הטכניון – מכון טכנולוגי לישראל

הפקולטה למדעי המחשב

מבוא למערכות לומדות (236756)

סמסטר חורף תשפ"ד – 09 ביוני 2024

מרצה: ד"ר יונתן בלינקוב

מבחן מסכם מועד ב'

הנחיות הבחינה:

- **משך הבחינה:** שלוש שעות + 12 דקות שנוספו במהלך הבחינה
- **חומר עזר:** המבחן בחומר סגור (ללא ספרים, מחברות, דפי נוסחאות).
- מחשבון: מותר.
- כלי כתיבה: עט בלבד.
- יש לכתוב את התשובות **על גבי שאלון זה**.
- מותר לענות בעברית או באנגלית.
- הוכחות והפרכות צריכות להיות פורמליות.
- קריאות:
  - סימונים לא ברורים בשאלות רב-ברירה ו/או תשובות מילוליות בכתב יד לא קריא יובילו לפסילת התשובה.
  - לא יתקבלו ערעורים בנושא.
- במבחן 16 עמודים ממוספרים סה"כ, כולל עמוד שער זה שמספרו 1 ושלושה עמודי טיוטה בסוף הגיליון.
- נא לכתוב רק את המבוקש ולצרף הסברים קצרים עפ"י ההנחיות.
- **בתום המבחן יש להגיש את שאלון זה בלבד.**
- **לזכאים להערכה חלופית מתאפשרת בחירה בין שאלות 3 ו-4.**

בהצלחה!

שאלה 1: רגרסיה לינארית ו-Generative models [18 נק']

נתון דאטה $S=\{(\mathbf{x}_i,y_i)\}_{i=1}^m$ שהגיע ממודל ליניארי $y_i=\mathbf{w}^\top\mathbf{x}_i+\varepsilon_i$ עם רעש אקראי מפילוג i.i.d. נורמלי: $\varepsilon_i\sim\mathcal{N}(0,1)$.

שימו לב: הדוגמאות $\mathbf{x}_i\in\mathbb{R}^d$ והתיוגים $y_i\in\mathbb{R}$ נתונים. וֶקטור המשקלים $\mathbf{w}\in\mathbb{R}^d$ לא ידוע ואותו אנו רוצים ללמוד.

תזכורת: הוכחנו שתחת הנחות אלה ה-likelihood שווה ל:

$$L(\mathbf{w};S)=\Pr\left(\left\{\left(\mathbf{x}_i,y_i\right)\right\}_i\middle|\boldsymbol{w}\right)=\prod_{i=1}^m\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\mathbf{w}^\top\mathbf{x}_i-y_i\right)^2\right\}$$

נניח בנוסף **שווקטור המשקלים** הלא ידוע $\mathbf{w}$ נדגם מהתפלגות גאוסיאנית רב ממדית בתוחלת $\boldsymbol{\mu}\in\mathbb{R}^d$ ושונות $\boldsymbol{\Sigma}\succ\mathbf{0}_{d\times d}$:

$$\mathbf{w}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})\ \implies\ f(\mathbf{w})=(2\pi|\boldsymbol{\Sigma}|)^{-1/2}\exp\left\{-\tfrac{1}{2}(\mathbf{w}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{w}-\boldsymbol{\mu})\right\}$$

א. [13 נק'] הוכיחו שתחת כלל ההנחות, MAP עם prior גאוסיאני רב-ממדי על המשקלים, משמע $\underset{\mathbf{w}\in\mathbb{R}^d}{\operatorname{argmax}}\Pr(\mathbf{w}\mid S,\boldsymbol{\mu},\boldsymbol{\Sigma})$, שקול לבעיית ה-Regularized LS הבאה (עבור $\mathbf{z}\in\mathbb{R}^d,\mathbf{M}\succ\mathbf{0}_{d\times d},\lambda>0$):

$$\hat{\mathbf{w}}=\underset{\mathbf{w}\in\mathbb{R}^d}{\operatorname{argmin}}\left(\frac{1}{m}\sum_{i=1}^m(\mathbf{w}^\top\mathbf{x}_i-y_i)^2+\lambda(\mathbf{w}-\mathbf{z})^\top\mathbf{M}(\mathbf{w}-\mathbf{z})\right).$$

ההוכחה צריכה לכלול פיתוח פורמלי מנומק. יש להתייחס ל-$S,\lambda,\mathbf{M},\mathbf{z}$ כנתונים ולמצוא ערכים מתאימים ל-$\boldsymbol{\mu},\boldsymbol{\Sigma}$.

