Theoretical Part:

1.1) Hard&Soft - SVM:

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

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

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

$$\operatorname{argmin} \frac{1}{2} v^T Q v + a^T v \quad s.t \quad Av \leq d$$

נשתמש בכתיב: $\|w\|^2 = w^T I w$ ונסמן: $\boldsymbol{v := \begin{bmatrix} w \\ b \end{bmatrix}}$

אנו רוצים למצוא:

$$\operatorname{argmin} \|w\|^2 = \operatorname{argmin} w^T I w \overset{\text{רצוי}}{=} \frac{1}{2} v^T Q v + a^T v$$

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

$\boldsymbol{Q := \begin{bmatrix} I_d & 0 \\ 0 & 0 \end{bmatrix}}$ בגודל מאחר וכשנכפול את המטריצות b לא ישפיע. ויוכפל ב-0 ונקבל $v^T Q v = \|w\|^2$ כנדרש.

בנוסף נוכל לבחור $\boldsymbol{a := 0_v}$ מאחר שבביטוי הנורמה אין גורם לינארי ונקבל בסך הכל:

$$\operatorname{argmin} \|w\|^2 = \operatorname{argmin} \frac{1}{2} v^T Q v + a^T v$$

לא מוסיף על התוצאה ומקל על הכתיבה בהמשך.

נמשיך בלנסח את אילוצי הבעיה בצורה אחרת (מטריציונית):

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

$$\Updownarrow$$

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

$$\Updownarrow$$

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

כעת נעביר את הכול לכתיב מטריציוני ונבנה את A, d:

$$-y_i(x_i^T w + b) \iff -y_i [x_i^T \ \ 1]\begin{bmatrix} w \\ b\end{bmatrix} = A_i V \le -1$$

נסמן

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

ו $d_i = -1$ לכל הקואורדינטות, 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 ביטוי לכל המרכיבים של התכנון הריבועי כנדרש ☒

1.2) PAC Learnability:

1)a) תהי D התפלגות מעל X, ו-S_m סט דגימות מגודל m (דגימות בלתי תלויות) מעל X. צ"ל שלכל $X'\subseteq X$ כך ש-

$P_D[x\in X']>\epsilon$ מתקיים $P_D[S_m\cap X'=\emptyset]\le e^{-m\epsilon}$

הוכחה:

לדגימה בודדת: $P_D[x\in X']>\epsilon \underset{\text{שיקולי הסתברות}}{\Rightarrow} P_D[x\notin X']<1-\epsilon$ ⊛

מאחר והדגימות בלתי תלויות:

$$P_D[\forall x_i\in S_m \notin X'] = \prod_{i=1}^{m} P_D[x_i\notin X'] \underset{⊛}{\le} (1-\epsilon)^m \underset{\text{א"ש ברנולי}}{\le} e^{-m\epsilon}$$

$\Updownarrow$

$P[S_m\cap X'=\emptyset]$

כנדרש ☒

b) נניח S סט אימון עם m דגימות, ובנוסף נניח שאנו במקרה הריאליזבילי ולכן $\hat f_S$ שנבחרה על ידי A אלג' עקבי בכלל על סט האימון, כלומר: $L_{D,f}(\hat f_S, S)=0$

עלינו לחסום את: $P_D[L_{D,f}(\hat f_S)>\epsilon]$

נשתמש ברמז הראשון ונגדיר סט $Z_{\hat f_S}$ כסט דגימות שבו $\hat f_S$ כושלת. כלומר, $Z_{\hat f_S}:=\{x\in X:\hat f_S(x)\neq f(x)\}$, כעת מאחר ו-$\hat f_S$ אלג' עקבי על סט האימון נקבל שאין בין שני הסטים דגימות חופפות כי אם היו אז שהפונקציה כושלת על סט האימון או

שהיא לא שונה וכן 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

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

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

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

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

הסיכוי לקבל דגימה שבה המודל טועה גדולה מ ε אמ"מ כל הטעויות
וכל האימון אינם בעלי איברים משותפים.

