Major HW2 – Report

Yaniv Valdman – 318932365

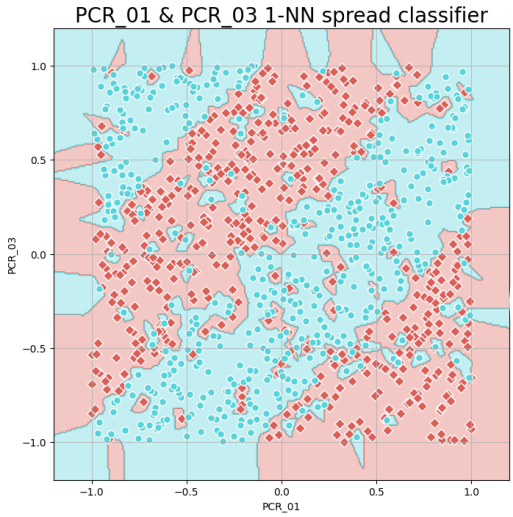
Itamar Friedman – 315173344

תוכן

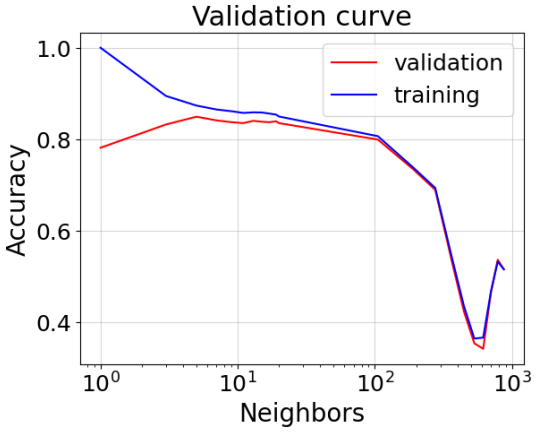
חלק 1

שאלה 1

להלן תוצאת מודל 1-NN בהסתמך על הפיצ'רים PCR_01&PCR_03:



שאלה 2



להערכתנו, $k = 5$ הוא ה-k האופטימלי. עבורו, הדיוק הממוצע על קבוצות האימון הוא 0.87 ועל קבוצת הוולידציה הוא 0.85 (התוצאות מעוגלות לשתי ספרות אחרי הנקודה העשרונית). זוהי הערכתנו כיוון שעבור ערך k זה, הדיוק הממוצע על קבוצות הוולידציה הוא הגבוה ביותר, ראינו בהרצאה שבוחרים את הערך הטוב ביותר בהסתמך על מידע זה.

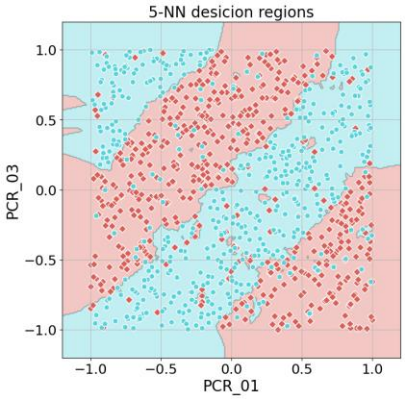
overfit מתרחש כאשר המודל מתבסס באופן מוגזם על ה-*training data* ולכן ביצועיו לא טובים על ה-*test set / validation set*. ניתן לראות תופעה זאת עבור ערכי k נמוכים כגון $k = 1,2$, ערכים אלו מתאפיינים בדיוק גבוה על ה-*train set* אך ביצועים גרועים על קבוצת הוולידציה. תופעה זו התרחשה כיוון שעבור ערכים אלו, הסיווג הותאם באופן מוחלט לקבוצת האימון, אך אינו מייצג באופן מהימן את ההתפלגות ממנה הגיעו הנתונים.

Underfit מתרחש כאשר המודל הינו בעל ביצועים לא טובים על קבוצת האימון, למשל כיוון שהמודל פשטני מדי ולא תופס את הניואנסים הקיימים במידע. ערכי k גבוהים (כגון $k = 530, 515$) גורמים ל-*underfit* כיוון שהמודל מתייחס לאזורים גלובאליים וגדולים מדי, מבלי לתפוס את הדקויות של תחומים מסוימים וקטנים יותר. כמו כן, בערכי k גדולים במיוחד, אשר מתקרבים למספר הדגימות בקבוצת האימון (800 במקרה זה בשל ה-*folding*), למעשה ההחלטה מתקבלת ע"י סיווג לפי הקבוצה הגדולה יותר, כי כל הנקודות במידע הן השכנות הקרובות ביותר של שאר הנקודות, מה שגורם לפשטנות יתר של המודל.

