הטכניון – מכון טכנולוגי לישראל

הפקולטה למדעי המחשב

מבוא למערכות לומדות (236756)

סמסטר אביב תשפ"ב – 12 ביולי 2022

מרצה: ד"ר ניר רוזנפלד

מבחן מסכם מועד א'

הנחיות הבחינה:

- **משך הבחינה:** 3 שעות.
- **חומר עזר:** המבחן בחומר סגור (ללא ספרים, מחברות, דפי נוסחאות).
- מותר להשתמש במחשבון.
- יש לכתוב בעט **בלבד**.
- יש לכתוב את התשובות **על גבי שאלון זה**.
- מותר לענות בעברית או באנגלית.
- קריאות:
  - תשובה בכתב יד לא קריא – **לא תיבדק**.
  - בשאלות רב-ברירה – הקפידו להקיף את התשובות בבירור. סימונים לא ברורים יביאו 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.**

מבנה הבחינה:

- **חלק א' [76 נק']:** 4 שאלות פתוחות.
- **חלק ב' [24 נק']:** 4 שאלות סגורות (אמריקאיות) [כל אחת 6 נק'].

בהצלחה!

חלק א' – שאלות פתוחות [76 נק']

שאלה 1: Feature selection and Classification models [20 נק']

נתון סֵט אימון עם $d \geq 3$ מאפיינים (features) ועם תיוגים בינאריים, משמע $\boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \{-1,1\}$.

בסט האימון אין שתי דוגמאות זהות.

נתון שהקורלציה בין שני מאפיינים מסוימים A, B בסט האימון היא בדיוק -1.

משמע, לכל דוגמה $\boldsymbol{x}_i$ מתקיים $x_i[B] = ax_i[A] + b$ עבור $a < 0$ וְ-$b \in \mathbb{R}$.

לומדים שני מסווגים:

- <u>בשלב הראשון</u>: לומדים מסווג על סט האימון המקורי ומחשבים עליו את דיוק האימון.
- <u>בשלב השני</u>:
  - מסירים את המאפיין B ומקבלים סט אימון מעודכן (שבו $d-1$ פיצ'רים).
  - מאמנים מסווג <u>חדש</u> על סט האימון המעודכן, ומחשבים עליו את דיוק האימון המעודכן.

עבור כל אלגוריתם למידה, סמנו האם דיוק האימון של המסווג <u>החדש</u> על סט האימון המעודכן זהה <u>בהכרח</u> לזה של המסווג <u>המקורי</u> על סט האימון המקורי.

הסבירו בקצרה את תשובותיכם (2-4 משפטים בכל סעיף).

הניחו שאין צעדים סטוכסטיים (אקראיים) בריצת האלגוריתמים ואין שגיאות נומריות (בפרט, בעיות קמורות מתכנסות לפיתרון האנליטי שלהן במדויק).

א. <u>kNN עם $k = 1$ (דוגמה לא נחשבת שכנה של עצמה).</u> דיוק האימון זהה בהכרח? **כן / לא**

הסבר: נניח שפיצ'ר B קובע את ~~המרחקים~~ הסיווגים של הנק' בגלל הערכים הגדולים שלו (יחסית) ולכן ~~הוא~~ הסרתו תשנה את השכנויות

ב. Hard-SVM ליניארי לא הומוגני בהנחה שהדאטה המקורי פריד. דיוק האימון זהה בהכרח? **כן** / **לא**

הסבר: הדאטה המקורי פריד לינארית (ובפרט הדאטה לאחר
הסרת הפיצ'ר ש.ישטן תלוי בפיצ'ר אחר

ג. Soft-SVM ליניארי לא הומוגני עם $\lambda = 1$ בהנחה שהדאטה המקורי פריד. דיוק האימון זהה בהכרח? **כן** / **לא**

הסבר: הפיצ'ר יכול להיות כבר שמוסיף "ממד" חדש בין הנקודות
השונות, מה שמאפשר לנו לנרמל נורמה קטנה יותר (רגולציה
משמעותית יותר) ובכך ישנה את דיוק האימון וישנה את המסווג (המפריד)

ד. ID3 המשתמש באנטרופיה ובונה עץ בעומק מירבי 4. דיוק האימון זהה בהכרח? **כן** / **לא**

הסבר: יש קורלציה שלילית בין הפיצ'ר שהסרנו לבין אחד הפיצ'רים שנשאר
ולכן העץ ישתמש במקום בפיצ'ר B, ויוכל להשתמש (ההפך) בפיצ'ר A
והדיוק ישאר זהה

שאלה 2: Kernel SVM [18 נק']

תזכורת: פונקציה מהווה קרנל חוקי אם ניתן לכתוב אותה בתור $K(u,v) = \langle \phi(u), \phi(v) \rangle = \sum_{i=1}^{n} \phi_i(u)\phi_i(v)$ עבור מיפוי כלשהו $\phi: \mathcal{X} \to \mathbb{R}^n$.

