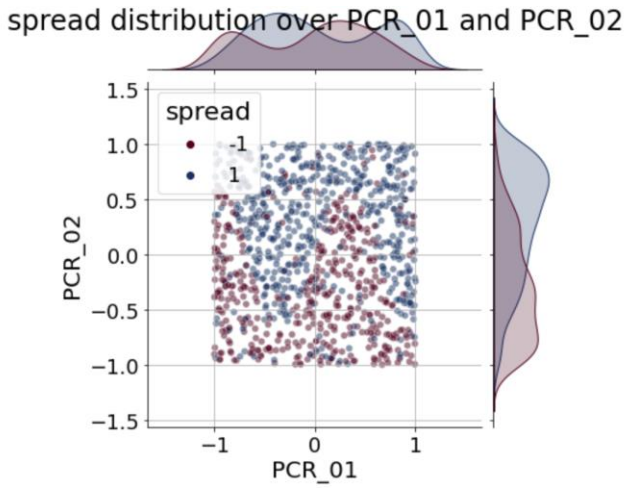
Major 02 Report

Part 1: Basic model selection with k-Nearest Neighbors

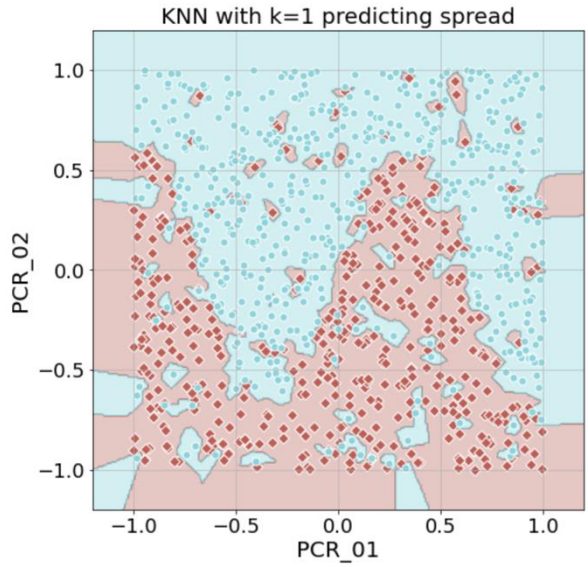
Visualization and basic analysis

שאלה 1:



שאלה 2:

אימנו מודל על סט האימון עם הפרמטר K=1:



Model selection

שאלה 3:

ערך ה-k שהתקבל כאופטימלי עבור kNN הינו k=7 .

נשים לב שעבור ערך זה של k אנחנו מקבלים דיוק גבוהה עבור סט האימון ובנוסף גם דיוק גבוהה על סט הוולדיציה, ניתן לראות לפי הגרף שהדיוק המתקבל בערך זה כבור שני הסטים קרוב מאוד.

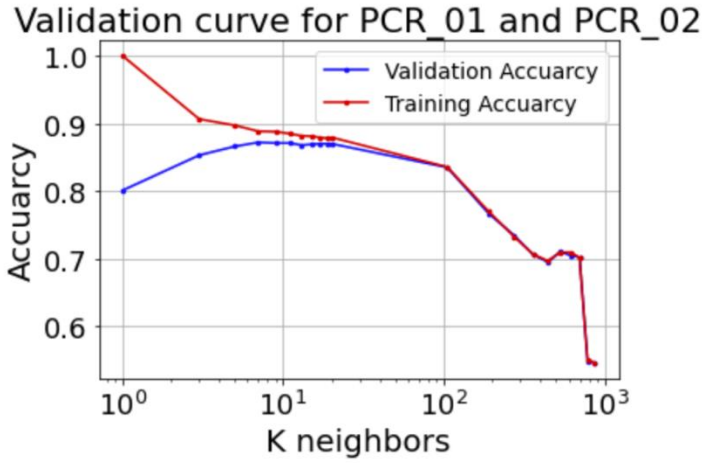
על מנת לבחור את הפרמטר האופטימלי עלינו לבחור פרמטר כזה שעבורו הדיוק בסט האימון הוא הגבוהה ביותר (וה- loss הוא הנמוך ביותר) וגם ההפרש בין הדיוק של סט הוולידיציה לדיוק של סט האימון הוא המינימאלי ביותר .

הדיוק המתקבל עבור דיוק סט הוולידציה היא : 0.872

ועבור סט האימון: 0.8884285714285713.

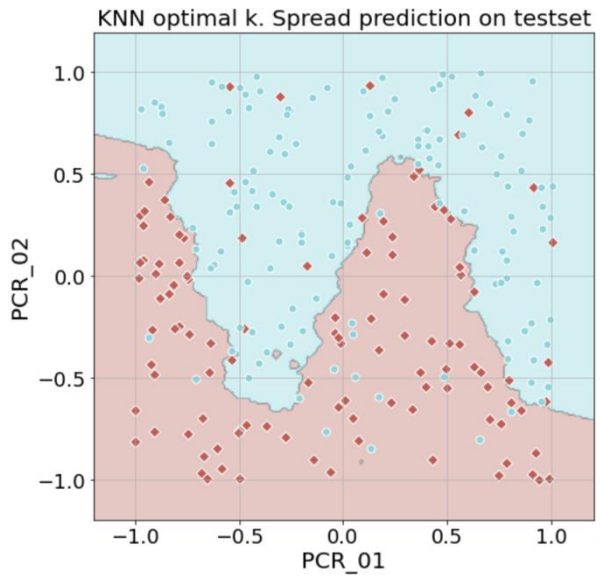
עבור הערכים K<5 מתקיים כי הדיוק של סט האימון גבוה לעומת סט הוולידציה(מבחן), לכן מתקיים בטווח ערכים זה - Overfitting.

עבור הערכים K>100 מתקיים כי דיוק האימון קטן ולכן בטווח ערכים מתקיים- Underfitting. .