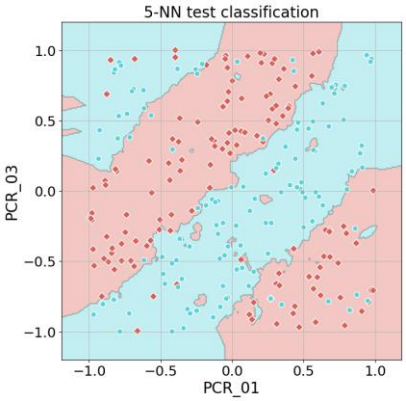
שאלה 3

דיוק המודל הנבחר על קבוצת ה *test*, *5-NN*, הינו 0.86

פיזור ה *train* לפיהן נוצרו אזורי ההחלטה:

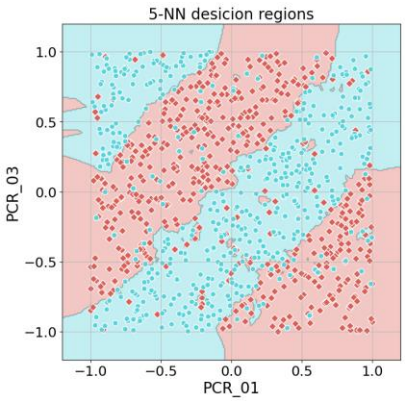


פיזור ה *test* על מפת אזורי ההחלטה:

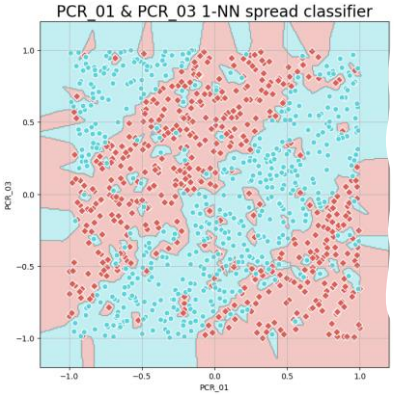


שאלה 4

אזורי החלטה של *5-NN*:



אזורי החלטה של *1-NN*:



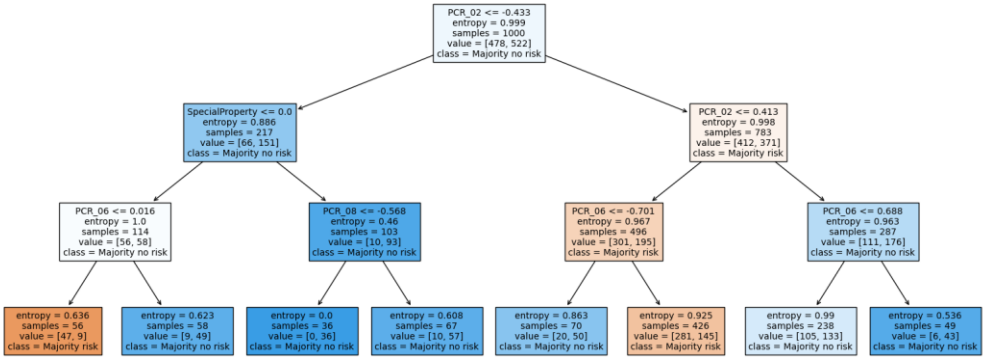
נתבונן באזורי ההחלטה שהתקבלו בשאלות 1 ו-3 בזה ליד זה. ניתן לראות שעבור חמישה שכנים, מתקבלים אזורי החלטה הרבה יותר "חלקים" שפחות מושפעים מרעש. מודל שמתחשב רק בשכן אחד הוא overfitted לtraining set, לכן, במודל 1-NN ניתן לראות תופעה של "איים" רבים מאוד, זהו אחד מהאופנים שה-overfit בא לידי ביטוי בהם במודל זה.

חלק 2

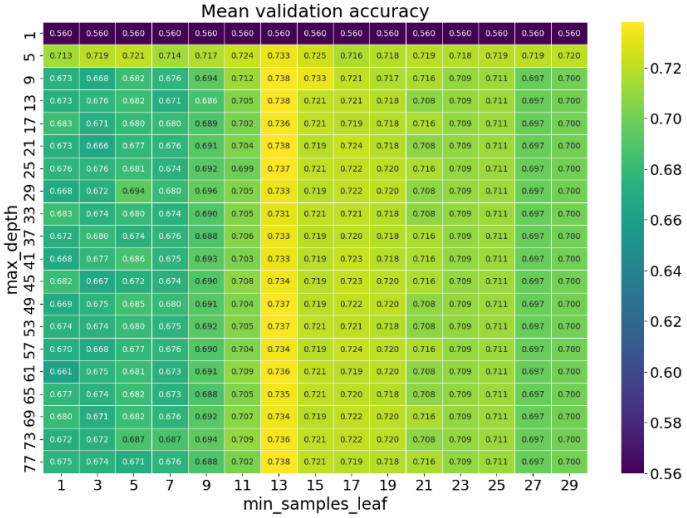
שאלה 5

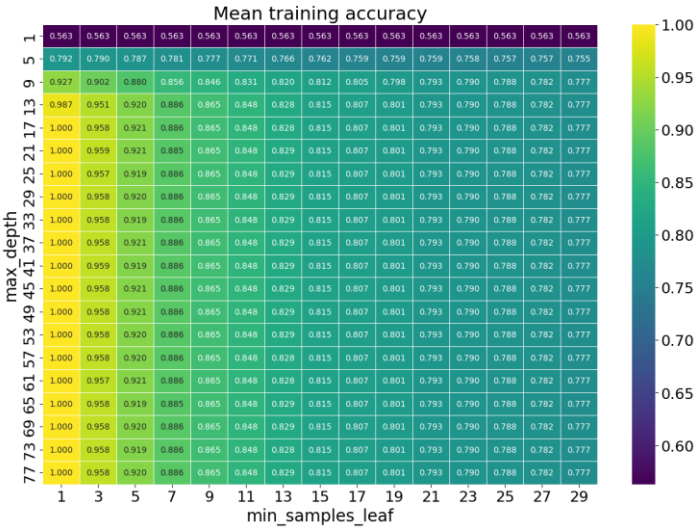
הדיוק על ה-train set של העץ בעל עומק 3 הינו 0.696 .