נתונות שתי פונקציות קרנל חוקיות $K_1, K_2: (\mathcal{X} \times \mathcal{X}) \to \mathbb{R}$ והמיפויים המתאימים להן $\phi, \psi: \mathcal{X} \to \mathbb{R}^n$. משמע, מתקיים:

$K_1(u,v) = \langle \phi(u), \phi(v) \rangle, \qquad K_2(u,v) = \langle \psi(u), \psi(v) \rangle$.

הוכיחו שהפונקציות הבאות מהוות קרנלים חוקיים (כפי שמוסבר בתזכורת לעיל).

בכל סעיף, הגדירו בבירור פונקציית מיפוי $\phi': \mathcal{X} \to \mathbb{R}^{n'}$ מתאימה והראו שמתקיים $K'(u,v) = \langle \phi'(u), \phi'(v) \rangle$ כנדרש.

א. [5 נק'] הפונקציה $K'(u,v) = \left(K_1(u,v)\right)^2$.

הוכחה:

$K'(u,v) = (K_1(u,v))^2 = \langle \varphi(u), \varphi(v) \rangle^2 =$

$= \left(\sum_{i=1}^{n} \varphi_i(u)\varphi_i(v)\right)^2 = \sum_{i=1}^{n} \varphi_i(u)\varphi_i(v) \sum_{j=1}^{n} \varphi_j(u)\varphi_j(v) =$

$= \sum\sum (\varphi_i(u)\varphi_j(u))(\varphi_i(v)\varphi_j(v)) \Rightarrow$

$\Rightarrow \varphi'(u) = ((\varphi_1(u)\varphi_1(u), \varphi_1(u)\varphi_2(u), \ldots, \varphi_1(u)\varphi_n(u), \varphi_2(u)\varphi_1(u), \ldots, \varphi_2(u)\varphi_n(u), \ldots, \varphi_1(u)\varphi_n(u), \ldots, \varphi_n(u)\varphi_n(u))$

$\varphi: X \to \mathbb{R}^{n^2}$

ניתן להפחית את המימד אך בגלל ישנה יותרמדי את הסעוריין

ב. [5 נק'] הפונקציה $K'(u,v) = K_1(u,v) + 3 \cdot K_2(u,v) + 1$.

הוכחה:

$K'(u,v) = \sum_{i=1}^{n} \varphi_i(u)\varphi_i(v) + 3\sum_{i=1}^{n} \psi_i(u)\psi_i(v) + 1 =$

$= \left(\sum_{i=1}^{n} \varphi_i(u)\varphi_i(v) + 3\psi_i(u)\psi_i(v)\right) + 1 \Rightarrow \varphi'(u) = [\varphi(u), \sqrt{3}\psi(u), 1]$

$\varphi': X \to \mathbb{R}^{2n+1}$

$(uv)^2 + 2uv + 1$

ג. [8 נק'] אילו מהפעולות הבאות אמורות להפחית overfitting ב-Kernel SVM?

סמנו את כל התשובות המתאימות (השאלה אינה עוסקת במקרי קצה אלא במקרה הסביר).

$(uv)^2$

a. לעבור מפונקציית הקרנל $K(\boldsymbol{u},\boldsymbol{v})=(\boldsymbol{u}^\top\boldsymbol{v})^p$ לפונקציית הקרנל $K(\boldsymbol{u},\boldsymbol{v})=(\boldsymbol{u}^\top\boldsymbol{v}+1)^p$, כאשר $p\in\mathbb{N}_{\geq 2}$.

b. להקטין את השונות σ^2 של קרנל RBF (שמוגדר בתור $K(\boldsymbol{u},\boldsymbol{v})=\exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{u}-\boldsymbol{v}\|_2^2\right\}$).

