# Theoretical Part:

## 1.1) Hard&Soft - SVM:

צריך להוכיח שבעיית האופטימיזציה של Hard-SVM היא בעיית
תכנות ריבועית:

$$\underset{(w,b)}{\arg\min}\|w\|^2 \quad s.t \quad \forall i\; y_i(\langle w,x_i\rangle+b)\ge 1$$

כלומר עלינו למצוא מטריצות $Q, A$ ווקטורים $a, d$ כך ש:

$$\arg\min \tfrac{1}{2}v^TQv+a^Tv \quad s.t \quad Av\le d$$

נשתמש בסימון: $v:=\begin{bmatrix}w\\b\end{bmatrix}$ ונסמן: $\|w\|^2=w^TIw$

אנו רוצים למצוא:

$$\arg\min\|w\|^2=\arg\min w^TIw \overset{\text{רצוי}}{=} \tfrac{1}{2}v^TQv+a^Tv$$

נשים לב שאנו רוצים לתת משמעות לb בכינוי וכן נרצה לתת
משמעות לw ולכן מאחר ו $v=\begin{bmatrix}w\\b\end{bmatrix}$, $Q$ תבקש להיות המטריצה:

$Q:=\begin{bmatrix}I_d&0\\0&0\end{bmatrix}$ בלוקים ⇐ מאחר וכשנעבור את כפל המטריצות b לא ישפיע. כי
יוכפל ב-0 ונקבל $v^TQv=\|w\|^2$ כנדרש.

בנוסף נוכל לבחור $a:=0_v$ מאחר שבביטוי הנורמה אין צורך
לינארי ונקבל בסך הכל:

$$\arg\min\|w\|^2=\arg\min\tfrac{1}{2}v^TQv+a^Tv$$

לא משפיע על
התוצאה ומקל
על הגזירה בהמשך.

נמשיך בלרשום את אילוצי הבעיה בצורה אחרת (מטריציונית):

$$\forall i\; y_i(\langle w, x_i\rangle + b) \geq 1$$

$$\Updownarrow$$

$$\forall i\; y_i x_i^T w + y_i b \geq 1$$

$$\Updownarrow$$

$$\forall i\; -(y_i x_i^T w + y_i b) \leq -1 \iff \forall i\; -y_i (x_i^T w + b) \leq -1$$

כעת נסדר את הכל בצורה מטריציונית ונגלה את $A, d$:

$$-y_i(x_i^T w + b) \iff -y_i \begin{bmatrix} x_i^T & 1 \end{bmatrix} \begin{bmatrix} w \\ b \end{bmatrix} = A_i \underbrace{V}_{\text{נסמן}} \leq -1$$

כלומר כל שורה $i$ במטריצה $A$ תהיה: $A_i = -y_i \cdot [x_i^T \; 1]$

ו $d_i = -1$ לכל הקורדינטות כלומר:

$$A = \begin{bmatrix} -y_1 \cdot [x_1^T \; 1] \\ -y_2 \cdot [x_2^T \; 1] \\ \vdots \\ -y_{d+1} [x_{d+1}^T \; 1] \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)} \quad \text{ו} \quad d = \begin{pmatrix} -1 \\ \vdots \\ -1 \end{pmatrix} \in \mathbb{R}^{d+1}$$

ומצאנו בכך את כל המרכיבים של התכנון הריבועי כנדרש ☑

# 1.2) PAC Learnability:

1)a) תהי D התפלגות מעל X , $S_m$ סט דגימות מגודל m
(דגימות בלתי תלויות) מעל X . צ"ל שלכל $X' \subseteq X$ כך ש
$\mathbb{P}_D[x \in X'] > \varepsilon$ מתקיים $\mathbb{P}_D[S_m \cap X' = \emptyset] \le e^{-m\varepsilon}$

הוכחה:

לדגימה בודדת: $\circledast$ $\mathbb{P}_D[x \in X'] > \varepsilon \Rightarrow \mathbb{P}_D[x \notin X'] < 1-\varepsilon$

שיקולי הסתברות

מאחר והדגימות בלתי תלויות:

$$\mathbb{P}_D[\forall x_i \in S_m \notin X'] = \prod_{i=1}^{m} \mathbb{P}_D[x_i \notin X'] \underset{\circledast}{\le} (1-\varepsilon)^m \le e^{-m\varepsilon}$$

א"ש ברנולי

$$\Updownarrow$$

$$\mathbb{P}[S_m \cap X' = \emptyset]$$

כנדרש $\blacksquare$

b) נניח S סט אימון עם m דגימות רנדומלי
נניח שאנו במקרה הריאליזבילי ולכן $\hat{f}_S$ שנבחרה ע"י
A לא טועה בכלל על סט האימון, כלומר: $L_{D,f}(\hat{f}_S, S) = 0$
עלינו לחסום את: $\mathbb{P}_D[L_{D,f}(\hat{f}_S) > \varepsilon]$

נשתמש ברמז הראשון ונגדיר סט $Z_{\hat{f}_S}$ כסט דגימות שבו
$\hat{f}_S$ טועה. כלומר, $Z_{\hat{f}_S} := \{x \in X : \hat{f}_S(x) \neq f(x)\}$, כעת מאחר ו $\hat{f}_S$
לא טועה על סט האימון נקבל שאין בין שני הסטים דגימות
חופפות כי אם היו אז שהפונקציה טועה על סט האימון וא