Risk Decision Tree



שאלה 6





הקומבינציה האופטימלית שמצאנו, עבורה ה-validation accuracy מגיע לערך המקסימלי של שהוא 0.738, הינה $max\_depth = 21, min\_samples\_leaf = 13$.

קומבינציה שגורמת ל-underfitting הינה $max\_depth = 1, min\_samples\_leaf = 1$ (ולמעשה כל ערך $min\_samples\_leaf$ כאשר העומק מוגבל ל-1). זהו מצב של underfitting כיוון שהתוצאות עבור הזוג הזה על קבוצת *האימון* לא טובות, מה שנובע מכך שהמודל פשטני מדי בשל עומק עץ שלא מאפשר למידה משמעותית (כלומר, אין "פיצ'ר קסם" בודד שעל פיו בלבד ניתן לנבא את ה-*risk*).

דוגמה לקומבינציה שגורמת ל-*overfitting* הינה $max\_depth = 61, min\_samples\_leaf = 1$. זהו מצב של *overfitting* כיוון שעל קבוצת האימון התוצאה מושלמת (1.0) ואילו על קבוצת הוולידציה התוצאה לא טובה, *0.661* בקירוב. כאשר נשתמש בערך $min\_samples\_leaf = 1$ אנו מאפשרים עלים שהם ספציפיים לדגימה בודדת מהדאטא. דגימה בודדת יכולה להיות רעש, או לא לשקף את התפלגות הדאטא, ולכן לא נרצה ליצור סיווג בעץ על סמך נקודה בודדת. דבר זה פוגע במודל כיוון שהוא נותן חשיבות רבה ומוטעית עבור רעשים בדאטא, מה שעלול לסווג נקודות אחרות בעלות פיצ'רים דומים (עד אותו פיצול בעץ) באופן שגוי.

[YV1] עם הערות: אנו מאפשרים לדגימה בודדת מהדאטא להיות עלה ובכך ליצור סיווג בעץ אולי ניסוח יותר טוב

הגדלת ה-*min_sample_leaf* "מחליקה" את המודל, כלומר מתעלמת מרעשים מה שעוזר להימנע מ-*overfitting*.

שאלה 7

ראשית, נציין כי בחרנו לבחון 20 עומקים שונים בטווח [1,80], ו-15 כמויות שונות של דגימות מינימליות בעלים בטווח [1,30]. סה"כ נוצר לנו (מהמכפלה הקרטזית) גריד בגודל 20X15, כלומר 300 תאים, ובהתאם המודל בחן 300 קומבינציות שונות של היפר-פרמטרים.

אילו רצינו לבחון היפר-פרמטר שלישי, היינו מתמודדים עם גריד תלת מימדי שמספר התאים בו הוא כגודל המכפלה הקרטזית בין 3 הקבוצות שהיינו מגדירים. כלומר, בהנחה ששתי הקבוצות הראשונות שלנו הן בגודל 20 ו-15, והשלישית בגודל n, היו בידינו $300n$ שלשות שונות של היפר-פרמטרים.

כל היפר-פרמטר נוסף יוסיף קבוצה חדשה למכפלה הקרטזית, ובהתאם יגרור הכפלה של מספר הבדיקות הקודמות בגודל הקבוצה החדשה. כמו כן, כל הוספה תוסיף מימד לגריד שנוצר.

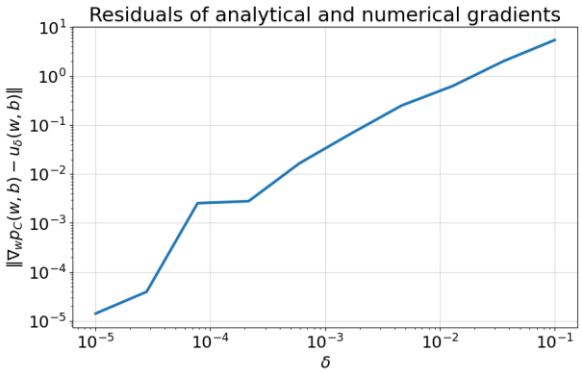
שאלה 8

הדיוק של מודל העץ ID3 כאשר העומק מוגבל ל-21 ומספר הדגימות המינימלי לעלה הוא 13 הניב דיוק של **0.736 על ה-test set**.

נבחין כי זוהי תוצאה דומה לזו שהתקבלה על קבוצת הוולידציה עבור אותו מודל.

חלק 3

שאלה 9



מהגרף נראה כי ככל ש-δ קטנה, ההפרש בין החישוב הנומרי והאנליטי קטן גם הוא. החישוב האנליטי של הגרדיאנט אינו תלוי ב-δ, כלומר, בהינתן $p_c(w,b)$ עם w ו-b קבועים, חישוב הגרדיאנט באופן אנליטי הינו דטרמיניסטי ואותה תוצאה תתקבל בכל הרצה (עד כדי שגיאות נומריות של *floats*). לעומת זאת, חישוב הנגזרת הנומרי מושפע למדי מערכה של δ, ככל ש-δ קטנה, אנו בוחנים שתי נקודות קרובות יותר בפונקציה (בכיוון w_i מסוים), לכן, חישוב השיפוע בהסתמך על שתי נקודות אלו מדויק יותר מאשר אותו חישוב כאשר δ גדולה.