$\sigma \to 0 \to -\infty \to e^{-\infty}$ 1 1 1NN

c. לפתור (במדויק) את הבעיה ה-primal במקום לפתור (במדויק) את הבעיה ה-dual.

$\sigma \to \infty \to 0 \to e^{0}$ 1 → Majority

d. להגדיל את מקדם הרגולריזציה λ (בהתאמה: להקטין את C).

e. להגדיל את סֶט האימון (באופן $i.i.d.$ מאותה התפלגות של הדאטה המקורי).

f. להגדיל את סֶט המבחן (באופן $i.i.d.$ מאותה התפלגות של הדאטה המקורי).

שאלה 3: VC-dimension [16 נק']

נגדיר את מחלקת ההיפותזות של מסווגים ליניאריים הומוגניים ב-d ממדים:

$$\mathcal{X} = \mathbb{R}^d,\ \ \mathcal{H}_{\text{lin}}^d = \{\boldsymbol{x} \mapsto \text{sign}(\boldsymbol{w}^\top \boldsymbol{x}):\ \boldsymbol{w} \in \mathbb{R}^d\}.$$

כדי למנוע טעויות, נקבע שלאורך כל השאלה $\text{sign}(0) = 0$ (ואחרת הפונקציה מחזירה ± 1 בהתאם לסימן).

צ: $\forall y_1, \dots, y_d \quad \exists h \in H \quad \forall i: y_i = h(x_i)$

א. [7 נק'] הוכיחו שמתקיים $\text{VCdim}(\mathcal{H}_{\text{lin}}^d) \geq d$.

הוכחה:

ויהיו $e_1 = [1,0,\dots,0]\ e_2 = [0,1,0,\dots,0], \dots, e_d = [0,\dots,0,1]$

נקח אותם C' כאלו, יהי $y_1, \dots, y_d \in \{\pm 1\}^d$ התיוגים שלהן

נגדיר $w = (y_1, \dots, y_d)$ ונשים לב כי

$\forall i:$ $w^T x_i = (y_1, \dots, y_i, \dots, y_d)\begin{pmatrix} 0 \\ \vdots \\ 1_i \\ 0 \\ \vdots \\ 0 \end{pmatrix} = y_i$

$h(x) = \text{Sing}(w^T x) = \text{Sign}(y_i) = y_i$ as wanted

H shatters C, $|C| = d \Rightarrow VCdim(\mathcal{H}_{lin}^d) \geq d$

ב. [9 נק'] הוכיחו שממד ה-VC הוא בדיוק d על ידי כך שתוכיחו שמתקיים $\mathrm{VCdim}(\mathcal{H}_{\mathrm{lin}}^{d}) < d+1$.

<u>רמז</u>: כל אוסף $\boldsymbol{x}_1, \dots, \boldsymbol{x}_d, \boldsymbol{x}_{d+1}$ של $d+1$ וקטורים כלשהם ב-$\mathbb{R}^d$ הינו תלוי ליניארית, ובהכרח אחד הווקטורים באוסף (בה"כ $\boldsymbol{x}_{d+1}$) מקיים $\boldsymbol{x}_{d+1} = \sum_{i=1}^{d} z_i \boldsymbol{x}_i$ עבור אוסף סקלרים $z_1, \dots, z_d \in \mathbb{R}$ שלפחות אחד מהם שונה מאפס.

הוכחה (לרשותכם דפי טיוטה בסוף הגיליון):

יהי $x_1, \dots, x_{d+1}$ אוסף נקו' כלשהו

מהרמז $x_{d+1} = \sum_{i=1}^{d} z_i x_i$ כאשר $z_1, \dots, z_d \in \mathbb{R}$ סקלרים שלפחות אחד מהם

שונה מ-0. נניח כי H מנתצת את הנק' הנ"ל

כלומר $\forall y_1, \dots, y_{d+1} \in \{+,-\} \ \exists h \in H \ \forall i: h(x_i) = y_i$

נגדיר את התיוג הבא: $y_i = \begin{cases} \mathrm{sign}(z_i) & i \le d \\ -1 & i = d+1 \end{cases}$

H מנתצת את אוסף הנק' הנ"ל ולכן קיים $w \in \mathbb{R}^{d+1}$ כך ש- $\forall i: \mathrm{sing}(w^T x_i) = y_i$

$$-1 = \mathrm{Sign}(w^T x_{d+1}) = \mathrm{Sign}\left(w^T \sum_{i=1}^{d} z_i x_i\right) = \mathrm{Sign}\left(\sum z_i (w^T x_i)\right) \ge 0 \Rightarrow \mathrm{Sign}(\,) = 1$$