שהיא לא שנפה וזו הדגימה לא הייתה ב $Z_{\hat{f}_S}$. כלומר מתקיים:

$$Z_{\hat{f}_S} \cap S = \phi$$

נביע את הביטוי התחלתי בצורה אחרת:

הגדרת $Z_{\hat{f}_S}$ - מכיל את כל הטעויות

$$\mathbb{P}_D[L_{D,f}(\hat{f}_S) > \varepsilon] \overset{(=)}{} \mathbb{P}_D[\hat{f}_S(x) \neq f(x) > \varepsilon] = \mathbb{P}[D(Z_{\hat{f}_S}) > \varepsilon]$$

כלומר, הפונקציה טועה בהסתברות גדולה מ $\varepsilon$ אמ"מ ההסתברות לקבל דגימה שהיא תטעה בה גדולה מ $\varepsilon$.

כעת מאחר ו $Z_{\hat{f}_S} \cap S = \phi$ נוכל להשתמש באי השיוויון מהסעיף הקודם ולקבל:

$$\mathbb{P}_D[L_{D,f}(\hat{f}_S) > \varepsilon] = \mathbb{P}[D(Z_{\hat{f}_S}) > \varepsilon] \underset{\text{הסיכוי לקבל}}{=} \mathbb{P}_D[Z_{\hat{f}_S} \cap S = \phi] \underset{\text{סעיף קודם}}{\leq} e^{-m\varepsilon}$$

דגימה שבה המודל טועה גדולה מ $\varepsilon$ אמ"מ כל הטעויות וכל האימון אינם בעלי איברים משותפים.

c)

בסעיף b הראנו שלכל פונקציה ממחלקת ההיפותזות מתקיים:

$$\mathbb{P}[L_{D,f}(\hat{f}_S) > \varepsilon] \leq e^{-m\varepsilon}$$

כעת, מ 1.1 Theorem אנו יודעים שמחלקת היפותזות סופית היא למידה PAC ולכן במקרה שלנו נסמן H מחלקת ההיפותזות.

מכאן האיחוד מתקיים:

$$\mathbb{P}_D[\underbrace{\exists_{h\in H} : L_D(h) > \varepsilon}_{\text{קיימת היפותזה "רעה"}}) \le \underbrace{|H|}_{\text{כמות ההיפותזות}} \cdot \underbrace{e^{-m\varepsilon}}_{\text{הסיכוי של כל היפותזה להיות עם שגיאה שגדולה מ } \varepsilon}$$

כעת, מאחר והאלגוריתם שלנו מחזיר שגיאה אפסית על הסט האימון הסיכוי שטעינו ובחר היפותזה שגויה חסום גם הוא על ידי אותו חסם איחוד מאחר והמאורע הזה כלול במקרה הכללי ונקבל:

$$\mathbb{P}_D\left(L_{D,f}(A(S_m)) > \varepsilon\right) \le |H| \cdot e^{-m\varepsilon}$$

כעת נגדיר $\delta$: $|H| \cdot e^{-m\varepsilon} \le \delta$

ונקבל:

$$\mathbb{P}_D\left(L_{D,f}(A(S_m)) > \varepsilon\right) \le \delta$$

$$\Big\Updownarrow \text{ שיקולי הסתברות}$$

$$\boxed{\mathbb{P}_D\left(L_{D,f}(A(S_m)) < \varepsilon\right) \ge 1-\delta}$$ כנדרש

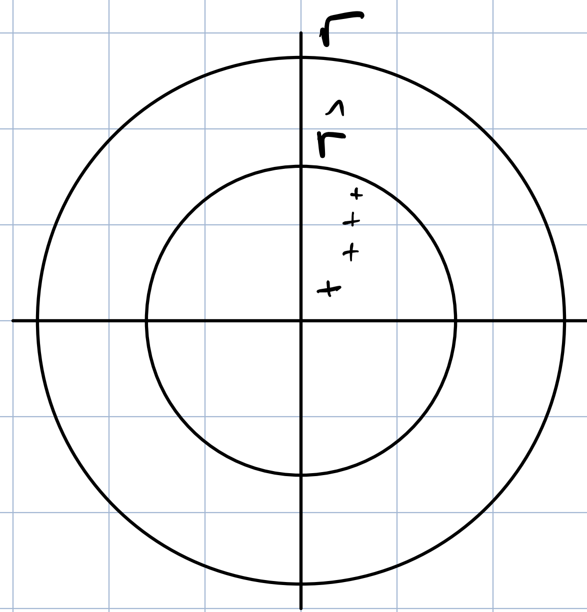
2. נפתור שאלה זו בדומה לדוגמת האינטרוולים שפתרנו בתרגול. כדי להראות ש-$H$ למידה PAC עלינו להראות אלגוריתם אשר מקיים: $\mathbb{P}_D\left(L_{D,f}(A(S_m)) < \varepsilon\right) \ge 1-\delta$.

האלגוריתם $A$ שנציע הוא: בחירת הרדיוס הקטן ביותר הכולל בתוכו את כל הדגימות שהחזירו 1 (חיוביות) בסט האימון.