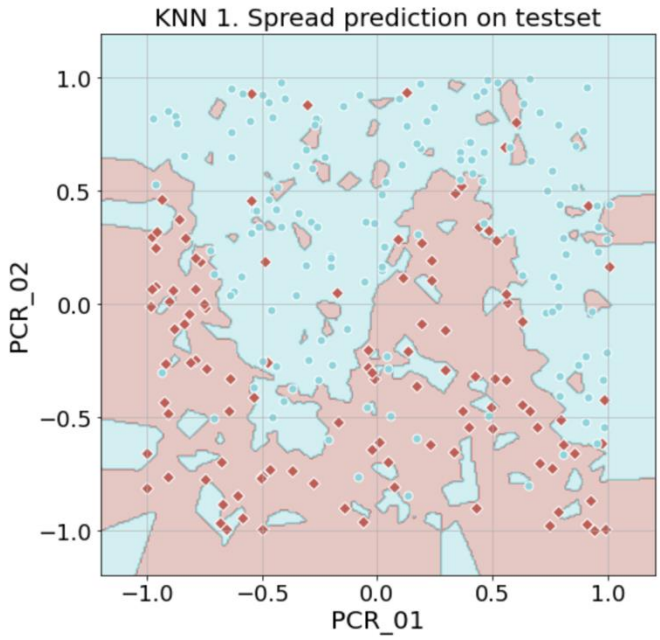
שאלה 4:

גרף הK האופטימלי -



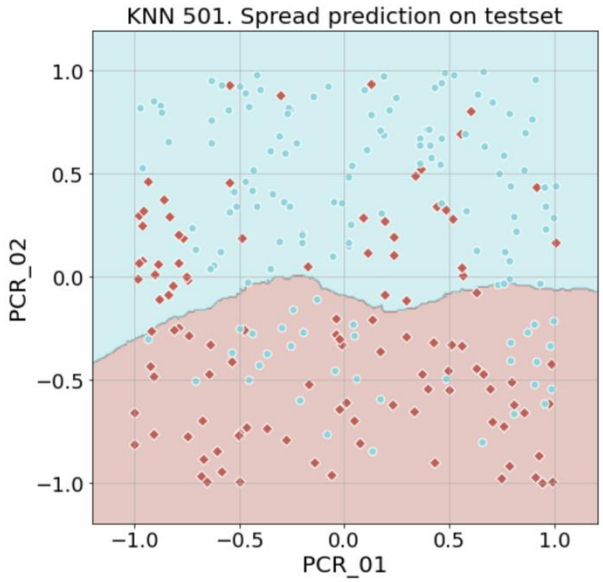
הדיוק של הK האופטימלי על המבחן הוא: 0.856.

שאלה 5:



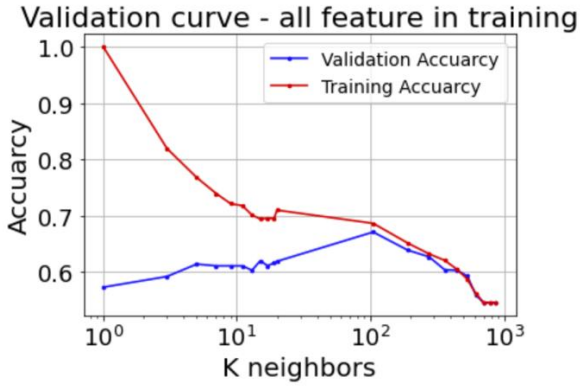
ניתן לראות שעבור k=1 הגבולות מאוד רגישים, הכוונה שהמודל מותאם מאוד לסט האימון, זאת אומרת שאנחנו נצפה ל Overfitting. וכן ניתן לראות מהצגת הגרף על k=1

כי ישנם טעויות על samples שלא נראה על ידי המודל - המודל אינו גמיש/מכליל עבור מידע אחר. ועל כן ניתן לראות כי דיוק מודל זה נמוכה מהאופטימלי: **0.796.**



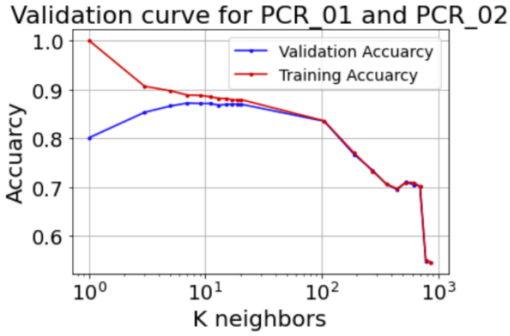
בנוסף ניתן לראות עבור k=501 כי הגבול בגרף הוא מאוד אדיש, ישנו קו יחסית ישר החותך את כל מרחב הsamples. מודל זה מבצע Underfitting שכן כפי שניתחנו את הפרמטר K (בשאלה 3), הדיוק על סט האימון הוא נמוך. וכן ניתן לראות כי הדיוק על סט המבחן נמוך גם כן: **0.688.**

שאלה 6:



בהרצאה בכתה הוסבר לנו שמודל KNN עובד בצורה פחות יעילה על דאטה סט בעל מספר פיצ'רים גבוה יחסית למספר הדגימות תחת ההסבר כי ככל שמספר המימדים של דגימה גדולה יותר ה"מרחק" בין 2 נק' קרוב יותר. על כן ככל שמספר העמודות גדול יותר אנו צופים שהנק' במרחב יהיו יותר קרובות אחת לשנייה ועל כן יתנו מס' רב של שגיאות על מדגם המבחן.

ובאמת, ההבדל המשמעותי בין גרף זה לגרף בשאלה 3 הוא שבגרף זה, כאשר אימנו על כל הפיצ'רים הדיוק גם של האימון וגם של הולידציה נמוכים יותר בממוצע וכן דיוק המקסימלי של הולידציה נמוך משמעותית מאימון 2 הפיצ'רים.