עם זאת, שיפוע ההפרש אינו קבוע, לפרקים אף המגמה משתנה והגרף יורד כאשר δ גדלה (בחרנו גרף "יפה" שלא מראה זאת), זה נובע מכך שהנקודות המופיעות בגרף הינן הממוצע של 100 דגימות הפרש לכל ערך δ, ובכל דגימה עבור δ מסוימת מוגרלים b ו $w-$. צמד *(w,b)* הם למעשה נקודה על גרף p_c. ייתכן שבסביבת נקודה *(w,b)* מסויימת יש שינוי חד במיוחד בשיפוע הגרף, ולכן עבור נקודה זו תדרש δ קטנה יותר על מנת להגיע לקירוב טוב של הגרדיאנט לעומת נקודות אחרות על הגרף.

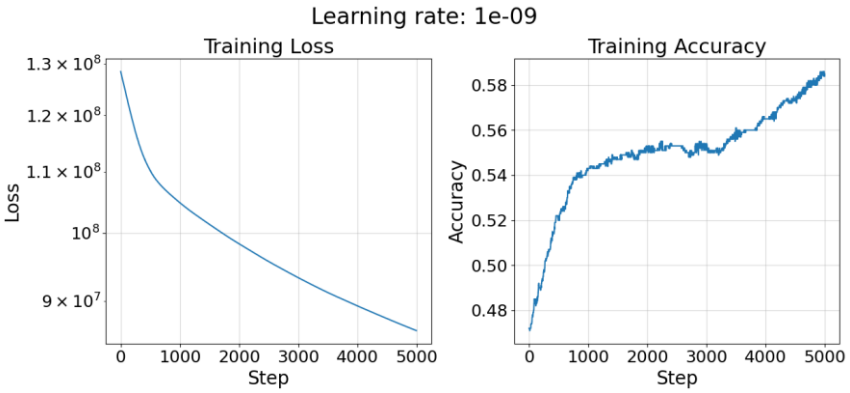
שאלה 10

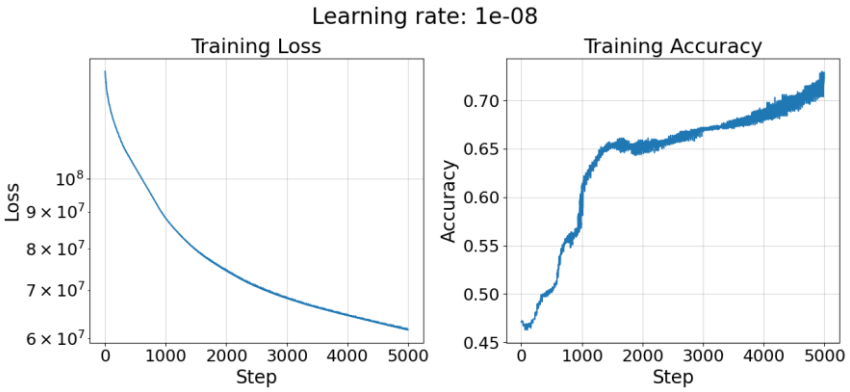
פעולת אימון (fit) על דאטא מבצעת אלגוריתם SGD, במהלכו, מנסים להביא למינימום את שיעור ה-loss function, שבמקרה שלנו הוא $p_c(w,b)$. כיוון שפונקציה זו הינה קמורה (ראינו זאת בהרצאה), המינימום הלוקאלי שלה מתלכד עם המינימום הגלובאלי. אי לכך, העובדה שבגרף המצורף לתרגיל שיעור ה-loss הולך ופוחת במגמת ירידה הוא הגיוני.

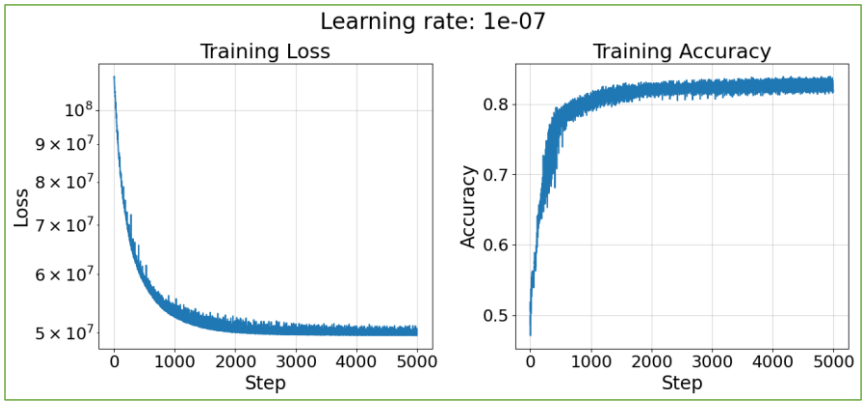
עם זאת, מאחר שהמידע לא פריד לינארית, הפסד מינימאלי לא מבטיח דיוק מינימאלי, וזאת כיוון שעבור מידע שכזה מפריד לינארי לא מתאים (כלומר, באופן אינהרנטי קיים bias עבור המודל, הכוונה היא לא ל- b ההיפר-פרמטר).