כלומר, $\hat{r} = \max\{\|x_i\|_2 : y_i = 1\}$, נסמן $r =$ הרדיוס האמיתי הנכון

כעת נראה שאלגוריתם זה מקיים את הנדרש:

תחילה נבחין כי $\hat{r} \subseteq r$ מתקיים
תמיד כי אחרת מאופן פעולת האלגוריתם
הייתה דגימה חיובית באיזור שמחוץ מ $r$
בסתירה להגדרת $r$. ולכן שגיאה יכולה
לקרות רק כאשר נקבל דגימה חיובית
שנמצאת ברדיוס בין $\hat{r}$ ל $r$. ולא הייתה
בסט האימון.




כלומר: $L_D(\hat{r}) = \mathbb{P}_{x \sim D}[\hat{r} < \|x\|_2 \le r]$, כעת נחסום את
פונקציית ה Loss הנ"ל. יהי $\varepsilon$, נרצה להראות כי:

$$\mathbb{P}_D(L_D(\hat{r}) < \varepsilon) > 1-\delta$$

נשים לב קודם כל שחייבת להיות מסה עם רדיוס בין $r$ ל $\hat{r}$
והסתברות $< \varepsilon$ לקבל דגימה מהמסה הזאת כי אם לא היה
אז המסה האמיתית הייתה שווה לשלנו והשגיאה הייתה 0
ולכן הביטוי היה נכון באופן ריק.

במקרה שהמסה בניהם אכן קיימת:
נשתמש באי השוויון מהסעיף הקודם, ההסתברות לקבל דגימה מהמסה
בין שני הרדיוסים $< \varepsilon$ ולכן חסומה על ידי: $(1-\varepsilon)^m$
כלומר:

$$\mathbb{P}(L_D(\hat{r}) > \varepsilon) < (1-\varepsilon)^m \le e^{-m\varepsilon}$$

הוכחנו בסעיף קודם.

כעת נקבע את $\delta$ להיות: $e^{-\varepsilon m} \le \delta$

ונקבל ע"י הוצאת ln:

$$-\varepsilon m \le \ln\delta$$

$$m \ge \frac{\ln\left(\frac{1}{\delta}\right)}{\varepsilon}$$

כלומר בסך הכל קיבלנו:

$$\mathbb{P}_D\left(L_{D,f}\left(A(S_m)\right) > \varepsilon\right) \le \delta$$

שיקולי הסתברות

$$\mathbb{P}_D\left(L_{D,f}\left(A(S_m)\right) < \varepsilon\right) > 1-\delta$$

כלומר H למידה PAC כאשר מספר הדגימות הדרוש חסום ע"י: $\frac{\ln\left(\frac{1}{\delta}\right)}{\varepsilon}$ כנדרש ∎

3)

נניח: H היא uniform convergence כלומר:

$$D^m\left(\{S\in (X\times Y)^m \mid \forall h\in H \ \ |L_S(h)-L_D(h)| < \varepsilon\}\right) \ge 1-\delta$$

נרצה להראות שH למידה APAC, כלומר צ"ל:

$$L_D(A(S)) \le \min_{h'\in H} L_D(h') + \varepsilon$$

נסמן, $h^* = \min_{h\in H} L_D(h)$ - שגיאה אמיתית מינימלית

$h_S = \min_{h\in H} L_S(h) \underset{\uparrow \text{ERM}}{=} A(S)$ - שגיאה אמפירית מינימלית

נסתכל על $\frac{\varepsilon}{2}$, מתכונות uniform convergence אז נקבל שמתקיים:

$$\forall h\in H \quad |L_S(h)-L_D(h)| \le \frac{\varepsilon}{2} \quad (*)$$

וכעת נקבל:

$$L_D(A(S)) \underset{ERM}{=} L_D(h_S) \overset{(*)}{\le} L_S(h_S)+\frac{\varepsilon}{2} \underset{h_S \text{ הוא המינימלי}}{\le} L_S(h^*)+\frac{\varepsilon}{2} \overset{(*)}{\le} L_D(h^*)+\frac{\varepsilon}{2}+\frac{\varepsilon}{2} = L_D(h^*)+\varepsilon$$

שלמנו בדיוק תכונות APAC כנדרש ◻

נשים לב שמספר הדגימות הדרוש הוא מספר הדגימות הדרוש להתכנסות uniform convergence כלומר סיבוכיות מספר הדגימות חסומה ע"י $m_H(\varepsilon,\delta) \le m^{UC}(\frac{\varepsilon}{2},\delta)$

## 1.3: VC-Dimension

1) נתון, $X=\{0,1\}^n$ , $Y=\{0,1\}$ , $h_I(x) = \left(\sum_{i\in I} x_i\right) \bmod 2$

כלומר כל פונקציה $h_i$ סופרת את 1 באינדקסים $i\in I$ בדגימות ומסווגת 1 או 0 בהתאם לתוצאה של הסכום mod 2. מאחר והאיברים בדגימות הם 1,0 נקבל ש $h_i$ בודקת ומחזירה 1 אם מספר ה-1-ים באינדקסים $i\in I$ אי זוגי ו-0 אם הוא זוגי. כעת נרצה לחשב את ה VC-dim של מחלקת ההיפותזות הנ"ל.