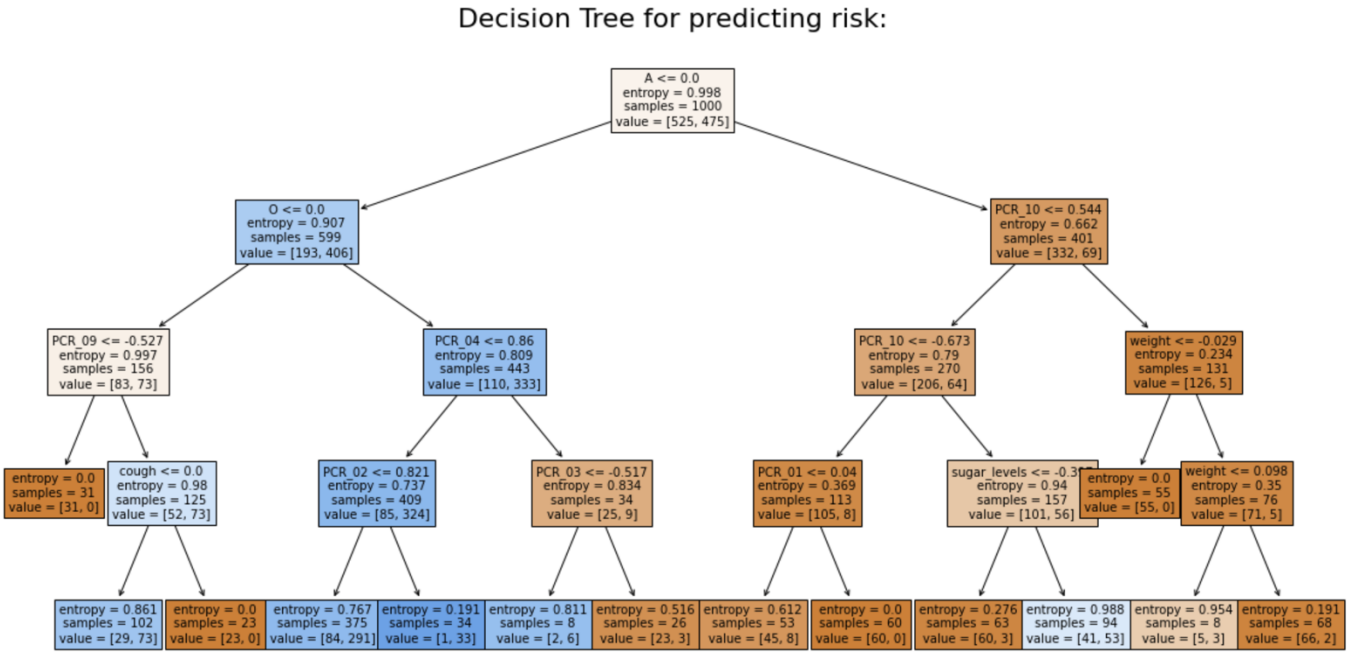


Decision trees :Part 2

Visualization

שאלה 7:

שגיאת האימון שקיבלנו היא : **0.78**



Model selection

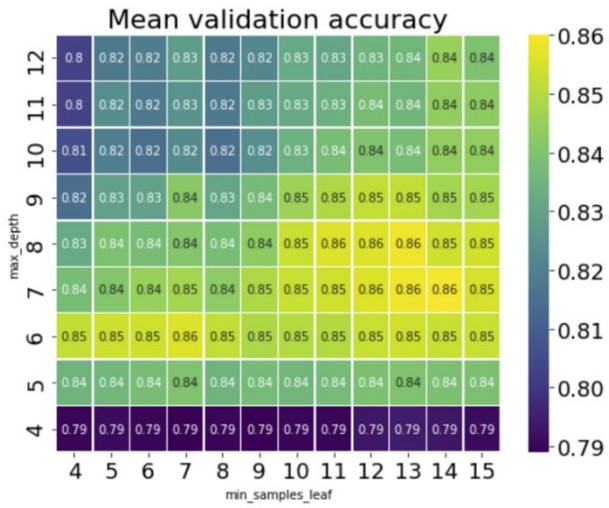
שאלה 8:

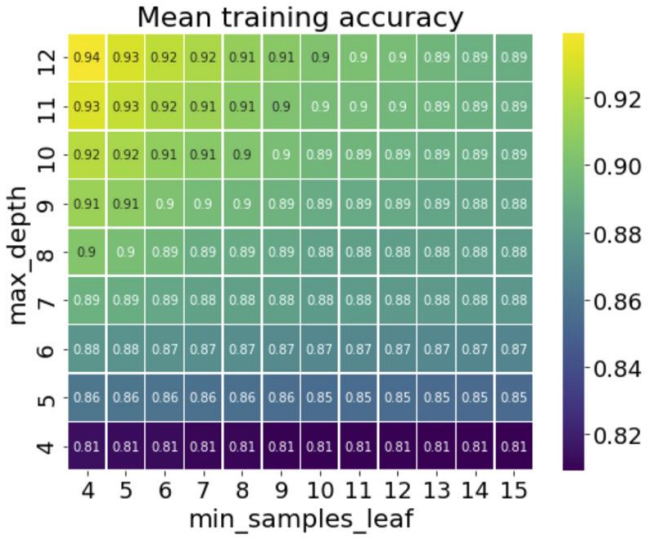
a) לאחר בדיקת טווחי ערכים שונים וחיפוש עבור ערכים עבורם שגיאת הוולידציה היא הגבוהה ביותר בחרנו להשתמש בטווחים הבאים :

Max_depth : [4,12]

Min_samples_leaf: [4,15]

b)





c) הערכים שעבורם נקבל את הדיוק האופטימלי הם :

max_depth=6 ו- min_samples_leaf=7 .

הדיוק ב-validation הוא 0.86

d) הערכים שעבורם נקבל Underfitting הם :

max_depth=4 ו- min_samples_leaf=15.

e) הערכים שעבורם נקבל overfitting הם :

Min_sample_leaf = 4 ו max_depth = 12

f) עבור הערכים בסעיף d נקבל שהדיוק בסט האימון הוא הנמוך ביותר (ושגיאה הכי גדולה), במצב זה מתקבל Underfitting.

עבור הערכים בשאלה e נקבל שההפרש בין הדיוק בסט הוולדיציה לסט האימון הוא הגדול, ביותר וזה המצב בו נקבל Overfitting.

עכשיו נבין את הקשר בין הפרמטרים לתוצאות שקבלנו.

ככל שבעץ שלנו יש יותר עלים אז המודל המתקבל מסובך יותר, הפרמטרים max_depth ו-min_samples_leaf שולטים על מספר העלים בעץ.

ככל שעומק העץ גדול יותר וככל שהמספר הדגימות המינימאלי בעלה קטן יותר אז נקבל עלים.

מודל "מסובך מידי" יקבל Overfitting ו-"מודל פשוט מידי" יקבל Underfitting.

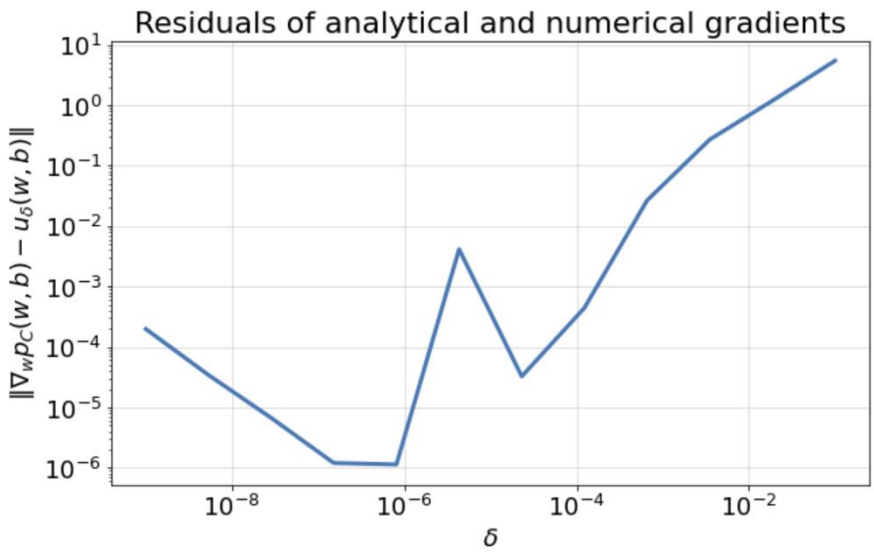
שאלה 9:

עבור עץ ההחלט בעל סט הערכים האופטימליים (משאלה 8c) נקבל כי הדיוק המתקבל עבור סט מבחן הוא 0.828.

