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' – פתרון חלקי

שימו לב: הפתרונות המופיעים כאן הם חלקיים בלבד ומובאים בשביל לעזור לכם בתהליך הלמידה.

ייתכנו כאן חוסרים / ליקויים / טעויות של ממש.

בהצלחה!

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

נתון דאטה $S=\{(\mathbf{x}_i,y_i)\}_{i=1}^m$ שהגיע ממודל לינארי $y_i=\mathbf{w}^\top\mathbf{x}_i+\varepsilon_i$ 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 $\mathbf{\Sigma}\succ\mathbf{0}_{d\times d}$:

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

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

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

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

$$\operatorname*{argmax}_{\mathbf{w}\in\mathbb{R}^d}\Pr(\mathbf{w}\mid S,\boldsymbol{\mu},\mathbf{\Sigma})\overset{Bayes}{=}\operatorname*{argmax}_{\mathbf{w}\in\mathbb{R}^d}\frac{\Pr(S\mid\mathbf{w},\boldsymbol{\mu},\mathbf{\Sigma})\Pr(\mathbf{w}\mid\boldsymbol{\mu},\mathbf{\Sigma})}{\Pr(S\mid\boldsymbol{\mu},\mathbf{\Sigma})}=\operatorname*{argmax}_{\mathbf{w}\in\mathbb{R}^d}\frac{\Pr(S\mid\mathbf{w})\Pr(\mathbf{w}\mid\boldsymbol{\mu},\mathbf{\Sigma})}{\Pr(S\mid\boldsymbol{\mu},\mathbf{\Sigma})}$$

$$\overset{indep.\,of\,w}{=}\operatorname*{argmax}_{\mathbf{w}\in\mathbb{R}^d}\Pr(S\mid\mathbf{w})\Pr(\mathbf{w}\mid\boldsymbol{\mu},\mathbf{\Sigma})\overset{monot.}{=}\operatorname*{argmax}_{\mathbf{w}\in\mathbb{R}^d}\ln(\Pr(S\mid\mathbf{w})\Pr(\mathbf{w}\mid\boldsymbol{\mu},\mathbf{\Sigma}))$$

$$=\operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d}(-\ln(L(\mathbf{w};S))-\ln(\Pr(\mathbf{w}\mid\boldsymbol{\mu},\mathbf{\Sigma})))$$

$$=\operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d}\left(-\ln\left(\prod_{i=1}^m\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}(\mathbf{w}^\top\mathbf{x}_i-y_i)^2\right\}\right)-\ln\left(\frac{1}{\sqrt{2\pi|\mathbf{\Sigma}|}}\exp\left\{-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})^\top\mathbf{\Sigma}^{-1}(\mathbf{w}-\boldsymbol{\mu})\right\}\right)\right)$$

$$=\operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d}\left(\frac{m}{2}\ln(2\pi)+\frac{1}{2}\ln(2\pi|\mathbf{\Sigma}|)+\frac{1}{2}\sum_{i=1}^m(\mathbf{w}^\top\mathbf{x}_i-y_i)^2+\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})^\top\mathbf{\Sigma}^{-1}(\mathbf{w}-\boldsymbol{\mu})\right)$$

$$=\operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d}\left(\sum_{i=1}^m(\mathbf{w}^\top\mathbf{x}_i-y_i)^2+(\mathbf{w}-\boldsymbol{\mu})^\top\mathbf{\Sigma}^{-1}(\mathbf{w}-\boldsymbol{\mu})\right)$$

מתקבלת שקילות כאשר $\boldsymbol{\mu}=\mathbf{z},\mathbf{\Sigma}=\frac{1}{m\lambda}\mathbf{M}^{-1}$.

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

$$\frac{1}{m}\sum_{i=1}^m(\hat{\mathbf{w}}^\top\mathbf{x}-y_i)^2\geq\frac{1}{m}\sum_{i=1}^m\left(\hat{\mathbf{w}}_{\text{LS}}^\top\mathbf{x}-y_i\right)^2.$$

הסבירו בקצרה: $\hat{\mathbf{w}}_{\text{LS}}$ מביא למינימום את שגיאת ה-MSE, ולכן לא ייתכן ש-$\hat{\mathbf{w}}$ מביא שגיאה נמוכה יותר.