נשתמש ברמז: גודל מחלקת ההיפותזות הוא $2^n$ - מספר תתי הקבוצות ב-$[n]$ בתרגול ראינו שלמחלקת היפותזות סופית מתקיים:

$VCdim(H) \le \log_2|H| = \log_2 2^n$ , כלומר $\boxed{VCdim(H)\le n}$

כעת נראה שאפשר לנתץ דגימה מגודל n: ניקח את הדגימה הבסיסית, כלומר: $x_1=\begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix}$, $x_2=\begin{pmatrix}0\\1\\0\\\vdots\\0\end{pmatrix}$, $x_3=\begin{pmatrix}0\\0\\1\\0\\\vdots\\0\end{pmatrix}$ ........ מגודל n.

כעת נשים לב שניתן לייצר כל תווית $y \in \{0,1\}^n$ על ידי בחירת תת הקבוצה של x עם האינדקסים בהם יש 1 התואמים ל-x-ים הבסיס. כלומר: $I=\{j: y_j=1\}$ וכך נוכל לייצר כל תווית שנרצה, כלומר ניתצנו דגימה מגודל n ולכן מתקיים $\boxed{VCdim(H)\geq n}$

וסך הכל נקבל ש $\boxed{VCdim(H)=n}$ כנדרש ☑

2) נתון $H_1 \subseteq H_2$ כלומר כל פונקציה ב-$H_1$ כלולה גם ב-$H_2$. VC dim בעצם אומר מה גודל הדגימה המקסימלי שמחלקת היפותזות יכולה לנתץ, כלומר להשיג כל תווית אפשרית לדגימה באמצעות הפונקציות שלה, מאחר וכל הפונקציות של $H_1$ נמצאות גם ב-$H_2$ היא תוכל לנתץ לכל הפחות את אותה דגימה מקסימלית שניתצה $H_1$ ואולי אף דגימה יותר גדולה אם יהיו בה את הפונקציות המתאימות, אך במילים אחרות:

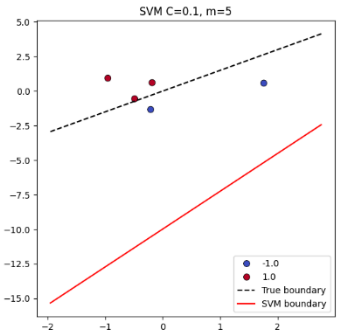
$$\boxed{VCdim(H_2)\geq VCdim(H_1)}$$

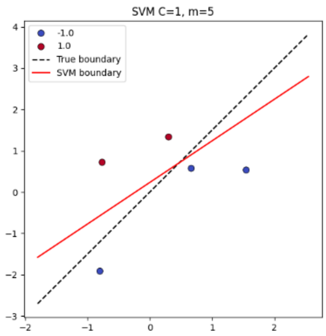
# Practical Part:

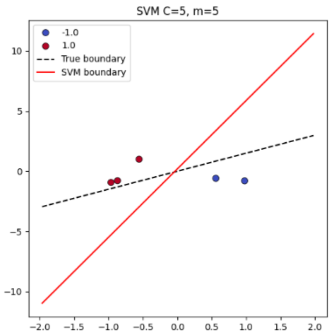
## SVM:

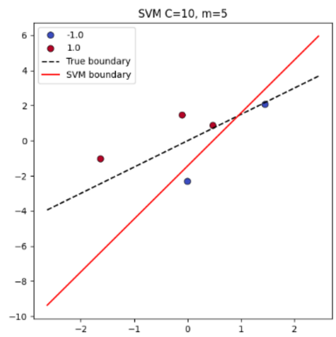
1)

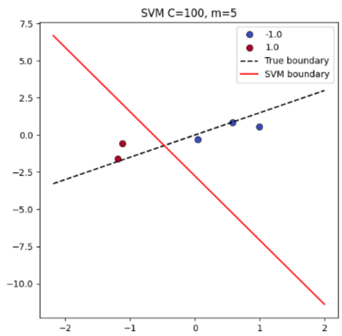
M = 5:



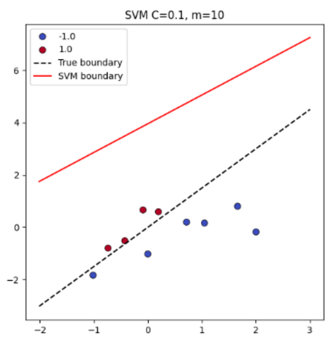



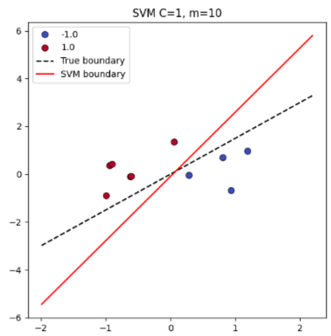



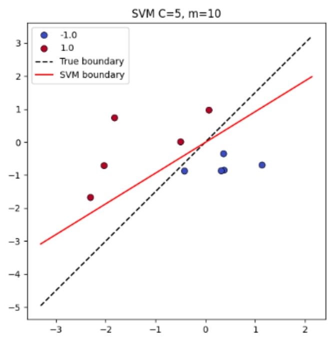



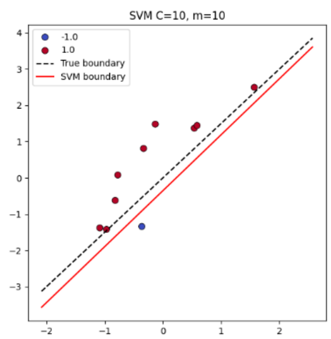



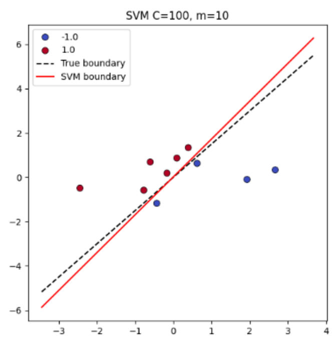

M = 10:



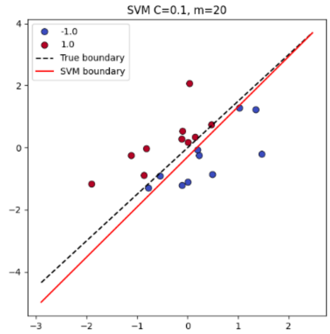



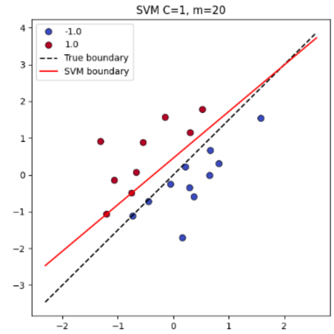



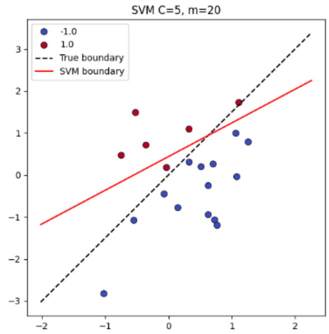



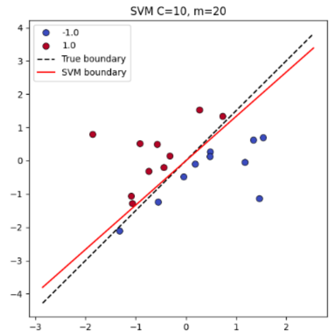



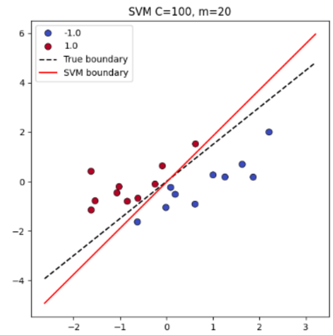
M = 20:


SVM C=0.1, m=20
-1.0
1.0
True boundary
SVM boundary

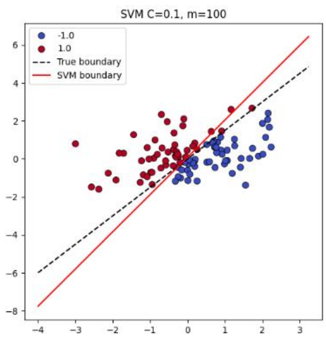

SVM C=1, m=20
-1.0
1.0
True boundary
SVM boundary

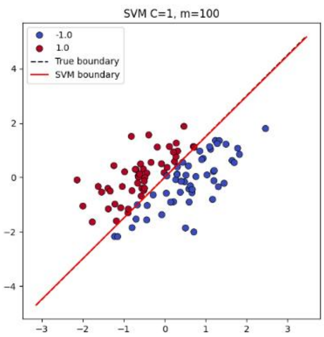

SVM C=5, m=20
-1.0
1.0
True boundary
SVM boundary

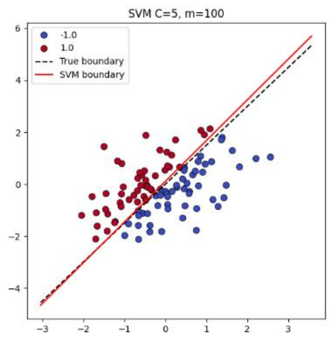

SVM C=10, m=20
-1.0
1.0
True boundary
SVM boundary

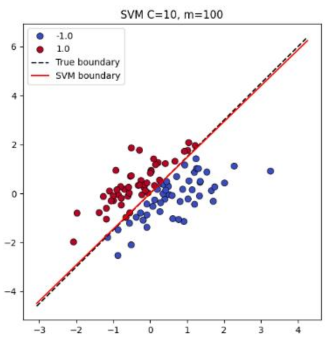

SVM C=100, m=20
-1.0
1.0
True boundary
SVM boundary

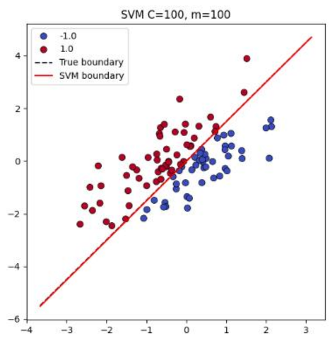
M = 100:


SVM C=0.1, m=100
-1.0
1.0
True boundary
SVM boundary


SVM C=1, m=100
-1.0
1.0
True boundary
SVM boundary


SVM C=5, m=100
-1.0
1.0
True boundary
SVM boundary


SVM C=10, m=100
-1.0
1.0
True boundary
SVM boundary

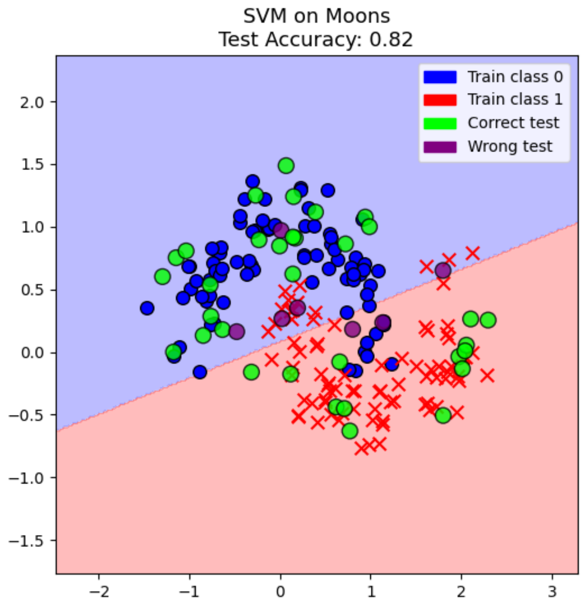

SVM C=100, m=100
-1.0
1.0
True boundary
SVM boundary

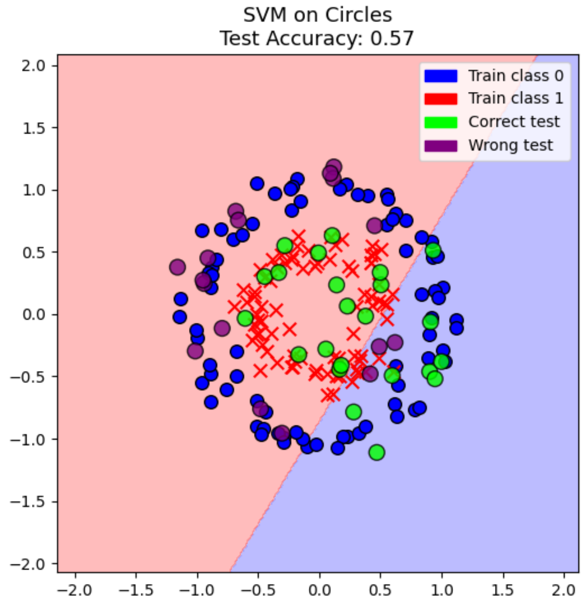
2) בSVM ערך הC מייצג כמה "עונש" אנו נותנים על כל טעות של המודל. כלומר ככל שהC יותר קטן המודל מאפשר יותר טעויות על מנת ליצור את המרווח הגדול ביותר. זה תואם את הגרפים שקיבלנו. ניתן לראות שככל שהC יותר גדול הקו עושה פחות טעויות בסיווג ומעביר קו עם מרווח קטן יותר מהדגימה הקרובה ביותר אליו. ובהתאמה ככל שהC קטן יותר הסיווג עושה יותר טעויות אך ממקסם את המרווח.

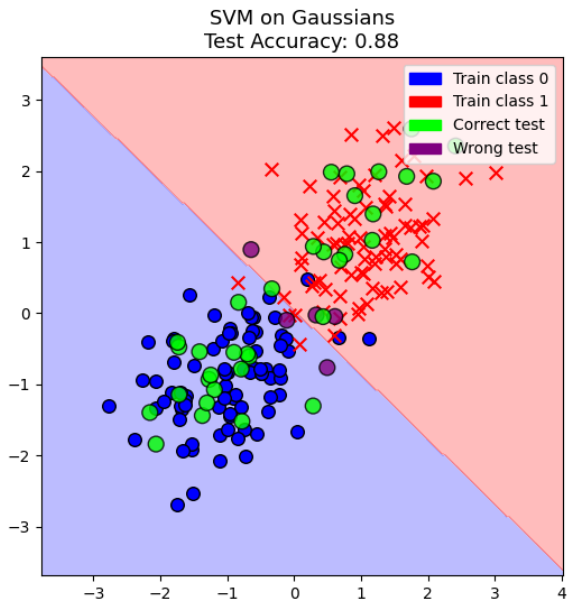
3) לפי עקרונות למידת PAC ככל שיש יותר דגימות כלומר M גדל מוביל לכך שאפסילון (מרווח הטעות) יקטן בהסתברות גבוהה יותר. זה מתבטא באמצעות סיבוכיות הדגימה שאומרת כמה דגימות צריך על מנת להבטיח שגיאה קטנה מאפסילון. ואכן נשים לב מהגרפים שקיבלנו שככל שM קטן יותר מודל הSVM יוצר סיווג פחות מדויק עם רעש שככל הנראה יחזיר טעויות על העולם האמיתי בהסתברות די גבוהה (ניתן לראות שקו הסיווג שהמודל העביר המסומן באדום רחוק מאוד מקו ההפרדה האמיתי המסומן בשחור). ואכן ככל שהמספר הדגימות עולה כלומר M גדל הקו שהמודל מעביר מסווג בצורה טובה יותר את הדגימות וככל הנראה יוביל להסתברות הרבה יותר גבוהה למזעור אפסילון (המייצג את טווח הטעות) ניתן לראות זאת גם על ידי הקרבה של קו הסיווג שהמודל העביר לקו המרווח האמיתי.