Part 3: Linear SVM and the Polynomial kernel

Verifying your implementation: Numerical vs. analytical gradients

שאלה 10:



ניתן לראות שבאיזור שבו דלתא הוא 10^{-6} נקבל ההפרש בין הנגזרת בנקודה והגרדיאנט האנליטי הוא המינימלי ולכן ישנה הערכה מדויקת כאשר דלתא בנק' זו.

בערכים הקטנים מהנק' הנ"ל נקבל כי ההפרש מתחיל לגדול הנובע משגיאות נומריות.

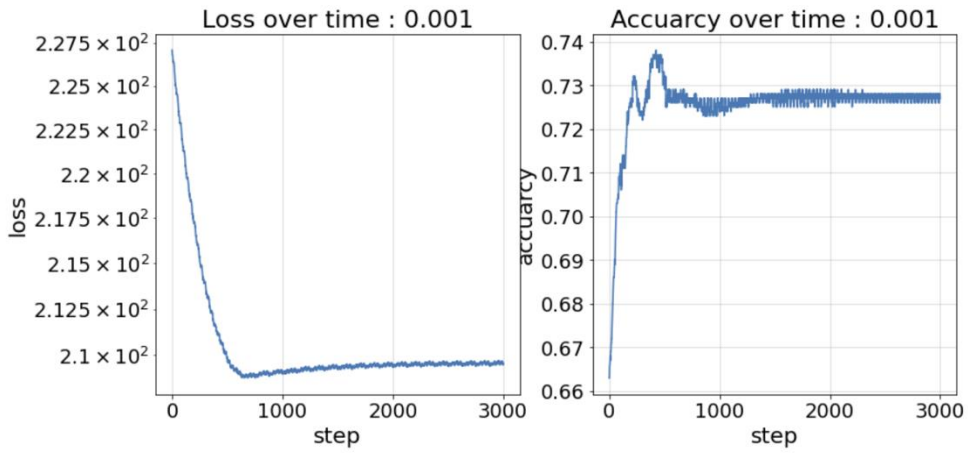
ובערכים הגדולים מאוד מהנק' הנגזרת פחות מדויקת ולכן ההפרש גדל בהתאם.

Solving Soft SVM problems using Stochastic Gradient Descent (SGD)

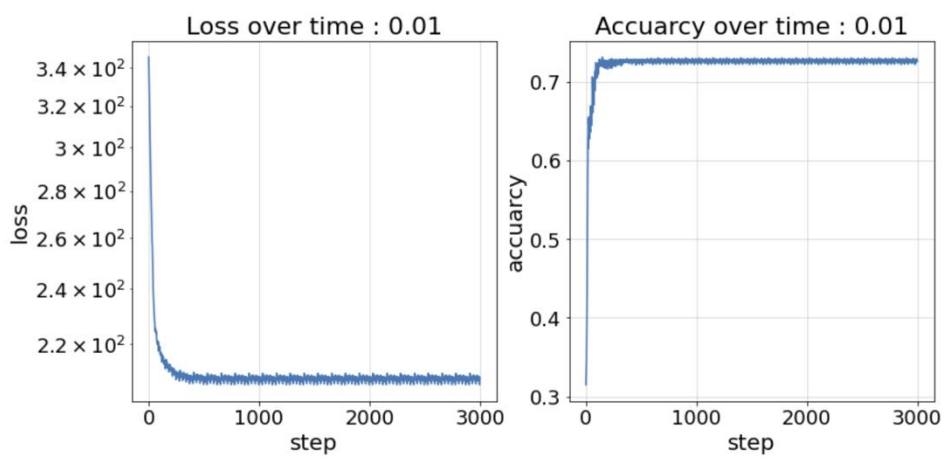
שאלה 11:

על מנת לקבל גרפים אינפורמטיביים ראינו כי עם C=0.1 אנחנו לא מקבלים תוצאות טובות ולכן לאחר בדירה של מספר ערכי C ראינו שעבור 0.3 ניתן לראות את ההבדלים בצורה טובה יותר בין קצבי הלמידה השונים.