הוכחה פורמלית:

(מקום נוסף בעמוד הבא)

המשך ההוכחה:

ב. [5 נק'] יהי $\widehat{\mathbf{w}}$ פתרון בעיית הרגולריזציה המוכללת כפי שהגדרנו לעיל. יהי $\widehat{\mathbf{w}}_{\mathrm{LS}}$ פתרון ה-Least Squares ללא רגולריזציה.
בטענה שלפניכם, כתבו את היחס ($\geq,\leq,<,>,=$) המדויק ביותר כך שתהיה נכונה בהכרח:

$$\frac{1}{m}\sum_{i=1}^{m}\left(\widehat{\mathbf{w}}_{\mathrm{LS}}^{\top}\mathbf{x}-y_i\right)^2 \underbrace{\qquad\qquad}_{\text{השלימו}} \frac{1}{m}\sum_{i=1}^{m}\left(\widehat{\mathbf{w}}^{\top}\mathbf{x}-y_i\right)^2.$$

הסבירו בקצרה:

שאלה 2: Kernel SVM [24 נק']

נתון $\gamma > 0$. נגדיר את ה-Gaussian kernel לקלט חד-ממדי: $K(x_i, x_j) = \exp\left(-\gamma\left(x_i - x_j\right)^2\right)$, $K: \mathbb{R} \times \mathbb{R} \to \mathbb{R}$.

תכונה אלגברית 1: $K(x_i, x_j) = \exp\left(-\gamma\left(x_i - x_j\right)^2\right) = \exp\left(-\gamma\left(x_i^2 + x_j^2\right)\right) \sum_{n=0}^{\infty} \frac{(2\gamma)^n \left(x_i x_j\right)^n}{n!}$.

u v

א. [6 נק'] הציעו פונקציית מיפוי $\phi: \mathbb{R} \to \mathbb{R}^p$ והוכיחו בעזרתה שהפונקציה K מהווה קרנל חוקי (בחד ממד).

שימו לב: יש לבחור $p \in \mathbb{N} \cup \{\infty\}$ מתאים, סופי או אינסופי. כדאי להשתמש בתכונה הנתונה.

תשובה:

$$K(u,v) = e^{-\gamma u^2} \cdot e^{-\gamma v^2} \cdot \sum_{n=0}^{\infty} \frac{(2\gamma)^{\frac{n}{2}} u^n}{\sqrt{n!}} \cdot \frac{(2\gamma)^{\frac{n}{2}} v^n}{\sqrt{n!}}$$

$$\varphi_n(u) = e^{-\gamma u^2} \cdot \frac{(2\gamma)^{\frac{n}{2}} u^n}{\sqrt{n!}} \quad . \quad \varphi(u) = \left[e^{-\gamma u^2},\ e^{-\gamma u^2} \cdot \frac{(2\gamma)^{\frac{1}{2}} \cdot u^1}{\sqrt{1!}},\ e^{-\gamma u} \frac{(2\gamma)^{\frac{2}{2}} \cdot u^2}{\sqrt{2!}}, \dots\right]$$

האיבר ה-n-י

$$\left(\varphi: \mathbb{R} \to \mathbb{R}^{\infty}\right) \qquad \varphi(u) \cdot \varphi(v) = K(u,v)$$

קל לראות

כזכור, בעיות SVM ניתן לפתור בצורת primal problem ובצורת dual problem.

עם זאת, בגלל קושי מובנה בפתרון ה-primal problem עם ה-feature mapping מהסעיף הקודם (חִשְׁבוּ מה הקושי), היינו רוצים לפתור את ה-dual problem במקום. כפי שנראה עתה, גם זה עלול להיות בעייתי.

ב. [6 נק'] נתון מדגם אימון $S=\{(x_i,y_i)\}_{i=1}^{10,000}$. הסבירו את הקושי העיקרי בפתרון ה-dual problem עם K שהוגדר.

