מבוא למערכות לומדות (236756)

סמסטר חורף תשפ"ב – 10 בפברואר 2022

מרצה: ד"ר יונתן בלינקוב

מבחן מסכם מועד א'

הנחיות הבחינה:

- **משך הבחינה:** 3 שעות.
- **חומר עזר:** המבחן בחומר סגור (ללא ספרים, מחברות, דפי נוסחאות).
- אין צורך במחשבון.
- מותר לכתוב בעט או בעיפרון, כל עוד הכתב קריא וברור.
- מותר לענות בעברית או באנגלית.
- יש לכתוב את התשובות **על גבי שאלון זה** בכתב יד קריא. תשובה בכתב יד לא קריא – לא תיבדק.
- 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' [75 נק']:** 4 שאלות פתוחות.
- **חלק ב' [25 נק']:** 5 שאלות סגורות (אמריקאיות) [כל אחת 5 נק'].

בהצלחה!

חלק א' – שאלות פתוחות [75 נק']

שאלה 1 [16 נק']

חוקרת מהטכניון עובדת על בעיית סיווג בינארי כלשהי. ברשותה dataset שבו $m = 150$ דוגמאות שונות (distinct).

החוקרת הריצה שלושה מודלים, ולכל מודל ביצעה hyperparameter tuning:

overfit → under
complex → simple

(a) מודל: kNN (נק' נחשבת שכנה של עצמה), היפרפרמטר: מספר השכנים k, טווח: 1 עד 97.

(b) מודל: עץ החלטה, היפרפרמטר: עומק מירבי, טווח: 1 עד 40.

under → over
simple → complex

(c) מודל: Kernel SVM, היפרפרמטר: חוזק הרגולריזציה λ, טווח: 10^{-3} עד 10^{7}.

hard → soft
complex → simple
overfitted → underfitted

how

לכל מודל, היא ציירה גרף של דיוק האימון ודיוק ההכללה (בעזרת 5-fold cross validation) בציר y כפונקציה של ערך ההיפרפרמטר בציר x (הערכים גדלים משמאל לימין).

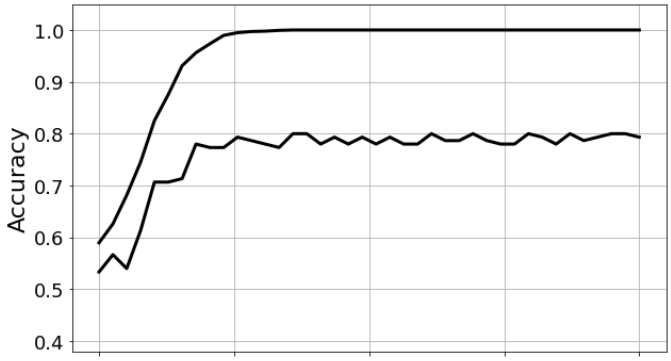
בעקבות תקלה, הכיתוב על ציר x נמחק מכל הגרפים.

בנוסף, הגרפים נשמרו בטעות בשחור לבן, כך שלא ניתן להבדיל בקלות בין דיוק האימון לדיוק ההכללה.

א. [8 נק'] במקום המתאים מתחת כל תרשים, כתבו את האות שמתאימה למודל ולהיפרפרמטר שיצרו אותו.

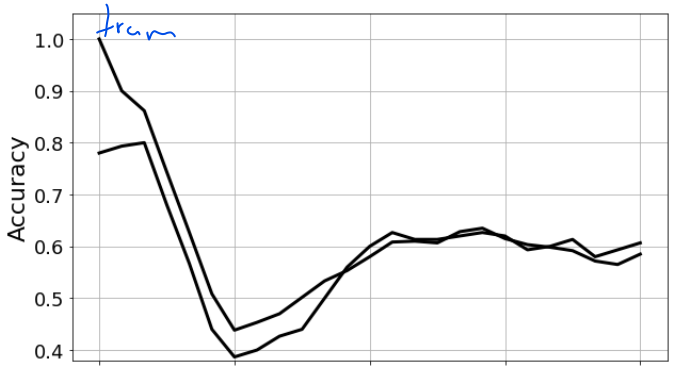
הערה: השאלה לא מוגדרת היטב ובדיעבד היו שתי תשובות נכונות לשאלה זו.