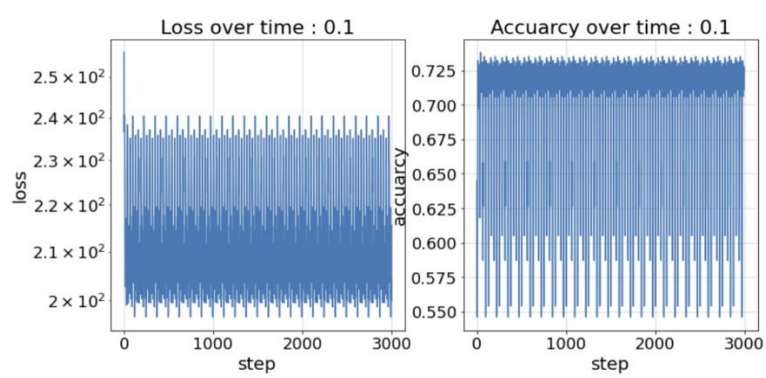
Learning Rate = 0.001



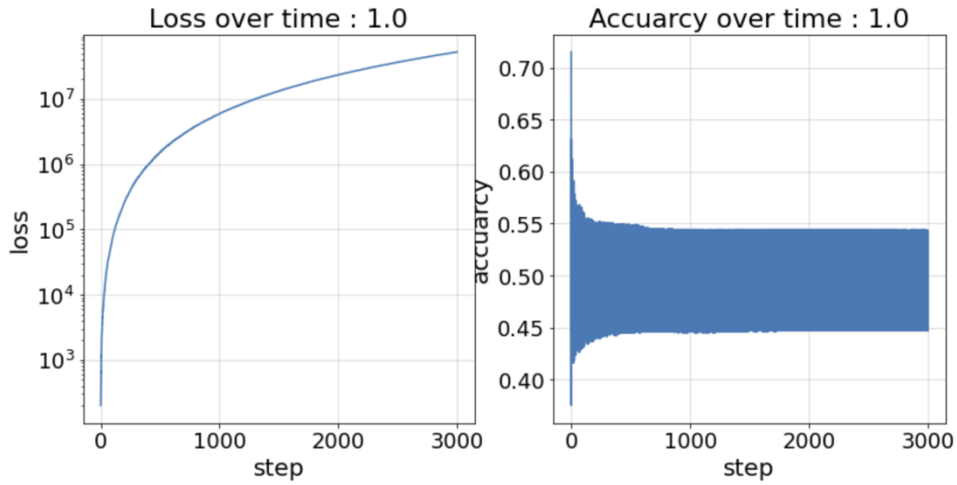
Learning Rate = 0.01



Learning Rate= 0.1



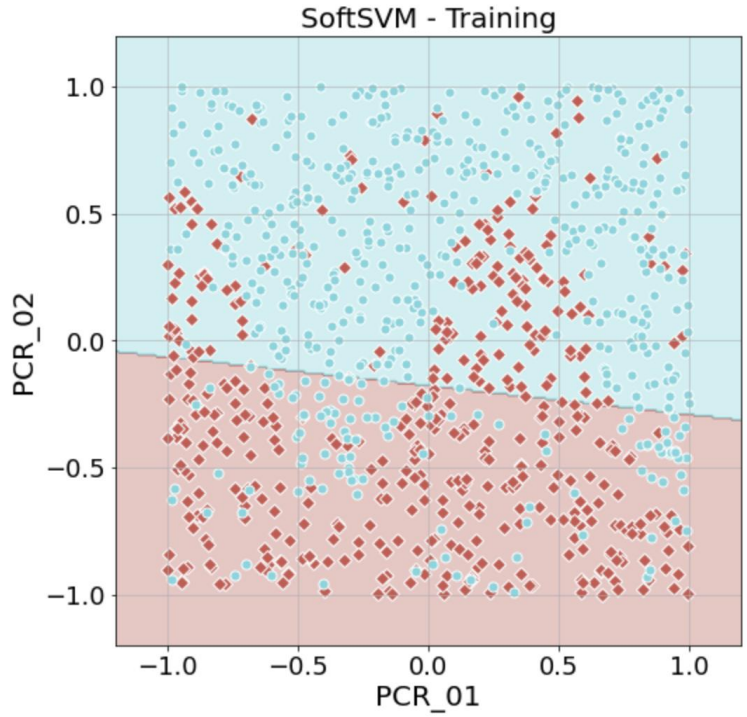
Learning Rate =1



a) נבחר קצב למידה 0.01 בכלל שניתן לראות שהוא מקבל דיוק גבוהה יחסית

(0.7 בערך) במספר נמוך של איטרציות וגם בעל השגיאה הנמוכה ביותר מכל הגרפים.

b)



c) עבור Learning rate = 0.01 הדיוק המקסימלי הוא ~ 0.75 לפי הגרף לעי"ל (0.73)

הloss המינימלי הוא ~ 200 (205).

ניתן לראות כי מס' הצעדים עד שה accuracy בערכו המקסימלי שונה במעט למספר הצעדים של ה loss המינימלי עבור learning rate הנ"ל(הloss ממשיכה להתמזער).

זאת מכיוון שלומרות ש hinge loss חוסם את loss 0/1 הוא לא חסם הדוק. ירידה ב hinge loss לא תבטיח ירידה ב0/1 וכתוצאה מכך אין הבטחה על עלייה ברמת הדיוק.

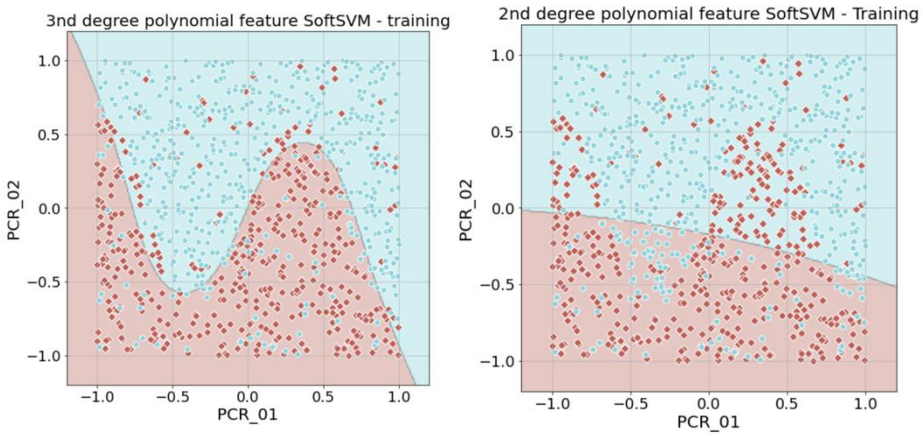
כתוצאה מכך הגרפים לא מתקבעים על ערך מקסימלי ומינימלי של דיוק וloss עבור אותו מספר צעדים.

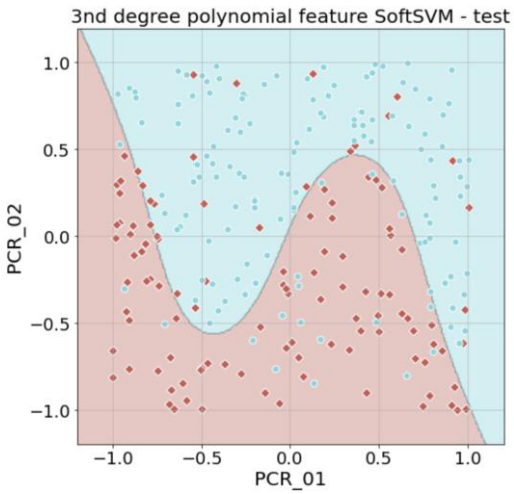
Using a feature mapping

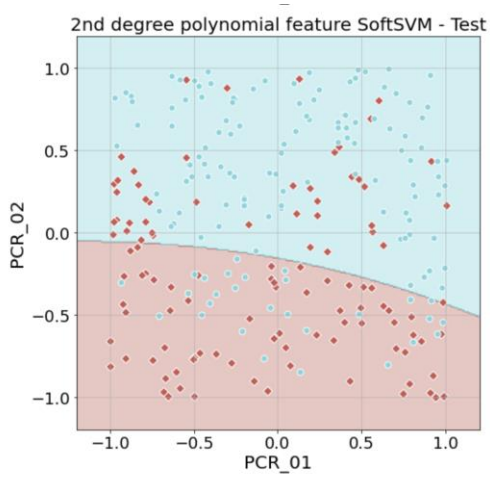
שאלה 12:

a) עבור 2nd-PolynomialFeatures שגיאת האימון היא 0.723 ושגיאת המבחן היא 0.72.
עבור 3nd-PolynomialFeatures שגיאת האימון היא 0.851 ושגיאת המבחן היא 0.832.

b)







c) ההבדל שקבלנו בין המפרידים עבור הפולינומים הוא שעבור הדאטה סט הנתון לא קיים פולינום מדרגה שנייה המסוגל להפריד אותו בצורה טובה מספיק לכן קבלנו את המפריד הטוב ביותר עבור פולינום מדרגה 2 וכפי שניתן לראות הוא אינו מפריד מספק.