עם זאת, יכול להתקבל שיוויון ב-datasets בהם $\mathbf{z}$ הנתון משיג שגיאת אימון (MSE) אפסית.

שאלה 2: Kernel SVM [24 נק']

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

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

א. [6 נק'] הציעו פונקציית מיפוי $\phi: \mathbb{R} \rightarrow \mathbb{R}^p$ 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: נגדיר מיפוי בממד אינסופי $p \rightarrow \infty$. נגדיר כל כניסה באופן הבא:

$$\forall n \in \mathbb{N} \cup \{0\}: \phi_n(x) = \exp(-\gamma x^2) \frac{(2\gamma)^{n/2} x^n}{\sqrt{n!}}$$

ומתקיים כנדרש

$$\phi(x_i)^\top \phi(x_j) = \sum_{n=0}^{\infty} \exp(-\gamma x_i^2) \frac{(2\gamma)^{n/2} x_i^n}{\sqrt{n!}} \exp(-\gamma x_j^2) \frac{(2\gamma)^{n/2} x_j^n}{\sqrt{n!}}$$

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

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

עם זאת, בגלל קושי מובנה בפתרון ה-primal problem עם ה-feature mapping מהסעיף הקודם (חִשְבוּ 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: בפיתרון הבעיה נדרשים לחשב את ה-kernel בין כל הזוגות, ויש $O(10^8)$ כאלה.

כמו כן פונקציית המטרה בניסוח הדואלי תכיל $O(10^8)$ מחוברים.

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

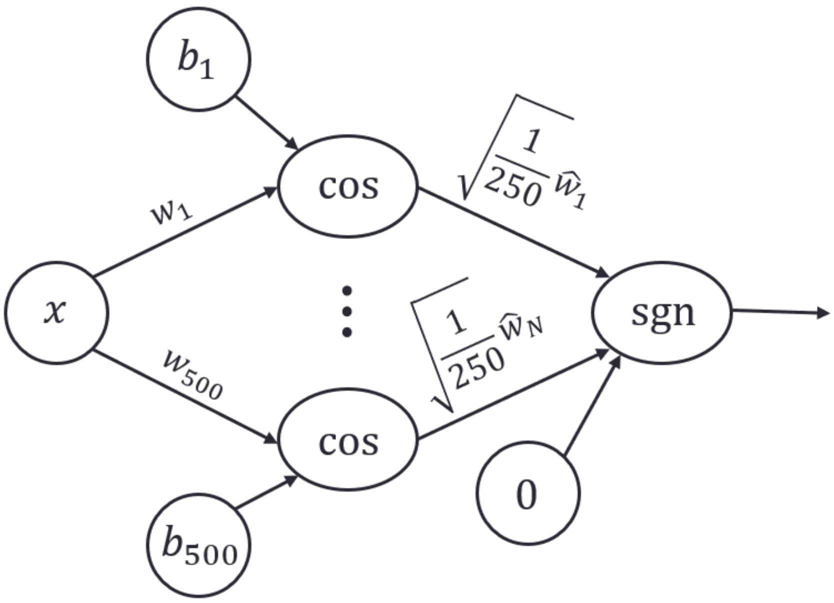
תשובה: ניצור מיפוי $\psi:\mathbb{R}\to\mathbb{R}^{500}$ לפי $\forall n\in[500]: \psi_n(x) = \frac{\cos(w_n x + b_n)}{\sqrt{250}}$.

נשים לב שמתקיים כנדרש $\psi(x_i)^\top\psi(x_j) = \frac{1}{250}\sum_{n=1}^{500}\cos(w_n x_i + b_n)\cdot\cos(w_n x_j + b_n) = K'(x_i,x_j)$.

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 השמה לרשת נוירונים עם שכבה חבויה אחת, כך שפעולתה על $x\in\mathbb{R}$ שרירותי תהיה זהה לזו של המסווג 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.
הבהרה: בהשמה תוכלו להשתמש בקבועים, פונקציות ו/או בנתונים שהתקבלו עד כה:

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

פתרון: בתרשים. ניתן היה לוותר על הכפל ב-$\sqrt{\frac{1}{250}}$ בשכבה השנייה (חִשְבוּ מדוע).