אותו סימן ≥ 0

~~בסת~~ סתירה! ולכן H לא מנתצת את אוסף הנק' הנ"ל $\Rightarrow \mathrm{VCdim}(H) < d+1$

Initialize $D^{(1)} = \left(\frac{1}{m}, \ldots, \frac{1}{m}\right)$

For t=1, …, T:

$$h_t = \mathcal{A}\left(S, D^{(t)}\right)$$

$$\epsilon_t = \sum_i D_i^{(t)} \cdot \mathbf{1}_{h_t(x_i) \neq y_i}$$

$$\alpha_t = \frac{1}{2}\log\left(\frac{1}{\epsilon_t} - 1\right)$$

$$D_i^{(t+1)} = \frac{1}{Z_t} D_i^{(t)} \exp\left(-\alpha_t y_i h_t(x_i)\right)$$

$$h_s(x) = \operatorname{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

שאלה 4: Bagging and Boosting [22 נק']

משמאל מופיע אלגוריתם AdaBoost.

א. [5 נק'] שימו לב שבכל איטרציה האלגוריתם מעדכן את ההתפלגות באופן $D_i^{(t+1)} = \frac{1}{Z_t} D_i^{(t)} \exp\left(-\alpha_t y_i h_t(x_i)\right)$

כאשר $Z_t \triangleq \sum_{j=1}^{m} D_j^{(t)} \exp\left(-\alpha_t y_j h_t(x_j)\right)$ הוא גורם נירמול.

הוכיחו שמתקיים $D_i^{(t+1)} = c \cdot \exp\left(-y_i \sum_{k=1}^{t} \alpha_k h_k(x_i)\right)$ עבור קבוע כלשהו $c > 0$ (שתלוי ב-t).

הוכחה (לרשותכם דפי טיוטה בסוף הגיליון):

בסיס:

$$D_i^{(2)} = \underbrace{\frac{1}{Z_1} D_i^{(1)}}_{c} e^{-\alpha_1 y_i h_1(x_i)}$$

נניח נכונות עבור t ונראה נכונות עבור $t+1$:

$$D_i^{(t)} = C \cdot e^{-y_i \sum_{k=1}^{t-1} \alpha_k h_k(x_i)} = C \cdot e^{-\alpha_1 h_1(x_i) y_i} \cdot \ldots \cdot e^{-\alpha_{t-1} h_{t-1}(x_i) y_i}$$

$$D_i^{(t+1)} = \frac{1}{Z_t} D_i^{(t)} \cdot e^{-\alpha_t y_i h_t(x_i)} = C' \cdot e^{-y_i \alpha_1 h_1(x_i)} \cdot \ldots \cdot e^{-y_i \alpha_t h_t(x_i)} =$$

$$C' = \frac{C}{Z_t}$$

$$= C' e^{-y_i \sum_{k=1}^{t} \alpha_k h_k(x_i)}$$

כנדרש

בסעיף הבא נוכיח שהבחירה של האלגוריתם ב- $\alpha_t = \frac{1}{2}\log\left(\frac{1}{\epsilon_t} - 1\right)$ הינה אופטימלית.

<u>תזכורת</u>: הגדרנו בתרגול את ההיפותזה ה-unthresholded ה"חזקה" באיטרציה t בתור: $h_s^{(t)} = \sum_{k=1}^{t} \alpha_k h_k(x_i)$,

והראינו שקיים קבוע חיובי C כך שערך ה-exp. loss של ההיפותזה הוא:

$$\mathcal{L}_{\exp}\left(h_s^{(t)}\right) \triangleq \frac{1}{m}\sum_{i=1}^{m} \exp\{-y_i \sum_{k=1}^{t} \alpha_k h_k(x_i)\} = \underbrace{C}_{>0} \cdot \left(e^{-\alpha_t} + \left(e^{\alpha_t} - e^{-\alpha_t}\right) \underbrace{\sum_i D_i^{(t)} 1_{h_t(x_i)\neq y_i}}_{R}\right),$$

כאשר C אינו תלוי בבחירה של המקדם α_t שנבחר בזמן t.