המפריד שקבלנו עבור הפולינום מדרגה 3 אכן מפריד בעל דיוק גבוה ומותאם להתפלגות של הדאטה ואכן מפריד את המידע בצורה יחסית טובה (ביחס לדאטה וביחס למפרידים שראינו עד כה).

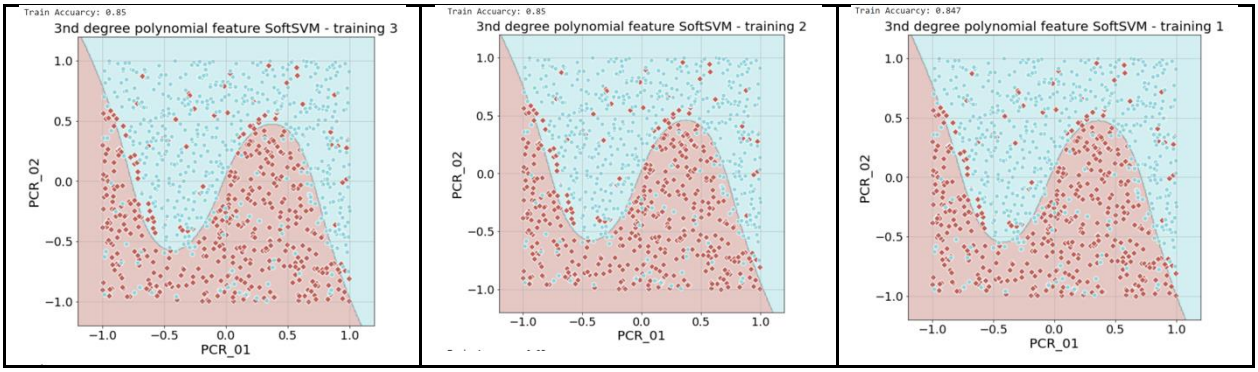
שאלה 13:

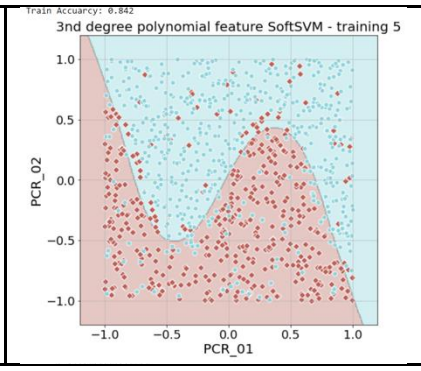
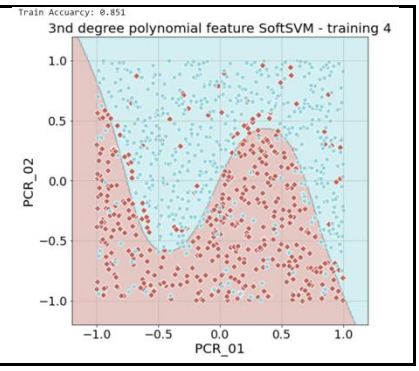
a) חמשת דיוקי המודל על סט האימון: [0.847 ,0.85 ,0.85 ,0.851 ,0.842]

הממוצע של המודלים הוא : 0.848

סטיית התקן של המודלים היא: 0.0032863353450309995

b)



mean: 0.8480000000000001 std: 0.0032863353450309995		

c) כפי שהוכחנו בעיית ה-SoftSVM היא בעיה קמורה ולכן מובטח שה-SGD יתכנס למינימום גלובאלי, ואכן כפי שניתן לראות כי סטיית התקן בין התוצאות השונות המתקבלות בכל אחת מן ההרצות נמוכה מאוד.

השונות מתקבלת בגלל שנקודת ההתחלה נבחרת באופן שרירותי (וקטור משקולות התחלתי w אקראי), האלגוריתם עוצר בנקודות קרובות מאוד למינימום ולא בהכרח במינימום (האלגוריתם עוצר אחרי מספר צעדים קבוע מראש).

Part 4: The RBF kernel

שאלה 14

נרצה להראות ש־

$$\lim_{\gamma \longrightarrow \infty} sign(\sum_{i\in[m],a_i>0} a_i y_i e^{-\gamma||x-x_i||_2^2}) = y_{i^*}, i^* = argmin||x-x_i||_2^2$$

לצורך הנוחות נסמן $d_i = ||x - x_i||_2^2$.
בנוסף מהנתון $\forall i \neq j : x_i \neq x_j$ נסיק כי לכל $i^* \neq i \in [m]$ מתקיים ש־$d_{i^*} < d_i$.
(נובע מהגדרת i^*, $i^* = argmin||x - x_i||_2^2$)

$$\lim_{\gamma \longrightarrow \infty} sign(\sum_{i\in[m],a_i>0} a_i y_i e^{-\gamma d_i}) =$$

$$\lim_{\gamma \longrightarrow \infty} sign(\sum_{i\in[m],a_i>0,i\neq i^*} a_i y_i e^{-\gamma d_i} + a_{i^*} y_{i^*} e^{-\gamma d_{i^*}}) =$$

נשתמש בכך ש־ $sign(t) = sign(\frac{t}{\alpha})$ עבור $a_{i^*} e^{-\gamma d_{i^*}} = \alpha > 0$ ונקבל :