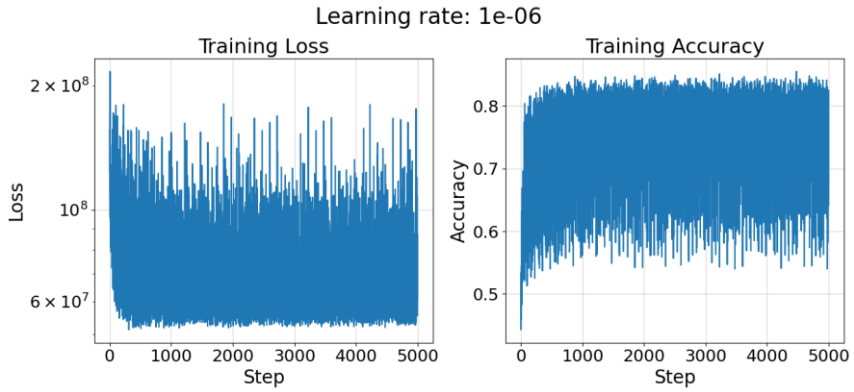
משום שאנחנו מנסים לבצע סיווג לינארי של דאטא שאינו פריד לינארית, ניתן לצפות שאלגוריתם SGD יגיע למינימום loss עם ערך loss יחסית גבוה, ובהתאם לכך, גם אחוז שגיאה גבוה. אכן ניתן לראות בשני הגרפים התכנסות למינימום של LOSS בקירוב על 1e14 עם שגיאה של 50%. כמו כן, ניתן לומר בקירוב שכמעט כל קו מפריד שנעביר ייתן דיוק של כ50% (לדוגמה ממבט על סיווג כל הנקודות בצבע אחד או העברת קו אקראי דרך ראשית הצירים).

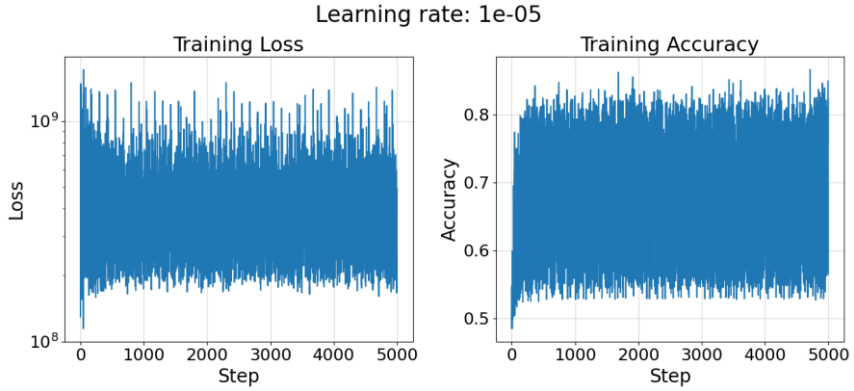
שאלה 11





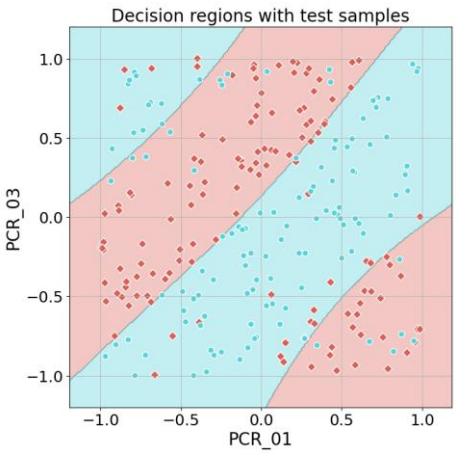




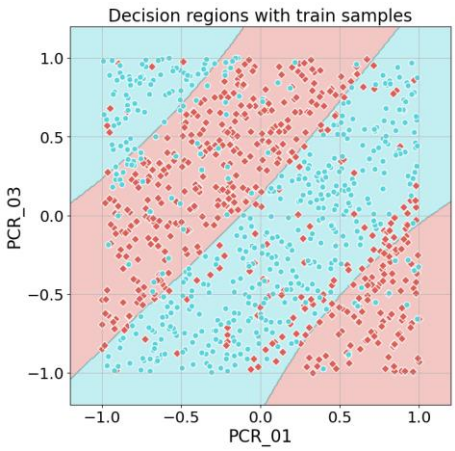


היינו בוחרים ב-learning rate = 1e-7 אשר מגיע למינימום הכי מהר (במספר הצעדים הקטן ביותר) וכן מגיע ל-training accuracy הגבוה ביותר.

שאלה 12



Test accuracy: 0.824



Train accuracy: 0.811

חלק 4

שאלה 13

סעיף א' - הוכחה:

$$\lim_{\gamma\to 0} sign\left(\sum_{i\in[m],\alpha_i>0} \alpha_i y_i \exp\left\{-\gamma\left||x-x_i|\right|_2^2\right\}\right) = (assumption\ i)$$

$$= sign\left(\lim_{\gamma\to 0}\sum_{i\in[m],\alpha_i>0} \alpha_i y_i \exp\left\{-\gamma\left||x-x_i|\right|_2^2\right\}\right) = (limit\ rule\ of\ sums)$$

$$= sign\left(\sum_{i\in[m],\alpha_i>0} \alpha_i y_i \lim_{\gamma\to 0}\exp\left\{-\gamma\left||x-x_i|\right|_2^2\right\}\right) = (assumption\ ii)$$