שאלה 3: PAC learnability [29 נק']

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 $m\ \geq\ m_{\mathcal{H}}(\epsilon,\delta)$, האלגוריתם מחזיר היפותזה שהיא (ϵ,δ)-PAC. משמע,

$$\underbrace{\Pr_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}\big(A(S)\big)>\epsilon\right]\leq\delta}_{\text{השלימו}}$$

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 $m\ \geq\ m_{\mathcal{H}}(\epsilon)$
- האלגוריתם מחזיר היפותזה עם שגיאת הכללה חסומה, משמע: $L_{\mathcal{D}}\big(A(S)\big)\leq\epsilon$.

סמנו את האפשרות הנכונה בהכרח לגבי הטענה שלעיל.

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

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

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

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

(הסבר: לא ניתן להבטיח באופן <u>דטרמיניסטי</u> שרואים את כל הדוגמאות, זה מוסבר בתחילת תרגול 5)

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 <u>במרחב</u> $\mathcal{X}$ 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:
- אם $x_+ \in S$ אזי $h_S = h_{x+}$
- אחרת, $h_S = h^0$
- נחזיר את h_S

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

הסבר: שגיאת <u>האימון</u> של האלגוריתם המוצע היא תמיד 0.

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

$$L_{\mathcal{D}}(h_S) = \begin{cases} 0, & x_+ \in S \\ \mathcal{D}(x_+), & x_+ \notin S \end{cases}$$

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 <u>בלבד</u>. תשובה שתלויה גם ב-$\mathcal{D},S,x_+$ (חלקם או כולם) תקבל ניקוד חלקי.
הראו את צעדי הפיתוח בצורה מנומקת.

תשובה ופיתוח: נתחיל את הפיתוח בהנחה שאנחנו יודעים את $\mathcal{D}(x_+)$ (זה לא המצב באופן כללי).

לפי הסעיף הקודם $L_{\mathcal{D}}(h_S)\geq\mathcal{D}(x_+)$, ולכן המצב $\mathcal{D}(x_+)>\epsilon>L_{\mathcal{D}}(h_S)$ בלתי אפשרי (ומספיק $m=1$).

כאשר $\epsilon\leq\mathcal{D}(x_+)$, ניתן לפתח במדויק את ההסתברות $\Pr_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(h_S)>\epsilon]$ אותה רוצים לחסום ע"י δ:

$$\Pr_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(h_S)>\epsilon]=\Pr_{S\sim\mathcal{D}^m}[x_+\notin S]=\left(1-\mathcal{D}(x_+)\right)^m$$

דרך א' (ניקוד חלקי):

$$\left(1-\mathcal{D}(x_+)\right)^m\overset{\text{נדרוש}}{\leq}\delta\iff m\underbrace{\ln\left(1-\mathcal{D}(x_+)\right)}_{<0}\leq\ln\delta\iff m\geq\frac{\ln(\delta)}{\ln(1-\mathcal{D}(x_+))}\Longrightarrow m_{\mathcal{H}_{\text{single}}}(\epsilon,\delta)=\left\lceil\frac{\ln(\delta)}{\ln(1-\mathcal{D}(x_+))}\right\rceil$$

דרך ב' (ניקוד מלא): בגלל שעוסקים במקרה ש-$\epsilon\leq\mathcal{D}(x_+)$, נדרוש דרישה חזקה יותר

$$\Pr_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(h_S)>\epsilon]=\left(1-\mathcal{D}(x_+)\right)^m\leq(1-\epsilon)^m\overset{!}{\leq}\delta\iff m\underbrace{\ln(\epsilon)}_{<0}\leq\ln\delta\iff m\geq\frac{\ln(\delta)}{\ln(1-\epsilon)}$$

כעת קל להיפטר מ-$\mathcal{D}(x_+)$ שאינו ידוע: $m_{\mathcal{H}_{\text{single}}}(\epsilon,\delta)=\left\lceil\frac{\ln(\delta)}{\ln(1-\epsilon)}\right\rceil\geq\begin{cases}1, & \epsilon>\mathcal{D}(x_+)\\ \frac{\ln(\delta)}{\ln(1-\epsilon)}, & \epsilon\leq\mathcal{D}(x_+)\end{cases}$.