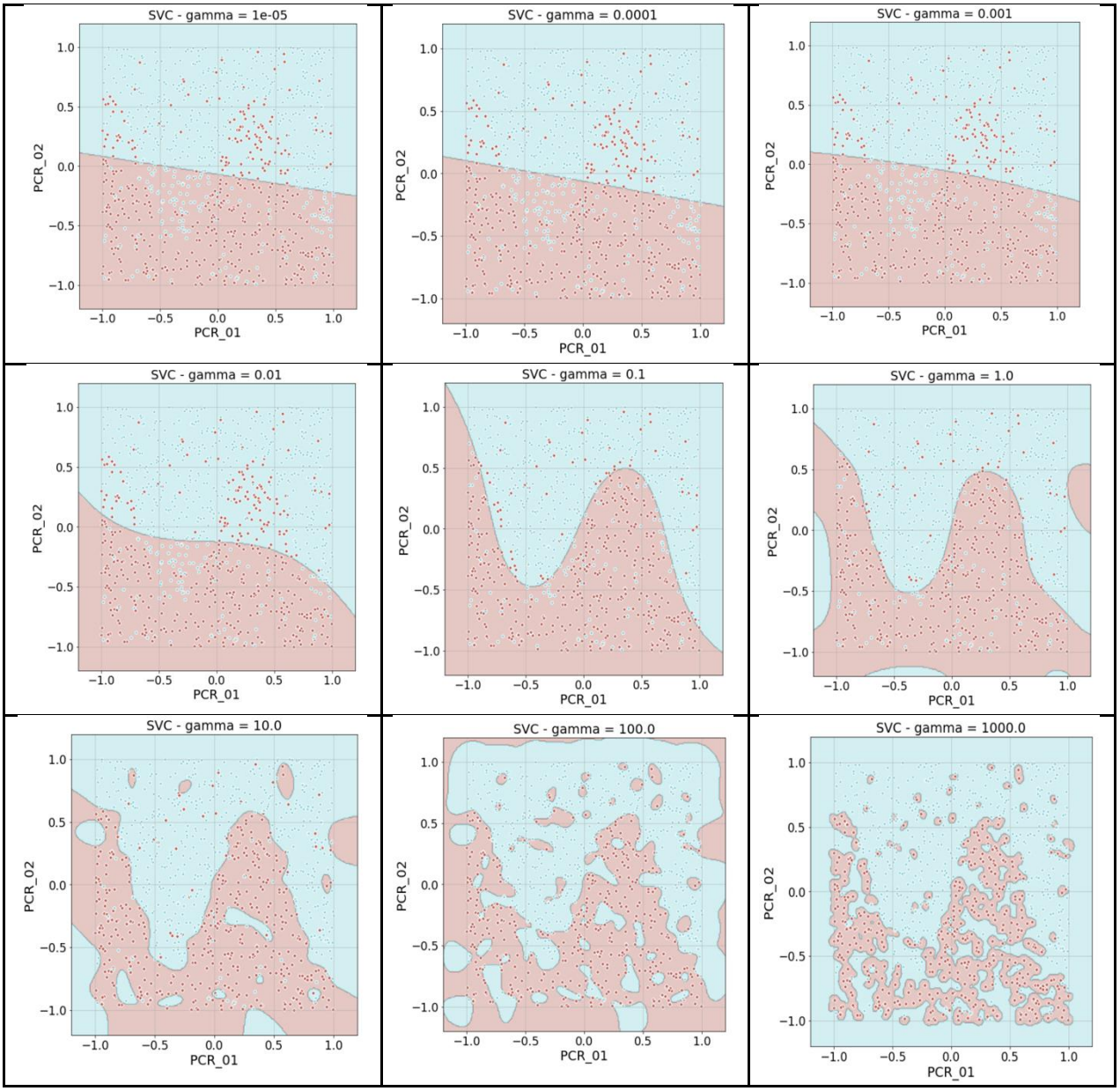
$$\lim_{\gamma \longrightarrow \infty} sign(\sum_{i\in[m],a_i>0,i\neq i^*} \frac{a_i}{a_i^*} {}_i y_i e^{-\gamma(d_i - d_{i^*})} + y_{i^*}) =$$

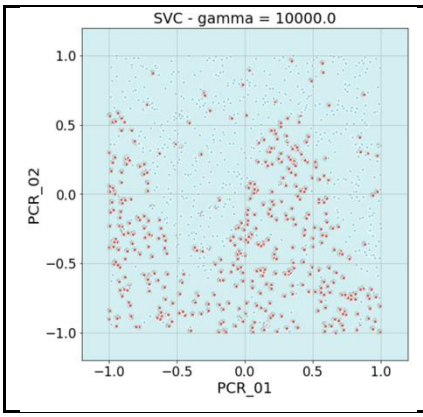
כאשר $\gamma \longrightarrow \infty$ אז $e^{\gamma(d_i - d_{i^*})} = 0$ ולכן $\lim_{\gamma \longrightarrow \infty} \sum_{i\in[m],a_i>0,i\neq i^*} \frac{a_i}{a_i^*} {}_i y_i e^{-\gamma(d_i - d_{i^*})} = 0$
ובסה"כ נקבל ש :

$$sign(y_i^*) = y_i^*$$

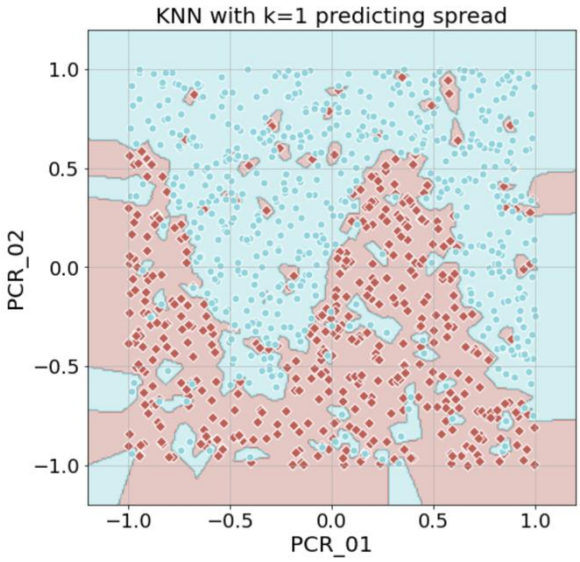
שהשוויון האחרון נובע מהגדרת y_i.