תרשים א'



B

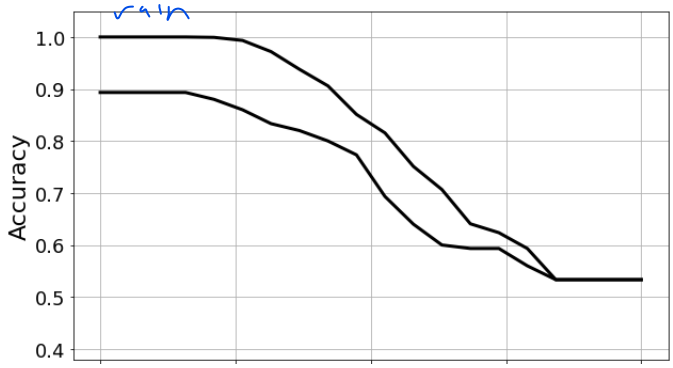
תרשים ב'



C/A

A

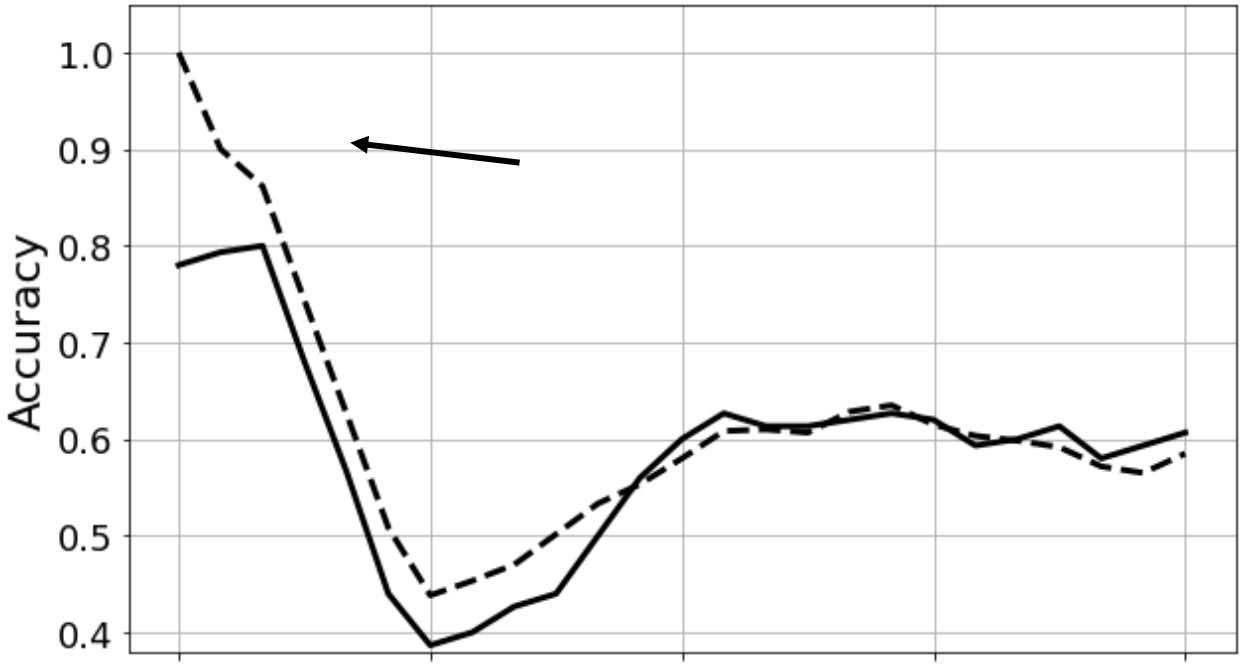
תרשים ג'



A/C

C

ב. [8 נק'] להלן תרשים ב' מוגדל.



הסתכלו על העקומה המְקוּוְקֶוֶת שבתרשים (מסומנת בחץ).

האם העקומה מתארת את דיוק האימון או את דיוק ההכללה?

הסבירו בקצרה. התבססו על התרשים ועל מאפייני המודל שיצר את עקומה זו (מבין שלושת המודלים).