ב. [5 נק'] הוכיחו שבהינתן h_t וְ- $\left\{D_i^{(t)}\right\}_{i\in[m]}$, הבחירה בְּ- $\alpha_t = \frac{1}{2}\log\left(\frac{1}{\epsilon_t} - 1\right)$ ממזערת את $\mathcal{L}_{\exp}\left(h_s^{(t)}\right)$.

<u>הערה</u>: שימו לב ש- $\epsilon_t \triangleq \sum_i D_i^{(t)} \cdot \mathbf{1}_{h_t(x_i)\neq y_i}$ והוא קבוע בהינתן h_t וְ- $\left\{D_i^{(t)}\right\}_{i\in[m]}$.

הוכחה:

$$\nabla_{\alpha_t} L_{exp} = C\left(-e^{-\alpha_t} + \left(e^{\alpha_t} + e^{-\alpha_t}\right)\varepsilon_t\right) = 0 \quad / :C$$

$$\varepsilon_t e^{-\alpha_t} - e^{-\alpha_t} + \varepsilon_t e^{\alpha_t} = 0 \quad / :$$

$$e^{-\alpha_t}(\varepsilon_t - 1) + \varepsilon_t e^{\alpha_t} = 0 \quad / \cdot e^{\alpha_t}$$

$$\varepsilon_t - 1 + \varepsilon_t e^{2\alpha_t} = 0$$

$$e^{2\alpha_t} = \frac{1-\varepsilon_t}{\varepsilon_t} \Rightarrow 2\alpha_t = \ln\left(\frac{1}{\varepsilon_t} - 1\right) \Rightarrow$$

$$\Rightarrow \alpha_t = \frac{1}{2}\ln\left(\frac{1}{\varepsilon_t} - 1\right)$$

L קמורה כסכום של קמורות (e קמורה) ולכן α_t נק' מינימום

ולכן α_t כנ"ל ממזערת את השגיאה.

הסעיף הבא לא קשור לאלגוריתם AdaBoost אלא לשיטת Ensemble שממשקלת היפותזות באופן כללי.

ג. [12 נק'] נתון סֶט אימון $\{(x_i, y_i)\}_{i=1}^{m}$ עם סיווגים בינאריים ($y_i \in \{-1, +1\}$).
צוות של 50 מומחים ומומחיות הגדיר ידנית 50 היפותזות בינאריות $h_1, \dots, h_{50}: \mathcal{X} \to \{-1, +1\}$.
המטרה שלנו היא למשקל אותן, משמע, ליצור היפותזה $h_\alpha(x) = \text{sign}\left(\sum_{k=1}^{50} \alpha_k h_k(x)\right)$.

כדי למנוע טעויות, נקבע שלאורך כל השאלה $\text{sign}(0) = 0$ (ואחרת הפונקציה מחזירה ± 1 בהתאם לסימן).

<u>נתון</u>: קיימים משקלים $\alpha_1^*, \dots, \alpha_{50}^* \in \mathbb{R}$ כך שהשגיאה האמפירית (שגיאת האימון) היא אפס.

כעת, נרצה <u>ללמוד</u> אוסף משקלים $\alpha_1, \dots, \alpha_{50} \in \mathbb{R}$.

(i) נסחו בעיית אופטימיזציה מתאימה **וקמורה** (ביחס למשקלים הנלמדים $\alpha_1, \dots, \alpha_{50}$) כך שהפיתרון שלה ממזער את השגיאה האמפירית.

תשובה (לרשותכם טיוטה בסוף הגיליון):

$h(x) = [h_1(x), \dots, h_{50}(x)]^T \in \{\pm 1\}^{50}$ $\quad \alpha = [\alpha_1, \dots, \alpha_{50}] \in R^{50}$

$$\underset{\alpha \in R^{50}}{\text{argmin}} \sum_{i=1}^{m} \max\{0, 1 - y_i \alpha^T h(x_i)\}$$

(ii) נמקו <u>בקצרה</u> מדוע הבעיה שהגדרתם קמורה (מומלץ להשתמש בטענות מההרצאה ומהתרגול).

תשובה (לרשותכם טיוטה בסוף הגיליון):

$1-z$ ישר ⇐ קמור, 0 ← קמור, max של קמורות ⇐ קמור

סכום של קמורות ⇐ קמור ⇐ לכן הבעיה קמורה

(iii) יהי אוסף המשקלים $\alpha_1, \ldots, \alpha_{50} \in \mathbb{R}$ פיתרון (אופטימלי) של הבעיה הקמורה <u>שהגדרתם</u> לעיל.