$$= sign\left(\sum_{i\in[m],\alpha_i>0} \alpha_i y_i\right)$$

$$= sign\left(\sum_{i\in[m],\alpha_i>0|y_i=+1} \alpha_i y_i + \sum_{i\in[m],\alpha_i>0|y_i=-1} \alpha_i y_i\right)$$

$$= sign\left(\sum_{i\in[m],\alpha_i>0|y_i=+1} \alpha_i - \sum_{i\in[m],\alpha_i>0|y_i=-1} \alpha_i\right) = (assumption\ iii\ \&\ *)$$

$$= \begin{cases} +1, & \sum_{i\in[m],\alpha_i>0|y_i=+1} \alpha_i \geq \sum_{i\in[m],\alpha_i>0|y_i=-1} \alpha_i \\ -1, & \sum_{i\in[m],\alpha_i>0|y_i=+1} \alpha_i < \sum_{i\in[m],\alpha_i>0|y_i=-1} \alpha_i \end{cases}$$

$$= argmax_{y\in\{-1,1\}}\left(\sum_{i\in[m]|y_i=y} \alpha_i\right)$$

* The norm of α is bounded, therefore :

$$\exists c_2 < \infty:\ \|\alpha\|_2 \leq c_2$$

$$\Rightarrow \sqrt{\alpha_1^2 + \cdots + \alpha_m^2} \leq c_2$$

$$\Rightarrow \alpha_1^2 + \cdots + \alpha_m^2 \leq c_2^2 =: c_3 < \infty$$

$$\Rightarrow \begin{matrix} 1.\ \forall i \in [m]: \alpha_i < c_3 < \infty \\ 2.\ \sum \alpha_i < c_3 < \infty \end{matrix}$$

סעיף ב'

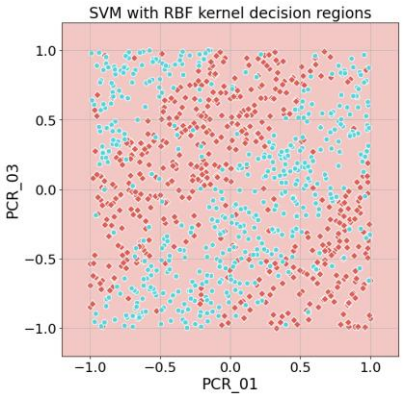
על סמך ההוכחה בסעיף א', אם כל כניסות הווקטור הינן 1, אז תוצאת הסיווג (שקול לתוצאת פונקציית ה-sign) הינה:

$$argmax_{y\in\{-1,1\}}\left(\sum_{i\in[m]|y_i=y} 1\right)$$

כלומר, הלייבל שיש ממנו יותר דגימות בדאטא.

הדבר דומה למודל *K-NN* כאשר *K=m* (מספר הדגימות) שבו החלטה על סיווג נקודה חדשה מתקבלת על סמך הלייבל שממנו הכי הרבה דגימות במידע (כיוון שכל הנקודות הינן שכנות של כל הנקודות). עם זאת, בשל אופי חישוב הסיווג, המודלים לא זהים.

שאלה 14

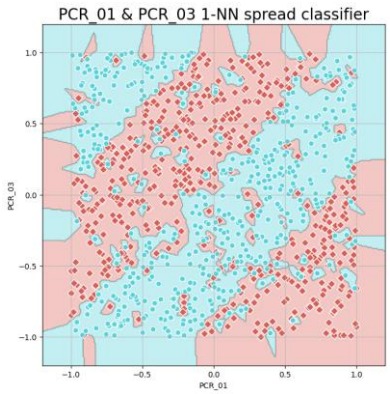


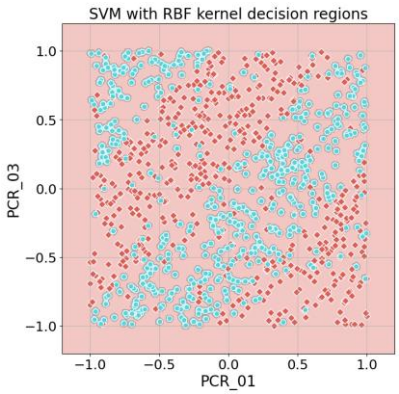
אכן, ה-decision regions שהתקבלו על ידי מסווג זה תואמים לכלל ההחלטה שדנו בו בשאלה 13b. בשאלה זו, דנו במצב בו $\gamma \to 0$ ולכל $i \in [m]$ בפתרון הדואלי לבעיה α מתקיים $\alpha_i = 1$. הראינו שהסיווג עבור כל נקודה יתבצע על ידי השמת הלייבל בעל מספר המופעים הגדול ביותר ב-train set.

אכן, במסווג שהגדרנו, בחרנו $\gamma = 1e-7$ שמדמה בקירוב טוב מצב בו $\gamma \to 0$, וכמו כן, בבחינת ה-members של המחלקה, אם נתבונן באיבר **dual_coef** , נבחין כי הוא מחזיק סדרה בגודל 1000 (מספר הדגימות) של ± 1, מהגדרת איבר זה במחלקה, הוא מחזיק את כניסות וקטור הפתרון הדואלי, כאשר הן נכפלות בלייבל המתאים לכל דגימה, כיוון שהלייבלים הינם ± 1 בלבד וכל כניסות α חיוביות, ניתן להסיק כי כל הכניסות הללו הן $+1$. אלו הם תנאי סעיף *13b* בדיוק ולכן ההתנהגות המתקבלת היא מתן הלייבל בעל המופע המקסימלי לכל הדגימות ב-*train set*.