הסבירו בקצרה:

"הקושי" הוא המחיר במקום, בגלל המט' שמייצגת את

הקרנל (בערך $\binom{10000}{2}$ תאים עבור כל $K(x_i,x_j)$ (סימטרי))

כלומר זה יהיה פתיר אך יקר חישובית

קירוב אפשרי: ידוע שמתקיים הקירוב הבא $K(x_i,x_j)\approx K'(x_i,x_j)\triangleq\frac{1}{500}\sum_{n=1}^{500}2\cdot\cos(w_n\overset{u}{x_i}+b_n)\cdot\cos(w_n\overset{v}{x_j}+b_n)$,

כאשר דוגמים 500 זוגות $w_n\sim\mathcal{N}(0,2\gamma),\ \ b_n\sim\mathrm{Uniform}[0,2\pi]$ פעם אחת ומשתמשים בהם לחישוב K' לכל הזוגות i,j.

ג. [6 נק'] הציעו פונקציית מיפוי $\psi:\mathbb{R}\to\mathbb{R}^p$ והוכיחו בעזרתה שהפונקציה K' מהווה קרנל חוקי.

שימו לב: יש לבחור $p\in\mathbb{N}\cup\{\infty\}$ מתאים, סופי או אינסופי.

תשובה:

$K'(u,v)=\frac{2}{500}\sum_{n=1}^{500}\cos(w_nu+b_n)\cdot\cos(w_nv+b_n)=$

$=\sqrt{\frac{2}{500}}\cdot\sqrt{\frac{2}{500}}\sum(\qquad)\cdot(\qquad)=\varphi(u)^T\cdot\varphi(v)$

$\varphi(u)=\sqrt{\frac{2}{500}}\left(\cos(w_1u+b_1),\cos(w_2u+b_2),\ldots,\cos(w_{500}u+b_{500})\right)$

$\forall n\in[500]:\ \varphi_n(u)=\sqrt{\frac{2}{500}}\cos(w_nu+b_n)\qquad\varphi:\mathbb{R}\to\mathbb{R}^{500}$

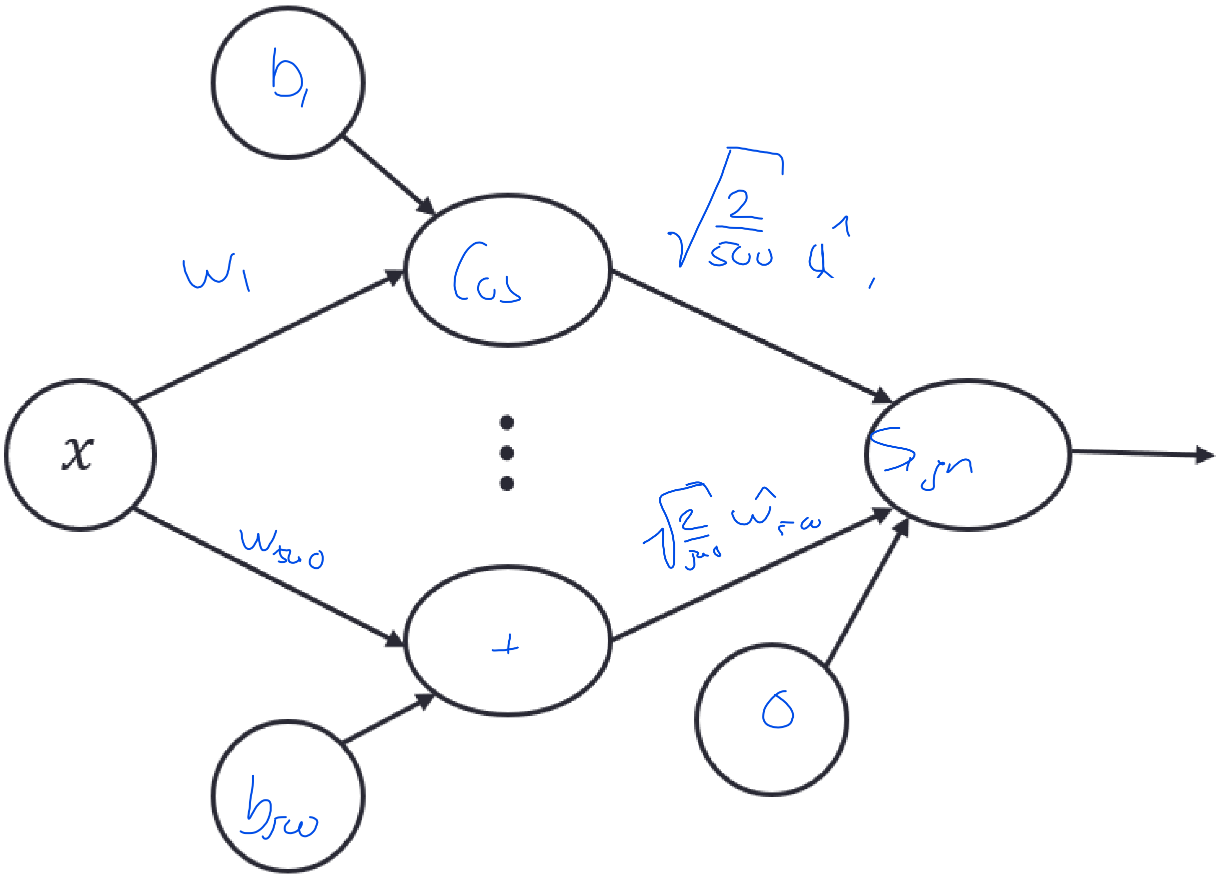
ד. [6 נק'] נתון מדגם אימון $S = \{(x_i, y_i)\}_{i=1}^{10,000}$. נפתור בעיית SVM (הומוגנית) עם המיפוי ψ מהסעיף הקודם בצורת ה-primal problem ונסמן את הפתרון בתור $\hat{\mathbf{w}}$.