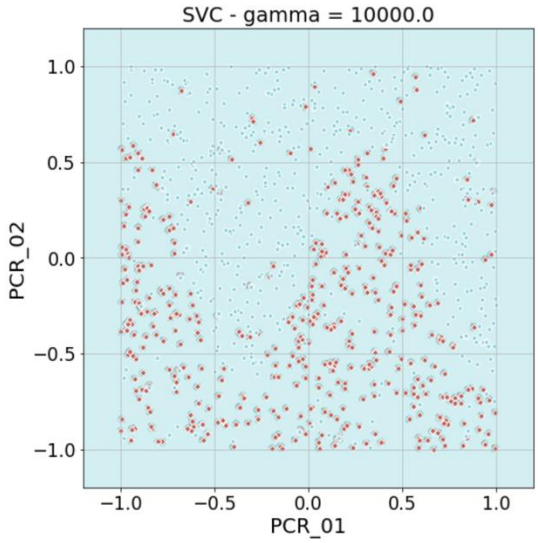
שאלה 15:


SVC - gamma = 1e-05
SVC - gamma = 0.0001
SVC - gamma = 0.001
SVC - gamma = 0.01
SVC - gamma = 0.1
SVC - gamma = 1.0
SVC - gamma = 10.0
SVC - gamma = 100.0
SVC - gamma = 1000.0
PCR_01
PCR_02



שאלה 16:





שני המודלים, גם SVC עם גאמה= 10000 וגם KNN עם k=1, מודלים מאוד דומים אך עם הבדל מסוים.

תחילה, 2 המודלים דומים בכך ששניהם מבצעים overfitting על סט האימון שלנו, בנוסף, ציפינו שהם יהיו זהים אחד לשני מכיוון שככל שגאמה גדולה יותר ישנה חשיבות לנק' קרובות יותר אליה בדומה לKNN.

אך לאחר בחינה במימוש כל אחד מהמודלים יש שוני גדול בין ה2 והוא הסיבה לשינוי הויזואליזציה והדיוק.

בRBF קרנל ככל שהגאמה גדלה הפונקציה במרחב נהיית מורכבת יותר ולכן מישור הTH במרחב חותך את הפונקצייה הגאוסיאנית קרוב לנק' האימון. בפשטות, ב SVM עם RBF קרנל ככל שהגאמה גדלה יש השפעה גדולה יותר לנק' שקרובות אחת לשנייה כפי שראינו בהרצאה. ועל כן כפי שרואים בגרף, הגאמה גדולה מאוד לכן מסווגת טוב מאוד את הנק' האימון אך מאבד את יכולת ההכללה שלו.

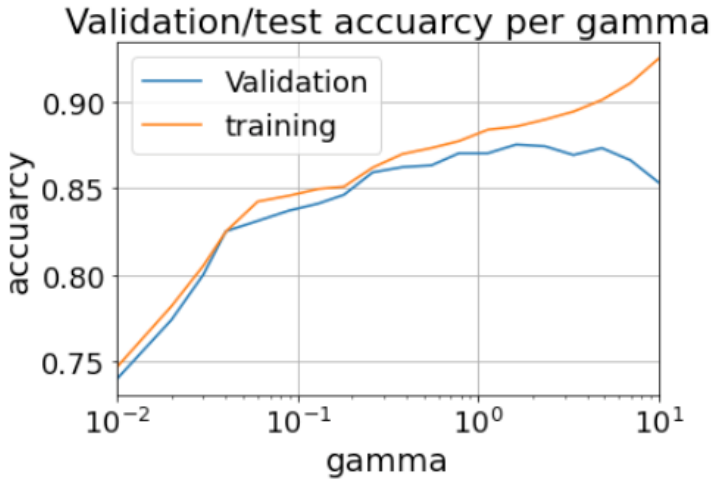
וכן ב KNN אנו מנבאים לפי המרחק במרחב. הנק' במרחק הקרובה לנקודת אימון יחידה תהיה בעלת אותו צבע של נק' האימון. וזה יוצר את ההבדל בויזואליזציות.

שאלה 17:

a) טווח הערכים שבחרנו הוא $\left[10^{-2}, 10^{1}\right]$ לאחר בדיקה של טווחים שונים של ערכי גאמה.

עבור גבול עליון גדול יותר (וגם כפי שניתן לראות בגרף) נקבל שהדיוק של סט האימון קטן וסט הוולידציה גדל לכן בהכרח לא נתעניין בערכים גדולים יותר כאלה ועבור ערכים הקטנים מהגבול התחתון נקבל שהדיוק של שני הסטים נמוך מאוד (אפילו עבור גבול תחתון של 0.1) .

b)



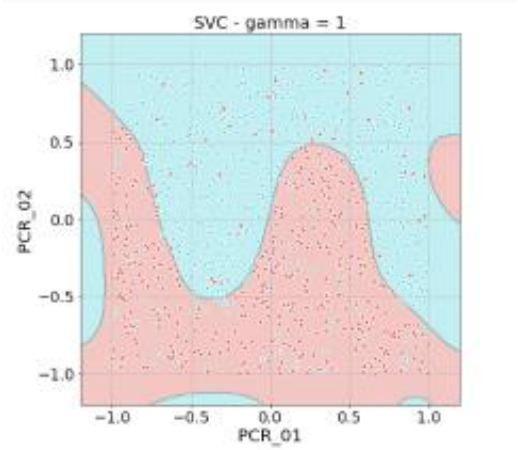
c) עבור גאמה 0.01 נקבל underfitting

d) עבור גאמה =10 נקבל overfitting

e) נשים לב שעבור ערכים קרובים וגדולים מ-10 נקבל שההפרש בין הדיוק של סט האימון והוולידציה הוא גדול ולכן במצב זה נקבל Overfiiting.

עבור ערכים הקטנים מ0.01 נקבל שהדיוק של סט האימון נמוך מאוד ולכן במצב זה נקבל Underfitting.

שאלה 18:



נשיב לב שלמודל שקבלנו יש דמיון למודל שקבלנו בשאלה מספר 4, אך הם גם שונים במספר דברים. המפריד שקבלנו ב- RBF חלק יותר לעומת המפריד שקבלנו ב-kNN, דבר שנובע מאופן פעולת האלגוריתם, המפריד ב-kNN נקבע לפי הסיווג של כל נקודה בהתאם למרחקה מהשכנים הקרובים אליה ולכן אין צפייה לחלקות. ב-RBF המפריד שקבלנו הינו החתך האוקלידי של פונ' הגאוסיין במימד גדול יותר, וכידוע זוהי פונקצייה חלקה.

נצפה שבמודל ה-RBF נקבל דיוק גבוהה יותר כי בנוסף להתחשבות במרחקים של הנקודות יש גם התחשבות למשקלים של כל נקודה.

קבלנו שהדיוק של המודל היא 0.868 (עבור בחירת גאמה להיות 1).

נזכר שבשאלה 4 עבור kNN עם k האופטימלי קבלנו דיוק של 0.856 ולכן נתן להסיק כי המודל RBF עדיף עבור המשימה של ניבוי ה-spread.