שאלה 15

נצרף פעם נוספת את הדיאגרמה מ-Q1 לשם נוחות.



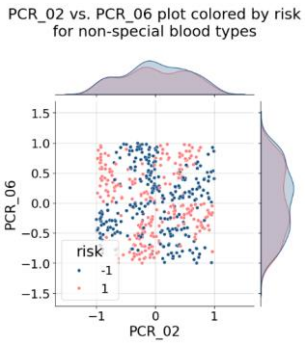


כפי שניתן לראות, בשתי הדיאגרמות ישנם "איים" בהם מתקבלת החלטה על סמך דוגמא בודדת (ניזכר כי במודל 1-NN כל החלטה מתקבלת על ידי שכן בודד קרוב ביותר וכל נקודה נחשבת שכן של עצמה), זהו מצב של overfitting. בדיאגרמה המתאימה ל-SVM עם RBF kernel כאשר $\gamma = 5000$ מתרחש מצב דומה, ערך γ מאוד גדול הינו פרופורציונלי ביחס הפוך ל-$\frac{1}{2\sigma^2}$, כלומר, ערך γ גדול מתאים לערך σ קטן. כפי שראינו במאמר המצורף, ככל שמקטינים את σ, קטן גם ה- region of similarity, משמעות הדבר הוא שנקודה חדשה "דומה" לנקודה קיימת רק אם הן קרובות במיוחד. דבר זה מסביר את ה"איים" הכחולים.

הבדל משמעותי בין שתי הדיאגרמות בא לידי ביטוי בשטחים שלא קרובים באופן משמעותי לאף נקודה. שטחים אלה מהווים במודל ה-*SVM* עם *RBF kernel* את ה-*region of dissimilarity* (נזכיר, σ שבו אנו משתמשים קטן כעת) והם מקבלים סיווג אדום וזאת בשל הזנב של פונקציית ה-*kernel*. העובדה שהסיווג של אזורים אלו הוא אדום נובעת מכך שישנם יותר דוגמאות עם סיווג המתאים לצבע זה ב-*train set*. הגרף של סכימת הגאוסיאנים האדומים יכסה את זה של הכחולים, ובכך, על אף שבאזורים הרחוקים לפי המאמר - "לא מתקבלת החלטה", בפועל הערך אינו אפס ובשל כיסוי הגרף האדום את הכחול ההחלטה שתתקבל תהיה לייבל אדום.

חלק 5

שאלה 16

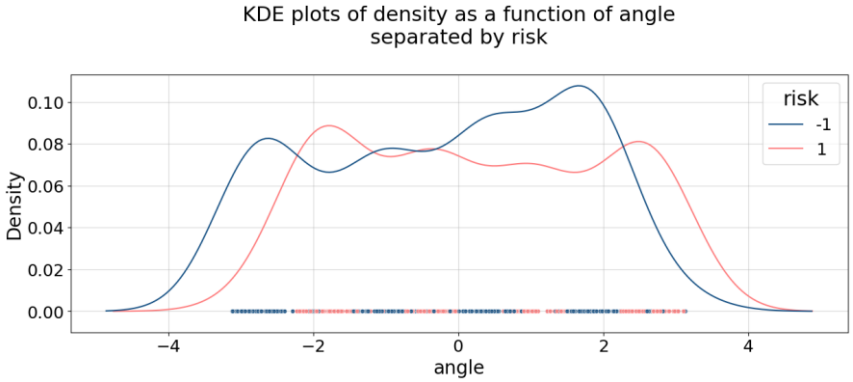


את התרשים לעיל ניתן לתאר איכותית באופן הבא –

נמקם ראשית הצירים של מערכת פולרית בנקודה (0,0). כעת, קיימים תחומים של θ עבורם לכל $0 \le r \le \sqrt{2}$ (כך שקיימות נק' בדאטה עבור הזוג (r,θ)) מרבית הנקודות הינן מאותו סוג, כחולות או ורודות, כלומר, בסיכון או לא בסיכון.

בהרחבה, עבור התחומים, $0 \le \theta \le \frac{\pi}{4}, \frac{\pi}{2} \le \theta \le \frac{3\pi}{4}, \pi \le \theta \le \frac{5\pi}{4}, \frac{3\pi}{2} \le \theta \le \frac{7\pi}{4}$, הרוב המוחלט של הנדגמים אינו בסיכון (הנקודות בעלות *risk = -1*). בשאר תחומי θ מרבית הנדגמים בסיכון.

שאלה 17



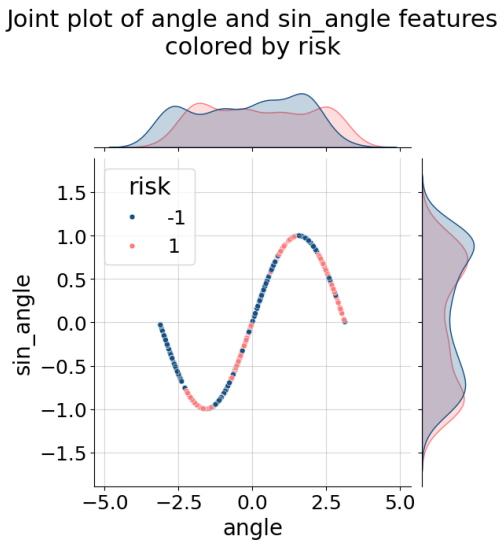
ניתן לראות הפרדה בציר האופקי (הזווית) לאינטרוולים, קיימות זוויות בהן מרבית תווית ה-risk

הינן 1- וכן זוויות בהן המרבית הינה 1.