הציעו <u>השמה</u> לרשת נוירונים עם שכבה חבויה אחת, כך שפעולתה על $x \in \mathbb{R}$ שרירותי תהיה <u>זהה</u> לזו של <u>המסווג</u> הלינארי ההומוגני שמושרה ע"י $\hat{\mathbf{w}}$ (ומחזיר ± 1).

כִּתְבוּ במפורש על הקשתות המתאימות בתרשים את המשקלים ועל הצמתים את ערכי ה-bias ופונקציות האקטיבציה.

<u>הבהרה</u>: בהשמה תוכלו להשתמש בקבועים, פונקציות ו/או בנתונים שהתקבלו עד כה:

$$S = \{(x_i, y_i)\}_{i=1}^{10,000},\ \ w_1, \dots, w_{500},\ \ b_1, \dots, b_{500},\ \ \hat{\mathbf{w}},\ \ K',\ \ \gamma$$



שאלה 3: PAC learnability [29 נק']

☐ לזכאים להערכה חלופית <u>בלבד</u> (כפי שהוגדרו באתר הקורס): סמנו את התיבה הזו אם ברצונכם **לדלג** על שאלה זו. המשקל של שאלות 1,2,4 יתפזר באופן יחסי על פני 100 נקודות. ניתן לדלג רק על שאלה אחת מתוך שאלות 3,4.

הבהרה: בכל השאלה הסימון $S\sim\mathcal{D}^m$ משמעו שדוגמים m דוגמאות מהתפלגות $\mathcal{D}$ באופן i.i.d.

א. [4 נק'] להלן הגדרת ה-PAC-למידות במקרה ה-realizable.

השלימו את החסר (בשורה האחרונה):

מחלקת היפותזות סופית $\mathcal{H}$ היא PAC-למידה אם קיימים פונקציה $m_{\mathcal{H}}:(0,1)^2\to\mathbb{N}$ ואלגוריתם למידה A כך ש:

- לכל $\epsilon,\delta\in(0,1)$ והתפלגות $\mathcal{D}$ שהיא realizable ע"י $\mathcal{H}$
- נסמן את שגיאת ההכללה של היפותזה h ע"י $L_{\mathcal{D}}(h)$
- עבור גודל מדגם $m\ \ge\ m_{\mathcal{H}}(\epsilon,\delta)$, האלגוריתם מחזיר היפותזה שהיא (ϵ,δ)-PAC. משמע,

$$\underbrace{\Pr_{S\sim\mathcal{D}^m}\text{(L\_D(A(S) <= \textbackslash epsilon) > 1- \textbackslash delta}}_{\text{השלימו}}$$

<u>שימו לב</u>: טעות בסעיף א' עלולה לגרור טעויות בחלק משאר הסעיפים ולעלות בניקוד רב! **ודאו את תשובתכם!**

ב. [5 נק'] יהי מרחב דוגמאות **<u>סופי</u>** $\mathcal{X}$ כלשהו. תהי מחלקת היפותזות סופית $\mathcal{H}$ שרירותית שהיא PAC-למידה.

