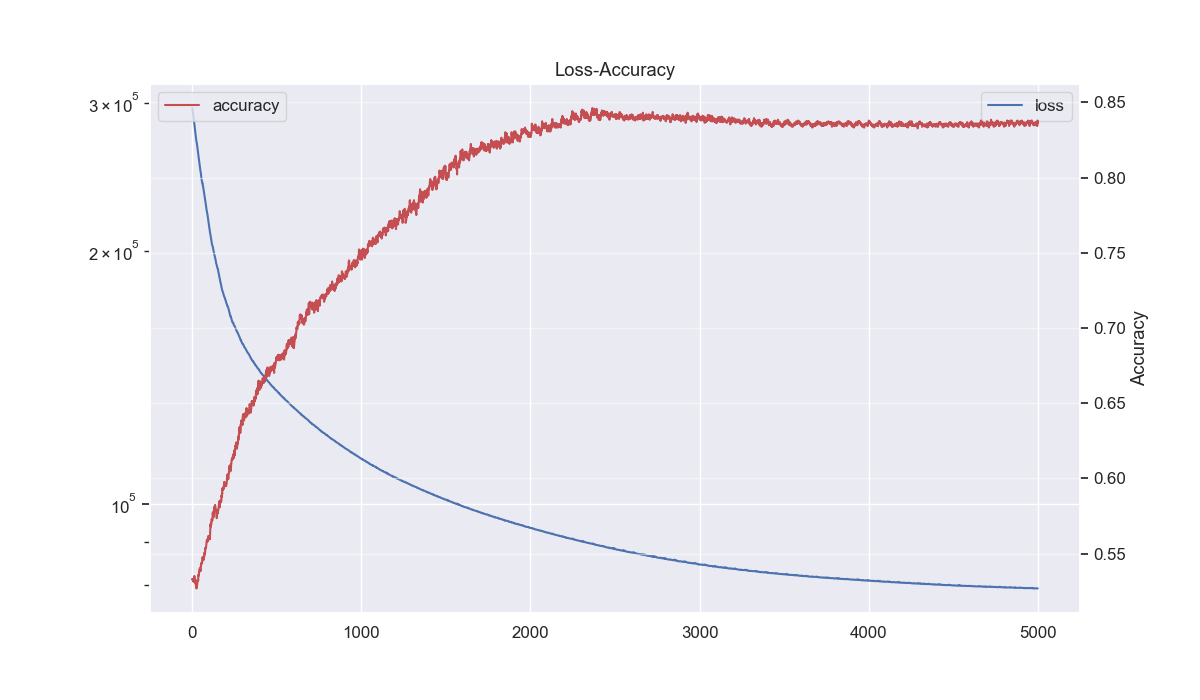
Chart, line chart

Description automatically generated

ככל שאנו מגדילים את הדלתא של ה גרדיאנט הנומארי הפער בינו לבין הגרדיאנט האנליטי גדל

A screenshot of a computer

Description automatically generated with medium confidence



הדיוק שקיבלנו הינו 85 אחוז,

Chart

Description automatically generatedText

Description automatically generated

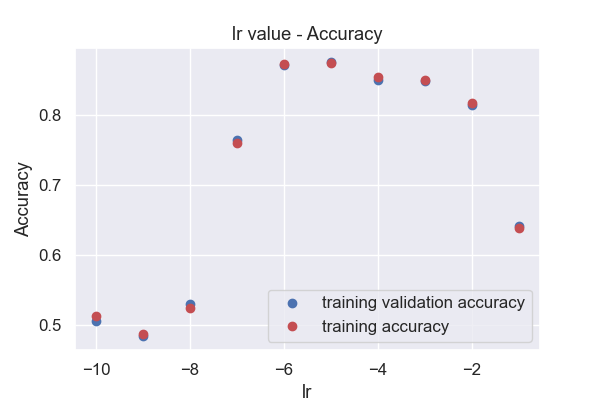
לפי, רמת הדיוק המקסימלית שניתן להגיע אליו עם softsvm על 3 התכונות pcr\_03,pcr\_07,pcr\_10

הוא 60 אחוז לעומת רמת דיוק של 92 אחוז בעת שימוש ב KNN עם K=9

softsvm במקרה של spread אינו יכול להתחרות עם מודל KNN בשל רמת דיוק מקסימלית הקטנה מרמת הדיוק המקסימלית של KNN

Graphical user interface, text

Description automatically generated



נשים לב כי אחוז הדיוק גם ב training וגם ב validation הם בערך אותו דבר,

כעת לגבי הלמידה,

נשים לב שכאשר ממעריך של ה lr הוא קטן , בין ל למעשה אנחנו מבצעים צעד קטן מאוד ולכן ייתכן כי אנחנו לא מספיקים להתכנס לפתרון האופטימלי של הבעיה, למעשה הבעיה הינה בעיה קמורה ולכן אנחנו כן אמורים להגיע לפתרון ולא להכנס לתוך עוקף.

כעת כאשר הגדלנו יותר את ה lr להיות גדול מ אז למעשה קיבלנו דיוק יותר רב ובכל שמגידלים אז אנחנו מקבלים דיוק רב יותר (וזה אומר שאנחנו מקבלים פתרון מהר יותר משום שגודל הצעד גדול יותר) עד שאנחנו מקבלים את המקסימום שהוא , לאחר מכן בכל שמגדלים את ה lr אז אנחנו יורדים באחוזי הדיוק וזאת משום שאנחנו מבצעים צעדים יותר מיידי גדולים, מה שגורם לפספוס הפתרון האופטימלי.

ישנה אפשרות לשפר את המודל בעזרת זה שבהתחלה נבצע lr יחסית גדולים, שנגיע לאיזור הפתרון האופטימלי ולאחר מכן להקטין את ה lr שנעשה צעדים קטנים יותר ויותר עד שבסופו של דבר נקבל את הפתרון האופטימלי עם אחוזי הדיוק הגבוהים ביותר.

Graphical user interface, text

Description automatically generated

ההבדל בין מה שעשינו בשאלה 19 ל kernel svm הוא שאנו ביצענו feature mapping ללא בדיקה שהתכונות מקיימות את הדרישה kernel ש ולכן ישנה אופציה שיצרנו תכונות חדשות אשר אינן ניתנות ליצירה מהתכונות הקיימות בלי להרחיב את מרחב התכונות שלנו

המטרה ב kernel להציג את המידע באופן אחר בלהסתמך על המידע הקיים בלבד.

המטרה ב kernel הוא להציב באופן אחר ללא הוספת מידע אלא הוספת תכונות המתבססות על התכונות הקיימות בלבד (בעצם תצוגה שונה)

A picture containing text

Description automatically generated

לגבי , ה SVM , אנחנו שומרים את W שהוא וקטור בגודל כמות ובנוסף שומרים וקטור המכיל את הסיווג של כלל הדוגמאות ולכן המודל צורך זיכרון יחסית נמוך

לגבי ה KNN , אנחנו שומרים את כל הדוגמאות בתוך המודל, ולכן הוא צורך זיכרון גדול יותר