שאלה 4: צירופים של מסווגים ליניאריים [27 נק']

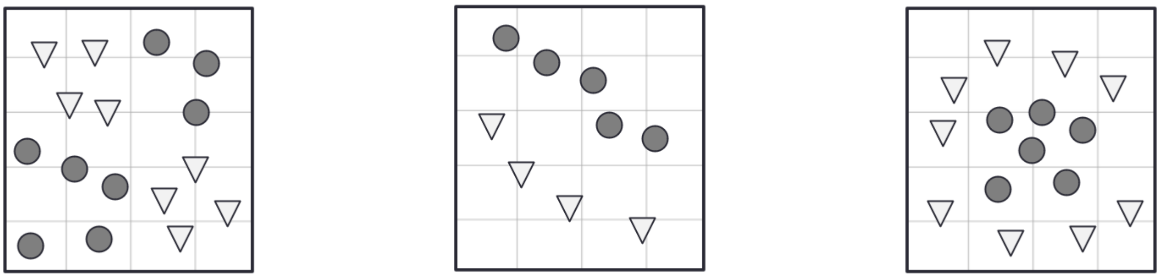
בכל השאלה מרחב הדוגמאות הוא $\mathcal{X} = \mathbb{R}^d$ ומרחב התיוגים הוא $\mathcal{Y} = \{-1, +1\}$ עבור $d \in \mathbb{N}$ שרירותי.

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

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

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.

(c) **(b)** (a)



הסבר: המסווג ליניארי, שהרי $\operatorname{sign}\left(\sum_{k=1}^K \alpha_k \left(\mathbf{w}_k^\top \mathbf{x} + b_k\right)\right) = \operatorname{sign}\left(\left(\sum_{k=1}^K \alpha_k \mathbf{w}_k\right)^\top \mathbf{x} + \sum_{k=1}^K \alpha_k b_k\right)$.

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: $h_\theta(\mathbf{x}) = \mathrm{sign}(\alpha \mathbf{w}^\top \mathbf{x})$.

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:

$$t\cdot\ell(\mathbf{w}_1,\alpha_1)+(1-t)\cdot\ell(\mathbf{w}_2,\alpha_2)\ge \ell(t\mathbf{w}_1+(1-t)\mathbf{w}_2,\ \ t\alpha_1+(1-t)\alpha_2\ )$$

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.

תשובה: כדי להוכיח צריך להראות בחירה אחת שלא מקיימת את אי-השיוויון.

למשל, נבחר $\alpha_1=1,\alpha_2=-1$ ו-$t=0.5$. נבחר $\mathbf{w}_1=\frac{1}{y\|\mathbf{x}\|^2}\mathbf{x}=-\mathbf{w}_2$.

מצד אחד, $\ell(t\mathbf{w}_1+(1-t)\mathbf{w}_2,\ \ t\alpha_1+(1-t)\alpha_2)=\ell(\mathbf{0},\ \ 0)=\max\big(0,\ \ 1-y(0\cdot\mathbf{0}^\top\mathbf{x})\big)=1$

ומצד שני, $t\ell(\mathbf{w}_1,\alpha_1)+(1-t)\ell(\mathbf{w}_2,\alpha_2)=\frac{\ell(\mathbf{w}_1,1)+\ell(-\mathbf{w}_1,-1)}{2}=\frac{2\max(0,\ \ 1-y\mathbf{w}_1^\top\mathbf{x})}{2}=\max\left(0,\ \ 1-\underbrace{\frac{y\mathbf{x}^\top\mathbf{x}}{y\|\mathbf{x}\|^2}}_{=1}\right)=0$

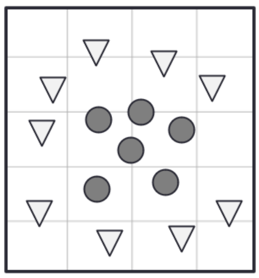
הערה: לא תיתכן דוגמה נגדית שבה $\alpha_1=\alpha_2$ או $\mathbf{w}_1=\mathbf{w}_2$, שהרי בכל אחד מהפרמטרים האלה בנפרד הפונקציה קמורה.