טענה: קיימים בהכרח פונקציה $m_{\mathcal{H}}:(0,1)\to\mathbb{N}$ ואלגוריתם למידה A כך ש:

- לכל $\epsilon\in(0,1)$ והתפלגות $\mathcal{D}$ שהיא realizable ע"י $\mathcal{H}$
- עבור מדגם $S\sim\mathcal{D}^m$ שמקיים $m\ \ge\ m_{\mathcal{H}}(\epsilon)$
- האלגוריתם מחזיר היפותזה עם שגיאת הכללה חסומה, משמע: $L_{\mathcal{D}}\big(A(S)\big)\le\epsilon$.

סמנו את האפשרות הנכונה בהכרח לגבי הטענה שלעיל.

a. X הטענה נכונה כי היא נובעת מהגדרת ה-PAC-למידות.

b. X הטענה נכונה כי תמיד ניתן לבחור $m_{\mathcal{H}}(\epsilon)$ גדול מספיק שמבטיח שגיאה קטנה כרצוננו במרחב דוגמאות סופי.

c. X הטענה שגויה כי היא אינה נובעת מהגדרת ה-PAC-למידות.

d. V הטענה שגויה כי לא ייתכן $m_{\mathcal{H}}(\epsilon)$ שמבטיח שהמדגם מכסה את המרחב כולו.

הנחיות לסעיפים הבאים: יהי $\mathcal{X}$ מרחב דוגמאות סופי ונניח $|\mathcal{X}| \geq 10$.

תהי התפלגות $\mathcal{D}$ לפיה ההסתברות להגריל $x \in \mathcal{X}$ כלשהו נתונה ע"י $\mathcal{D}(x) > 0$ (משמע, לכל x יש הסתברות חיובית ממש).

תהי מחלקת היפותזות סופית $\mathcal{H}_{\text{single}} = \{h_z : z \in \mathcal{X}\} \cup \{h^0\}$ כאשר לכל $x, z \in \mathcal{X}$ נגדיר $h_z(x) = \begin{cases} 1, & x = z \\ 0, & x \neq z \end{cases}$ וכן $h^0(x) = 0$.

ידוע שיש בדיוק דוגמה אחת במרחב $\mathcal{X}$ שמתויגת חיובית. נסמן אותה ע"י x_+ (מתקיים $y(x_+) = 1$).

ג. נציע אלגוריתם למידה שיבצע ERM עבור $\mathcal{H}_{\text{single}}$ ומדגם אימון $S \sim \mathcal{D}^m$ עבור $m \geq 1$ (דגימה i.i.d.).

a. [4 נק'] השלימו את האלגוריתם.

קלט: מדגם $S = \{(x_i, y_i)\}_{i=1}^m$ שנדגם i.i.d. מההתפלגות $\mathcal{D}$ המתוארת.

פלט: מסווג שנלמד בעזרת ERM ומסומן בתור h_S.

אלגוריתם:

if x+ \in S => h_s = h_(x+)
else h_s = h^0(x)

- נחזיר את h_S

b. [4 נק'] הסבירו בקצרה מדוע מדובר באלגוריתם ERM.

הסבר:

ERM because the Training error will always be 0

c. [4 נק'] כִּתְבוּ ביטוי מתמטי לשגיאת ההכללה של האלגוריתם שהצעתם (ניתן להשתמש ב-$\mathcal{D}, S, x_+$; ללא הוכחה):

$L_{\mathcal{D}}(h_S) =$ { 0 ; x+ \in S
{ D(x+)   x+ \notin S

ד. [8 נק'] מצאו ביטוי מפורש לפונקציה $m_{\mathcal{H}_{\text{single}}}(\epsilon,\delta)$ של האלגוריתם שהצעתם בסעיף הקודם.
משמע, מצאו את ה-sample complexity שיבטיח את פעולת האלגוריתם במובני (ϵ,δ)-PAC.
תשובה מלאה צריכה להיות תלויה ב-ϵ,δ בלבד. תשובה שתלויה גם ב-$\mathcal{D},S,x_+$ (חלקם או כולם) תקבל ניקוד חלקי.
הראו את צעדי הפיתוח בצורה מנומקת.