הנקודה האחרונה באה לידי ביטוי בשני אופנים, הראשון הוא שבתחומים בהם מרבית הנקודות בעלות risk=-1, גרף ה-KDE המסומן בכחול גבוה יותר מאשר הוורוד, וגם להיפך (באזורים בעלי risk=1 הגרף הוורוד בעל ערך גדול מהכחול). האופן השני שבו הנקודה באה לידי ביטוי הוא שבאותם אזורים, גרף הפיזור (לאחר שכל הנקודות הוטלו לציר האופקי) נראה כחול או ורוד כמעט בבלעדיות, דבר זה נובע מכך שדחיסות הנקודות מצבע מסוים גדולה יותר מאשר של הצבע השני, ובכך הן "מסתירות" את הנקודות האחרות.

עם זאת, לאחר מיפוי זה **המידע עדיין איננו פריד לינארית.**

שאלה 18



עבור ערך בטא 1, הפיצ'רים הללו לא נראים פרידים לינראית.

שאלה 19

רעיון:

נבצע grid search עבור שני היפר-פרמטרים –

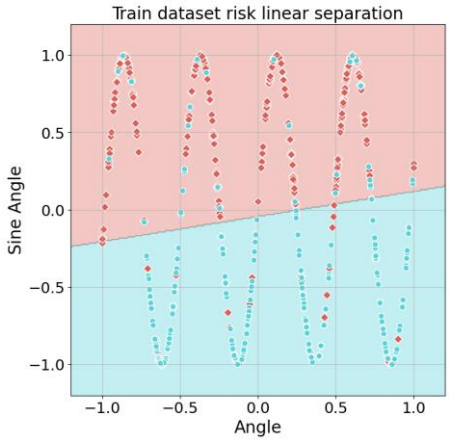
1. C של soft SVM (מקדם סך ההפסד) בעל kernel לינארי, שכן מיד נגדיר טרנספורמציה בעצמנו.
2. β, מידת המתיחה בציר האנכי של הפונקציה sin.

כיוון ש-β איננו ארגומנט של מחלקת SVC, לא נוכל לבצע את פונקציית grid search של sklearn בצורה פשוטה, נגדיר Pipeline לשם כך. ה-pipeline יקבל dataframe בו מופיע הייצוג הפולארי של דגימות המידע (למעשה רק את הזווית כיוון שרק היא משפיעה על הסיווג), יבצע את פונקציית הסינוס עם פרמטר β בטווח הערכים [3.0,5.0] עם קפיצות של 0.1 ינרמל את ערכי הסינוס באמצעות min max scaler ולבסוף יתאים ישר לנקודות לאחר הטרנספורמציה באמצעות *SVM* עם פרמטר *C* בטווח הערכים [0.4, 15].

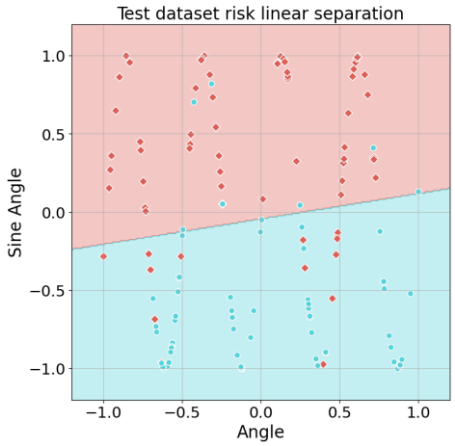
נציין כי לשם בחירת טווחי הערכים ביצענו שילוב של הגיון מתמטי וחיפושים חוזרים למציאת ה-*sweet-spot*. בהרחבה, כיוון שהבחנו ב-8 תחומי זוויות שונים (4 לכל קבוצה), חיפשנו פרמטר β (אשר מבצע כיווץ של הציר האופקי) באופן כזה שהמידע יופרד לשניים – מעל הציר האופקי ומתחתיו. עבור פרמטר *C*, התחלנו בבחינת ערכי *C* שונים והבחנו שהחל מערך *C=13* אין שיפור בביצועים, להערכתנו, דבר זה מתרחש בשל פרידות לינארית טובה יחסית של המידע, מה שמעניש באופן חמור יותר על חריגה. כמו כן, נציין כי לכל ערך $0 < C \leq 1$ (עד כדי קפיצות של *0.1*) מתקבל דיוק פחות טוב מאשר ערכי *C* הגדולים מאחת.

ביצוע התהליך הנ"ל מוצא לנו באמצעות חיפוש גריד את ההיפר-פרמטרים הבאים:

$$\boldsymbol{\beta = 4.1}\,, \qquad \boldsymbol{C = 13}$$



Train Score: 0.8689788053949904



Test Score: 0.8421052631578947

נשווה את הנתונים למודל הלינארי שהיה נתון:

	מודל נתון	מודל חדש	השוואה

Training set	58%	86.9%	שיפור משמעותי באחוזי ההצלחה (עליה בכ30 אחוזים)
Test set	53%	84.2%	גם כן שיפור (עליה בכ*30* אחוזים)

כלומר, המודל החדש שמצאנו יותר טוב באופן משמעותי.