סעיף קודם

c)

בסעיף b הראנו שלכל פונקציה ממחלקת ההיפותזיות 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.

מהגדרה האחרונה מתקיים:

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

כעת, מאחר והאלגוריתם שלנו מחזיר שגיאה אפסית על סט האימון הסיכוי שנטעה ונבחר היפותזה שגויה חסום גם הוא על ידי אותו חסם ביחוד מאחר והמאורע הזה כלול במקרה הכללי ונקבל:

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

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

ונקבל:

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

$\Updownarrow$ שיקולי הסתברות

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

ב. נפתור שאלה זו בדומה לדוגמת האינטרוולים שפגשנו בתרגול.

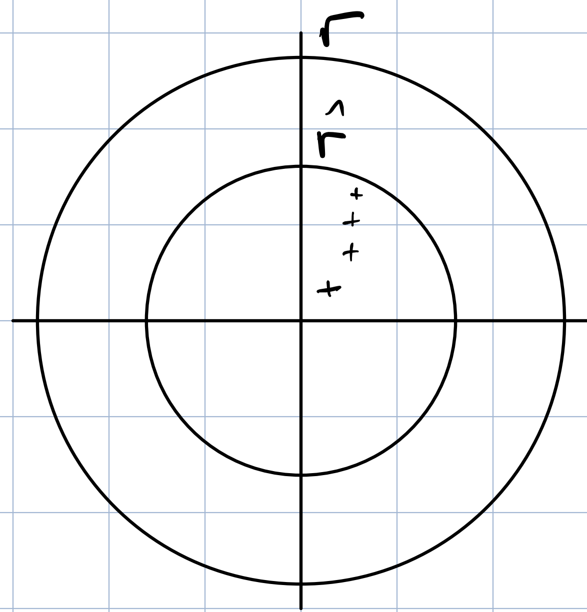
כדי להראות שH למידה PAC עלינו להראות אלגוריתם אשר מקיים: $\mathbb{P}_D(L_{D,f}(A(S_m))<\varepsilon)\ge 1-\delta$.

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

היתה דגימה חיובית באיזור שגדול מ r

בסתירה להגדרת r. ולכן שגיאה יכולה

להיות רק כאשר נקבל דגימה חיובית

שנמצאת ברדיוס בין $\hat{r}$ ל r. ולא היתה

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 ε, נרצה להראות כי:

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

נשים לב קודם כל שחייבת להיות טבעת עם רדיוס בין r ל $\hat{r}$

והסתברות $<\varepsilon$ לקבל דגימה מהטבעת הזאת כי אם לא היה

אז הטבעת האמיתית היתה שווה לטבעת שלנו והשגיאה היתה 0

ולכן הניסוי היה נכון באופן ריק.

במקרה שהטבעת בניהם אכן קיימת:

נשתמש באי השיוויון מהסעיף הקודם, ההסתברות לקבל דגימה מהטבעת

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 δ להיות: $e^{-\varepsilon m} \le \delta$

ונקבל ע"י הוצאת ln:

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

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

כלומר בסך הכל הגענו:

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

שיקולי הסתברות

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

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

3) נשים: 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 A PAC, כלומר, כך:

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

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

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

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

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

וכעת נקבל:

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

שמצאנו בדיוק תכונת APAC כנדרש ☑

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

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 סוכמת את i האיברים $i\in I$ בדגימה ומסווגה 1,0 בהתאם לתוצאה של הסכום mod2.

מאחר והאיברים בדגימות הם 1,0 נקבל ש h_I בעצם מחזירה 1 אם מספר ה1-ים באיברים $i\in I$ אי זוגי ו 0 אם הוא זוגי. כעת נרצה למצוא 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: $I=\{j: y_j=1\}$ וככה נוכל לייצר כל תווית,

שנרצה, כלומר ניתצנו דגימה מגודל n ולכן מתקיים $\boxed{VCdim(H)\ge n}$

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 שאפשר היפותזה

יכולה לנתץ, כלומר להשיג כל תווית אפשרית לדגימה באמצעות

הפונקציות של ה, 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, או במילים אחרות:

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