# Classification Boundaries and Model Comparison:

## SVM:

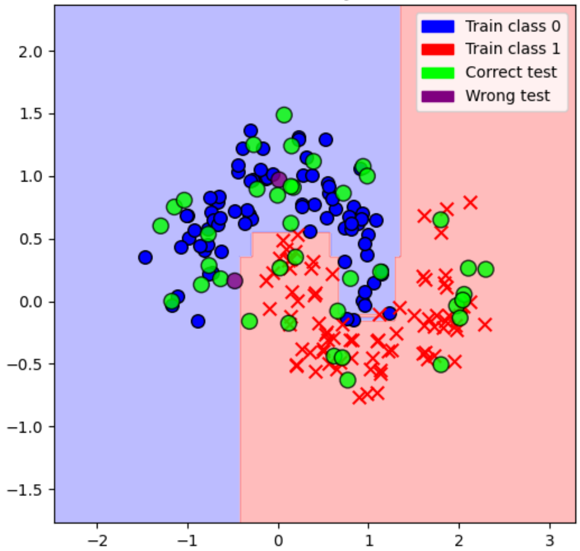


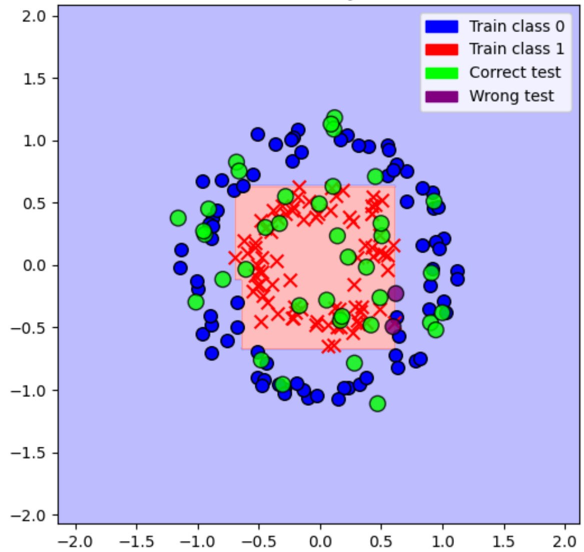



# **Decision Tree:**

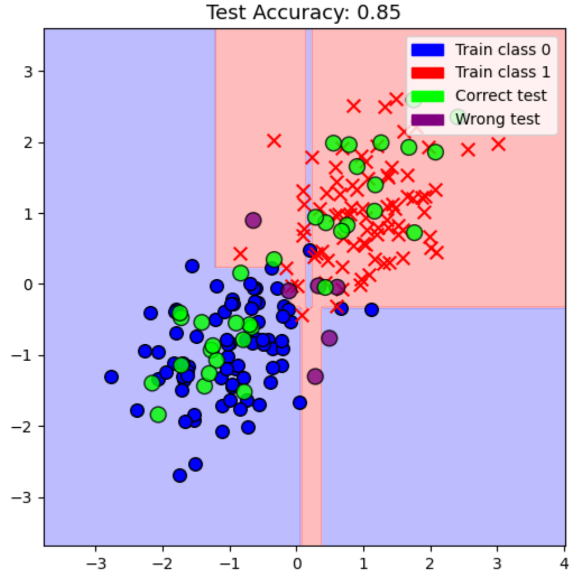
Decision Tree on Moons
Test Accuracy: 0.95