תשובה ופיתוח (לרשותכם דפי טיוטה בסוף הגיליון):

m_H(\epsilon \delta) <= (VC(H)log(1/epsilon) + log(1/delta)) / epsilon

שאלה 4: צירופים של מסווגים לינאריים [29 נק']

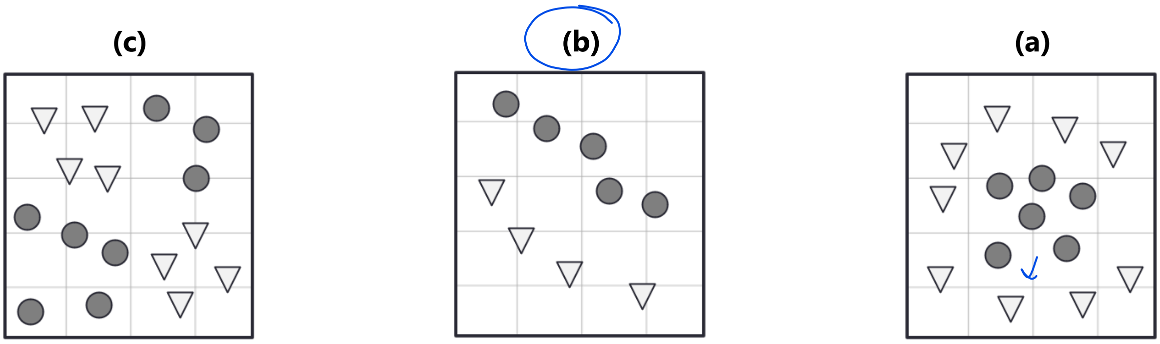
☐ לזכאים להערכה חלופית בלבד (כפי שהוגדרו באתר הקורס): סמנו את התיבה הזו אם ברצונכם **לדלג** על שאלה זו. המשקל של שאלות 1,2,3 יתפזר באופן יחסי על פני 100 נקודות. ניתן לדלג רק על שאלה אחת מתוך שאלות 3,4.

בכל השאלה מרחב הדוגמאות הוא $\mathcal{X}=\mathbb{R}^d$ ומרחב התיוגים הוא $\mathcal{Y}=\{-1,+1\}$ עבור $d\in\mathbb{N}$ שרירותי.

עבור $K\geq 1$ שרירותי, נגדיר מחלקת היפותזות ראשונה:

$$\mathcal{H}_1^{(K)}=\left\{h_\theta \,\middle|\, \theta=\left(\underbrace{\mathbf{w}_1,\dots,\mathbf{w}_K}_{\in\mathbb{R}^d},\ \underbrace{b_1,\dots,b_K}_{\in\mathbb{R}},\ \underbrace{\boldsymbol{\alpha}}_{\in\mathbb{R}^K}\right)\right\},\qquad \text{where}\qquad h_\theta(\mathbf{x})=\operatorname{sign}\left(\sum_{k=1}^K\alpha_k\left(\mathbf{w}_k^\top\mathbf{x}+b_k\right)\right).$$

א. [4 נק'] אילו מה-datasets הדו-ממדיים הבאים ניתן לסווג באופן מושלם עם היפותזות מהמחלקה $\mathcal{H}_1^{(K)}$ (עבור K גדול מספיק)? הקיפו בבירור את האותיות המתאימות והסבירו בקצרה.



הסבר קצר: המויים לינארי

בסעיף הבא, נגדיר גרסה "פשוטה" של $\mathcal{H}_1^{(K=1)}$ ללא bias ונשמיט את האינדקסים של $\mathbf{w}_1, \alpha_1$.

משמע, ההיפותזות הן מהצורה: $h_\theta(\mathbf{x}) = \operatorname{sign}(\alpha \mathbf{w}^\top \mathbf{x})$.