הוכיחו שהשגיאה האמפירית של פיתרון זה היא מינימלית.

הוכחה:

יהי $\alpha^* = [\alpha_1, \ldots, \alpha_{50}]$ ש. $\forall i: \text{sign}(\alpha^{*T} h(x_i)) = y_i \in \{\pm 1\}$

ולכן $y_i \alpha^{*T} h(x_i) > 0$ נסמן $(= R)$

נגדיר $\beta = \frac{2}{R} \Leftarrow \forall i: y_i \beta \alpha^{*T}(h(x_i)) = 2 > 1$

כלומר ה"מ פתרון $\beta\alpha^*$ המאפס את ה objective

של ה hinge loss וכיוון ש ה hinge loss הוא ה surrogate ל

השגיאה האמפירית (בפרט) גם השגיאה האמפירית תתאפס

עבור כל פתרון אופטימלי של הבעיה הנ"ל

חלק ב' – שאלות אמריקאיות [24 נק']

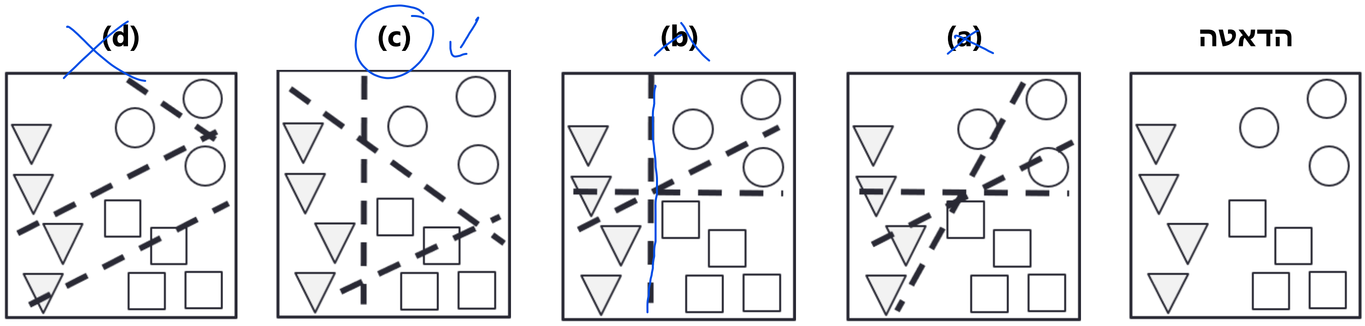
בשאלות הבאות סמנו את התשובות המתאימות (לפי ההוראות). בחלק זה אין צורך לכתוב הסברים.

א. [6 נק'] נתון דאטה דו-ממדי עם שלושה תיוגים אפשריים – משולש, מעגל וריבוע.

מפעילים One-vs-All (OVA) עם פרספטרון בתור מסווג בסיס. לכל בעיה בינארית מריצים את הפרספטרון מספיק איטרציות כך שאם הבעיה פרידה ליניארית – הפרספטרון יתכנס (חישבו מה קורה אחרת).

בכל תרשים מצוירים שלושה גבולות החלטה שאמורים לתאר את הגבולות שהתקבלו ע"י מסווגי הבסיס (הבינאריים).

הקיפו את האות שמתאימה לתרשים <u>היחיד</u> שמתאר גבולות החלטה שיכולים להתקבל ע"י השיטה שתוארה לעיל.



ב. [6 נק'] נתון דאטה d-ממדי $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m}$ עם סיווגים בינאריים (± 1).

סמנו את <u>כל</u> הטענות הנכונות **בהכרח** ביחס **לשיטות להורדת ממד** לממד נמוך כלשהו.

<u>הבהרה</u>: זו שאלת חשיבה ולא שאלת טריוויה.

a. הדאטה המקורי פריד ליניארית ⟸ הדאטה פריד ליניארית לאחר הורדת ממד ליניארית עם PCA.

b. הדאטה המקורי פריד ליניארית ⟸ בהסתברות גבוהה, הדאטה פריד ליניארית לאחר הורדת ממד ליניארית עם Random Projections.

c. הדאטה המקורי פריד ליניארית ⟸ הדאטה פריד ליניארית לאחר הורדת ממד עם Kernel-PCA (עם קרנל לא ליניארי).

d. הדאטה פריד ליניארית <u>לאחר</u> הורדת ממד ליניארית עם PCA ⟸ הדאטה המקורי פריד ליניארית.

e. PCA היא שיטה Supervised בעוד ש-Random Projections היא שיטה Unsupervised.