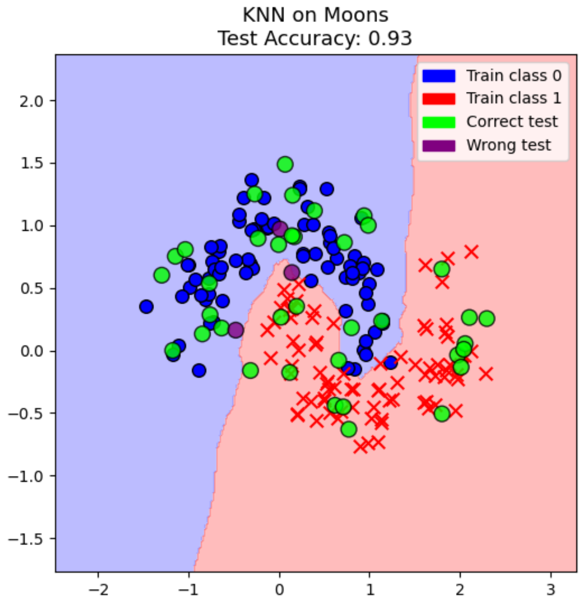

Decision Tree on Circles
Test Accuracy: 0.95

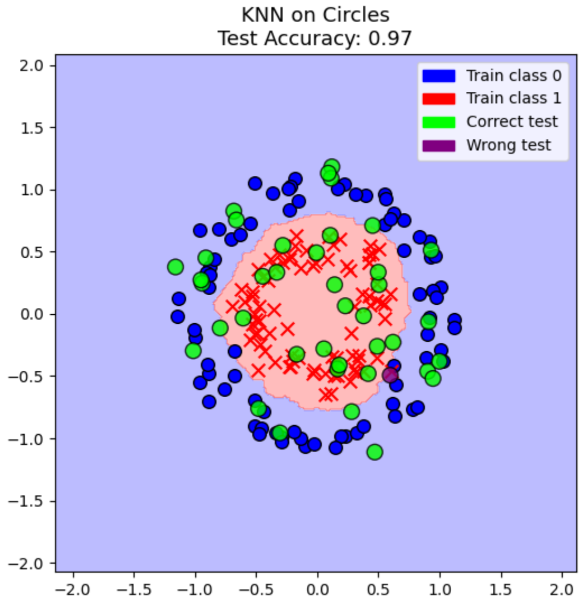


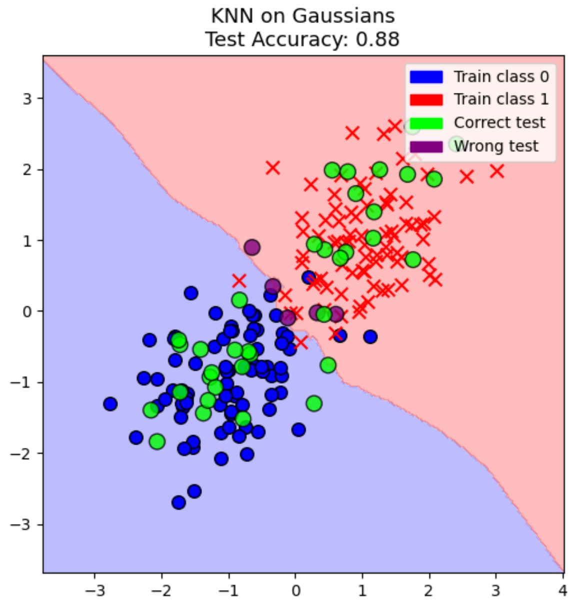

Decision Tree on Gaussians
Test Accuracy: 0.85



## KNN:


KNN on Moons
Test Accuracy: 0.93
Train class 0
Train class 1
Correct test
Wrong test


KNN on Circles
Test Accuracy: 0.97
Train class 0
Train class 1
Correct test
Wrong test


KNN on Gaussians
Test Accuracy: 0.88
Train class 0
Train class 1
Correct test
Wrong test

2) Moons-דגימות עם הפרדה בצורת ירח-  לדאטה סט הזה המודל הטוב ביותר היה עץ החלטות לאחר מכן KNN ולבסוף SVM. תוצאה זאת הגיונית מאחר ולסוג דאטה זה יש הכי הרבה "התערבבות" מרחבית של שני הקלאסים, ולכן הגיוני שלSVM יהיה הכי קשה לסווג מאחר ולא קיים קו לינארי שיעשה הפרדה טובה בין הקלאסים לגבי עץ החלטות וKNN הגיוני שהתוצאה תהיה טובה יותר ודומה כי יותר פשוט להגיע לצורה הנדרשת ע"י סוג הפעולה שלהם.

Circles- דגימות עם הפרדה מעגלית – בדומה לMOONS המודל הגרוע ביותר הוא SVM מאותה סיבה, הדאטה לא מחולק בצורה לינארית ולכן לא יהיה אפשרי להעביר קו ישר שיעשה סיווג טוב. KNN ועשה את הסיווג הטוב ביותר וזה הגיוני בשל כך שהוא יכול ליצור הפרדה דינאמית מעגלית שאינה תלויה בצורת פיזור הדאטה ועץ ההחלטות התקשה טיפה יותר בשל כך שהוא יוצר מלבנים שיותר קשה לחפוף אותם עם צורת המעגל של הדאטה.

Gaussians- שלושת סוגי הסיווג קיבלו פה תוצאה טובה וזאת בשל העובדה שהדאטה מסודר בצורה נוחה לשלושת המודלים להוציא תוצאה אופטימלית, יש הפרדה וקרבה מספקת בין הדגימות כך שכל שלושת במודלים מצליחים להוציא סיווג טוב.

3) SVM- האלגוריתם מייצר גבול לינארי ישר המחלק את המרחב לשני חצאים לפי המרחק הגדול ביותר מהנקודות הקרובות אליו ביותר משני הקלאסים.

עץ החלטות- האלגוריתם יוצר מלבנים לפי שני הצירים ועובר על הדגימות תוך חלוקה של המרחב לשניים בצורה רקורסיבית וסיווג על ידי רוב במלבן שנוצר עד לעצירה על ידי קביעת פרמטר שמסמן מתי לעצור את החלוקה. אלגוריתם זה מחלק את המרחב ל"קופסאות" בצורת מלבן שלכל אחת סיווג אחר. עובד טוב כשהדאטה "מרובע" באופי.

KNN – האלגוריתם מקבל פרמטר K ומסווג כל נקודה לפי סיווג K השכנים הקרובים אליה ביותר ולכן הוא יוצר גבולות מאוד גמישים ובצורות לא סדירות (לעתים סוג של מובלעות) בהתאם לדאטה עצמו.