ב. [8 נק'] בהינתן מדגם $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, נגדיר בעיית אופטימיזציה ללמידת פונקציה "פשוטה" כמתואר לעיל:

$$\min_{\mathbf{w}\in\mathbb{R}^d, \alpha\in\mathbb{R}} \sum_{i=1}^m \max\big(0,\ \ 1 - y_i(\alpha \mathbf{w}^\top \mathbf{x}_i)\big).$$

נבחן את הקמירות של הבעיה (לפי $\mathbf{w}, \alpha$ יחדיו) ע"י בחינת $\ell(\mathbf{w}, \alpha) = \max\big(0,\ \ 1 - y(\alpha \mathbf{w}^\top \mathbf{x})\big)$.

תזכורת: הפונקציה קמורה אמ"מ לכל $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d,\ \alpha_1, \alpha_2 \in \mathbb{R}$ ולכל $t \in [0,1]$ מתקיים:

$$t \cdot \ell(\mathbf{w}_1, \alpha_1) + (1-t) \cdot \ell(\mathbf{w}_2, \alpha_2) \geq \ell(\underbrace{t\mathbf{w}_1 + (1-t)\mathbf{w}_2}_{\alpha w},\ \ \underbrace{t\alpha_1 + (1-t)\alpha_2}_{\alpha})$$

הוכיחו: הפונקציה $\ell(\mathbf{w}, \alpha)$ **__אינה__** קמורה. ההוכחה צריכה להיות פורמלית.

__רמז:__ בהוכחה ניתן לבחור $\mathbf{x}, y$ מסוימים לבחירתכם או להוכיח עבור $\mathbf{x}, y$ כלליים.

הוכחה (לרשותכם דפי טיוטה בסוף הגיליון):

$$t \cdot \max\{0, \overbrace{1 - y(\alpha_1 w_1^T \underline{x})}^{0}\}_{1\ w_1} + (1-t)\max\{0, \overbrace{1 - y(\alpha_2 w_2^T \underline{x})}^{0}\}_{-1\ (-w_1)} \overset{\text{want}}{<}$$

$$\max\{0, \underbrace{1 - y \cdot \overbrace{[t\alpha_1 + (1-t)\alpha_2]}^{\alpha} \cdot \overbrace{[t w_1 + \underbrace{(1-t) w_2}_{0}]}^{w} \underline{x}}_{1}\}$$

$t = \frac{1}{2}$ $\alpha_1 = 1$ $\alpha_2 = -1$ $w_1 = -$ $_2 = \frac{1}{(y)\|x\|^2}$

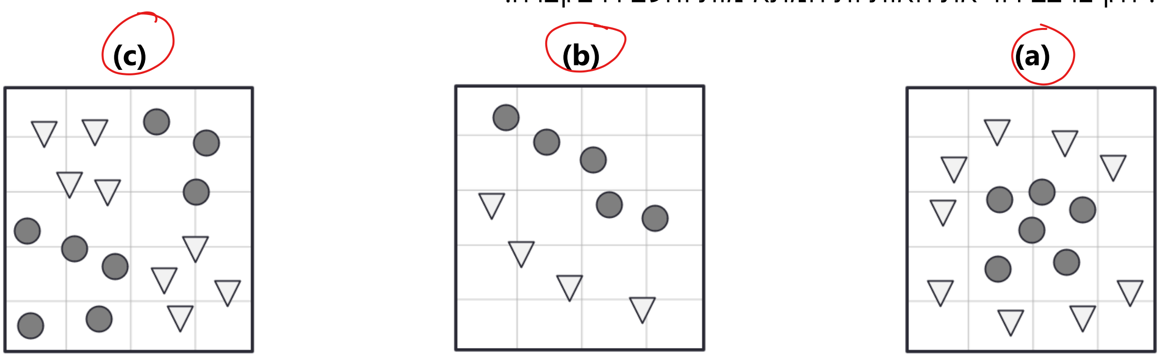
(מקום נוסף בעמוד הבא; **ניתן לפתור את הסעיפים בהמשך גם בלי לפתור את הסעיף הנוכחי!**)

המשך ההוכחה:

עבור $K \geq 1$ שרירותי, נגדיר מחלקת היפותזות שנייה:

$$\mathcal{H}_2^{(K)} = \left\{ h_\theta \;\middle|\; \theta = \Big(\underbrace{\mathbf{w}_1, \dots, \mathbf{w}_K}_{\in\mathbb{R}^d}, \underbrace{b_1, \dots, b_K}_{\in\mathbb{R}}, \underbrace{\boldsymbol{\alpha}}_{\in\mathbb{R}^K} \Big) \right\}, \quad \text{where} \quad h_\theta(\mathbf{x}) = \operatorname{sign}\Big(\textstyle\sum_{k=1}^K \alpha_k \operatorname{sign}\big(\mathbf{w}_k^\top \mathbf{x} + b_k\big)\Big).$$

ג. [4 נק'] אילו מה-datasets הדו-ממדיים הבאים ניתן לסווג באופן מושלם עם היפותזות מהמחלקה $\mathcal{H}_2^{(K)}$ (עבור K גדול מספיק)? הקיפו בבירור את האותיות המתאימות והסבירו בקצרה.



הסבר קצר:

ensemble

ד. [13 נק'] בהינתן מדגם $S=\{(\mathbf{x}_i,y_i)\}_{i=1}^m$, נגדיר בעיית אופטימיזציה ללמידת פונקציה מ-$\mathcal{H}_2^{(K)}$:

$$\min_{\substack{\mathbf{w}_1,\dots,\mathbf{w}_K\in\mathbb{R}^d,\\ \alpha_1,\dots,\alpha_K\in\mathbb{R}}} \sum_{i=1}^m \max\left(0,\ \ 1-y_i\left(\sum_{k=1}^K \alpha_k \operatorname{sign}\left(\mathbf{w}_k^\top \mathbf{x}_i+b_k\right)\right)\right).$$

במטרה ללמוד עם gradient descent, נגזור את הרכיב של פונקציית המטרה שתלוי בדוגמה ה-i לפי פרמטרים שונים.

הקלה: בכל מקום בו יש לגזור פונק' שאינה גזירה בנקודה יחידה, תוכלו להתעלם מנק' זו ולהניח שלעולם לא נגיע אליה.

a. עבור $j\in[K]$ נתון, כתבו את הנגזרת לפי α_j. נדרשת תשובה סופית בלבד (לרשותכם דפי טיוטה בסוף הגיליון).

$$\frac{\partial}{\partial\alpha_j}\max\left(0,1-y_i\overbrace{\left(\sum_{k=1}^K \alpha_k \operatorname{sign}(\mathbf{w}_k^\top\mathbf{x}_i+b_k)\right)}^{Z}\right)=$$

$$\begin{cases} 0 & ; \quad Z\geq 1 \\ -y_i\operatorname{sign}(w_k x_i+b_k) & ; \quad Z<1 \end{cases}$$

b. עבור $j\in[K]$ נתון, כתבו את הגרדיינט לפי $\mathbf{w}_j$. נדרשת תשובה סופית בלבד (לרשותכם דפי טיוטה בסוף הגיליון).

$$\nabla_{\mathbf{w}_j}\max\left(0,1-y_i\overbrace{\left(\sum_{k=1}^K \alpha_k \operatorname{sign}(\mathbf{w}_k^\top\mathbf{x}_i+b_k)\right)}\right)= 0_d$$

$$=\begin{cases} 0 & ; \quad Z\geq 1 \\ 0 & \end{cases}$$

(באופן דומה ניתן לחשב גם את הנגזרת לפי b_j, אך נדלג על זה כעת)

c. בהינתן התשובות לעיל (וללא קשר לקמירות), מה הבעייתיות במינימיזציה של הבעיה עם gradient descent?

תשובה: לא נוכל לעצור לפיה ש מכיוון שהנגזרות מתאפסות
ואין כל פעם הדגמה אחת, והיה גדול סיכוי

מסגרת נוספת (יש לציין אם מדובר בטיוטה או בהמשך לתשובה אחרת):

מסגרת נוספת (יש לציין אם מדובר בטיוטה או בהמשך לתשובה אחרת):

מסגרת נוספת (יש לציין אם מדובר בטיוטה או בהמשך לתשובה אחרת):