ג. [6 נק'] נתונה בעיית רגרסיה לינארית (עבור $\lambda \geq 0,\ p \in \{0.5, 1, 2\}$):

$$\operatorname{argmin}_{\boldsymbol{w}} \left(\frac{1}{m}\sum_{i=1}^{m}(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2 + \lambda\|\boldsymbol{w}\|_p^p\right) = \operatorname{argmin}_{\boldsymbol{w}} \left(\frac{1}{m}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{w}\|_p^p\right).$$

סמנו את כל הטענות הנכונות.

$2\frac{1}{m}X^\top|$ $|$

$2\frac{1}{m}X^\top X$

הערה: בכל הסעיפים, שואלים על קמירות ביחס ל-$\boldsymbol{w}$.

תזכורת: $\|\boldsymbol{w}\|_p^p \triangleq \sum_{k=1}^{d}|w_k|^p$.

a. כאשר $\lambda = 0$: ההסיאן הוא $\frac{2}{m}\boldsymbol{X}^\top\boldsymbol{X}$ שהיא מטריצה PSD אבל לא בהכרח PD, ולכן הבעיה לא בהכרח קמורה.

b. כאשר $\lambda > 0$ וגם $p = 2$: הבעיה קמורה בהכרח.

c. כאשר $\lambda > 0$ וגם $p = 1$: הבעיה קמורה בהכרח.

d. כאשר $\lambda > 0$ וגם $p = 0.5$: הבעיה קמורה בהכרח.

e. כאשר $\lambda > 0$ וגם $p = 1$: הבעיה אינה גזירה ביחס ל-$\boldsymbol{w}$ ולכן לא ניתן להפעיל שיטות גרדיינט (או סאב-גרדיינט).

ד. [6 נק'] רוצים לאמן רשת נוירונים בתצורת Autoencoder.

נסמן את אוסף כל הפרמטרים של הרשת בתור Θ. לכל דוגמה $\boldsymbol{x}_i \in \mathbb{R}^d$ נסמן את הפלט של הרשת בתור $\hat{\boldsymbol{x}}_i \in \mathbb{R}^d$.

בהרצאה ראינו שניתן לאמן את הרשת ע"י פיתרון הבעיה הבאה: $\operatorname{argmin}_\Theta \left(\lambda\|\Theta\|_2^2 + \frac{1}{m}\sum_{i=1}^{m}\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|_2^2\right)$.

אילו מבין הבעיות הבאות עשויות גם כן להתאים לאימון רשת Autoencoder (עבור $\lambda > 0$)?

סמנו את כל התשובות המתאימות.

a. $\operatorname{argmin}_\Theta \left(\lambda\|\Theta\|_2^2 + \frac{1}{m}\sum_{i=1}^{m}\|\boldsymbol{x}_i + \hat{\boldsymbol{x}}_i\|_2^2\right)$

b. $\operatorname{argmin}_\Theta \left(\lambda\|\Theta\|_2^2 + \frac{1}{m}\sum_{i=1}^{m}\exp\{\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|_2^2\}\right)$

c. $\operatorname{argmin}_\Theta \left(\lambda\|\Theta\|_2^2 - \frac{1}{m}\sum_{i=1}^{m}\|\boldsymbol{x}_i\|_2^2 + \frac{1}{m}\sum_{i=1}^{m}\|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|_2^2\right)$

d. $\operatorname{argmin}_\Theta \left(\lambda\|\Theta\|_2^2 + \frac{1}{m}\sum_{i=1}^{m}\max\{0, 1 - \operatorname{sign}(\boldsymbol{x}_i^\top\hat{\boldsymbol{x}}_i)\}\right)$

e. $\operatorname{argmin}_\Theta \left(-\lambda\|\Theta\|_2^2 + \frac{1}{m}\sum_{i=1}^{m}\frac{\boldsymbol{x}_i^\top\hat{\boldsymbol{x}}_i}{\|\boldsymbol{x}_i\|_2\|\hat{\boldsymbol{x}}_i\|_2}\right)$

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):