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

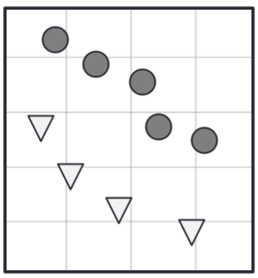
$$\mathcal{H}_2^{(K)}=\left\{h_\theta\ \middle|\ \theta=\left(\underbrace{\mathbf{w}_1,\dots,\mathbf{w}_K}_{\in\mathbb{R}^d},\ \underbrace{b_1,\dots,b_K}_{\in\mathbb{R}},\ \underbrace{\boldsymbol{\alpha}}_{\in\mathbb{R}^K}\right)\right\},\quad \text{where}\quad h_\theta(\mathbf{x})=\mathrm{sign}\left(\textstyle\sum_{k=1}^K\alpha_k\,\mathrm{sign}\big(\mathbf{w}_k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

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.

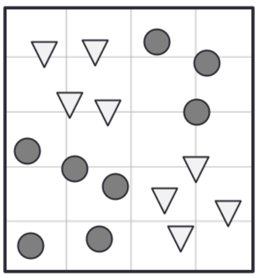
(a)



(b)



(c)



הסבר קצר: זוהי מחלקה עשירה מאוד. כל אחד מהרכיבים $\mathrm{sign}(\mathbf{w}_k^\top\mathbf{x}+b_k)$ הוא מפריד ליניארי.

לכן מדובר ב-ensemble של מסווגים ליניאריים (יותר עשירים מ-decision stumps שראינו ב-Adaboost).

אפשר גם להסתכל על המחלקה כרשתות נוירונים עם שכבה חבויה אחת ואקטיבציה מסוג threshold

וראינו בהרצאה על למידה עמוקה הדגמות שמראות שזה מאפשר לעשות fit לדאטה מורכב.

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.

<u>הקלה:</u> 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 <u>סופית</u> בלבד (לרשותכם דפי טיוטה בסוף הגיליון).

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

$$=\begin{cases} 0, & 1<y_i\left(\sum_{k=1}^K \alpha_k\operatorname{sign}\left(\mathbf{w}_k^\top\mathbf{x}_i+b_k\right)\right)\\ -y_i\operatorname{sign}\left(\mathbf{w}_j^\top\mathbf{x}_i+b_j\right), & 1\ge y_i\left(\sum_{k=1}^K \alpha_k\operatorname{sign}\left(\mathbf{w}_k^\top\mathbf{x}_i+b_k\right)\right)\end{cases}$$

b. עבור $j\in[K]$ נתון, כתבו את הגרדיינט לפי $\mathbf{w}_j$. נדרשת תשובה <u>סופית</u> בלבד.

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

$$=\begin{cases} \mathbf{0}_\mathrm{d}, & 1<y_i\left(\sum_{k=1}^K \alpha_k\operatorname{sign}\left(\mathbf{w}_k^\top\mathbf{x}_i+b_k\right)\right)\\ \nabla_{\mathbf{w}_j}(1)-y_i\alpha_j\nabla_{\mathbf{w}_j}\operatorname{sign}\left(\mathbf{w}_j^\top\mathbf{x}_i+b_j\right), & 1\ge y_i\left(\sum_{k=1}^K \alpha_k\operatorname{sign}\left(\mathbf{w}_k^\top\mathbf{x}_i+b_k\right)\right)\end{cases}$$

$$=\begin{cases} \mathbf{0}_\mathrm{d}, & 1<y_i\left(\sum_{k=1}^K \alpha_k\operatorname{sign}\left(\mathbf{w}_k^\top\mathbf{x}_i+b_k\right)\right)\\ \mathbf{0}_\mathrm{d}-y_i\alpha_j\underbrace{\frac{\partial\operatorname{sign}(\mathrm{z})}{\partial\mathrm{z}}}_{=0}\nabla_{\mathbf{w}_j}\left(\mathbf{w}_j^\top\mathbf{x}_i+b_j\right), & 1\ge y_i\left(\sum_{k=1}^K \alpha_k\operatorname{sign}\left(\mathbf{w}_k^\top\mathbf{x}_i+b_k\right)\right)\end{cases}=\mathbf{0}_\mathrm{d}$$

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

תשובה: הגרדיינטים לפי וקטורי המשקל $\mathbf{w}_k$ (וגם לפי ה-biases b_k) יהיו תמיד אפסים.