תשובה תמציתית:

אימון

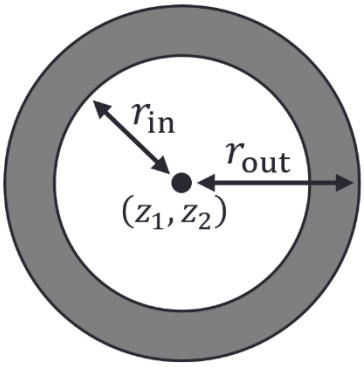
שאלה 2 – PAC learning [21 נק']

תהי $\mathcal{H}$ מחלקת היפותזות של Bagels/donuts בדו-ממד:

$$\mathcal{H} = \{h_\theta : \mathbb{R}^2 \to \pm 1 \mid \theta = (z_1, z_2, r_{\text{out}}, r_{\text{in}}),\ r_{\text{out}} > r_{\text{in}} \geq 0\}$$

כאשר היפותזה בודדת מוגדרת באופן הבא:

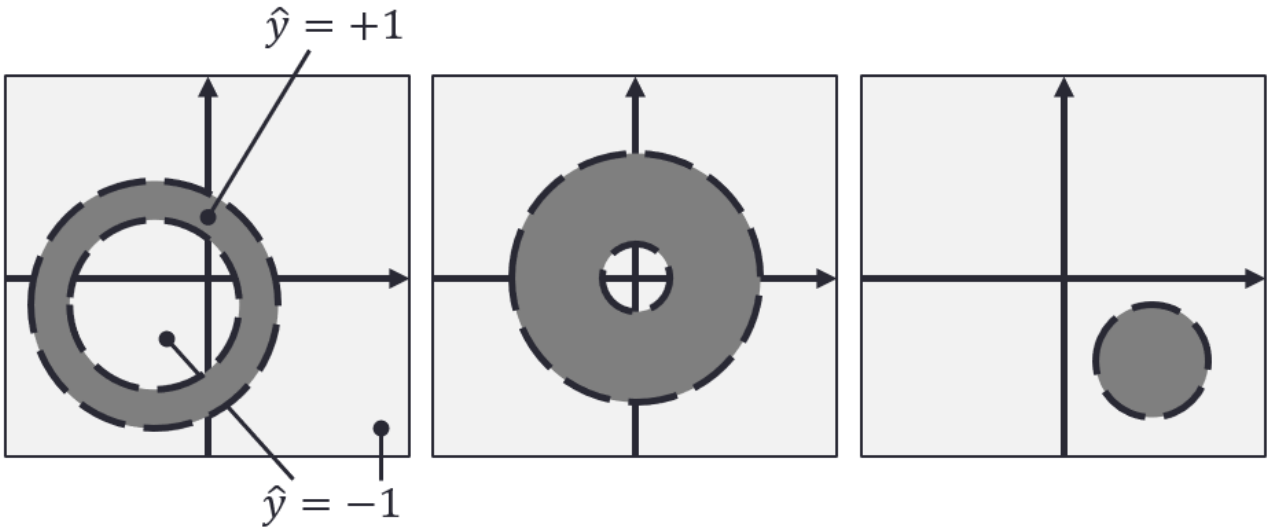
$$h_\theta(x) = \begin{cases} +1, & r_{\text{out}} \geq \sqrt{(x_1 - z_1)^2 + (x_2 - z_2)^2} \geq r_{\text{in}} \\ -1, & \text{otherwise} \end{cases}$$



דגשים לגבי <u>כל</u> היפותזה $h_\theta \in \mathcal{H}$:

- המרכזים של המעגלים <u>משותפים</u> ולא בהכרח בראשית הצירים.
- הרדיוס של המעגל הפנימי יכול להיות אפס.
  השטח שבתוך ה-donut לא יכול להיות אפס.
- האזור בין שני המעגלים מסווג כחיובי, והאזורים האחרים כשליליים.
- מדובר אך ורק במעגלים ולא באליפסות.

דוגמה לשלוש היפותזות מתוך $\mathcal{H}$:



א. [3 נק'] להלן ההגדרה של "ניתוץ". השמטנו מההגדרה את הַכַּמֻּתים.

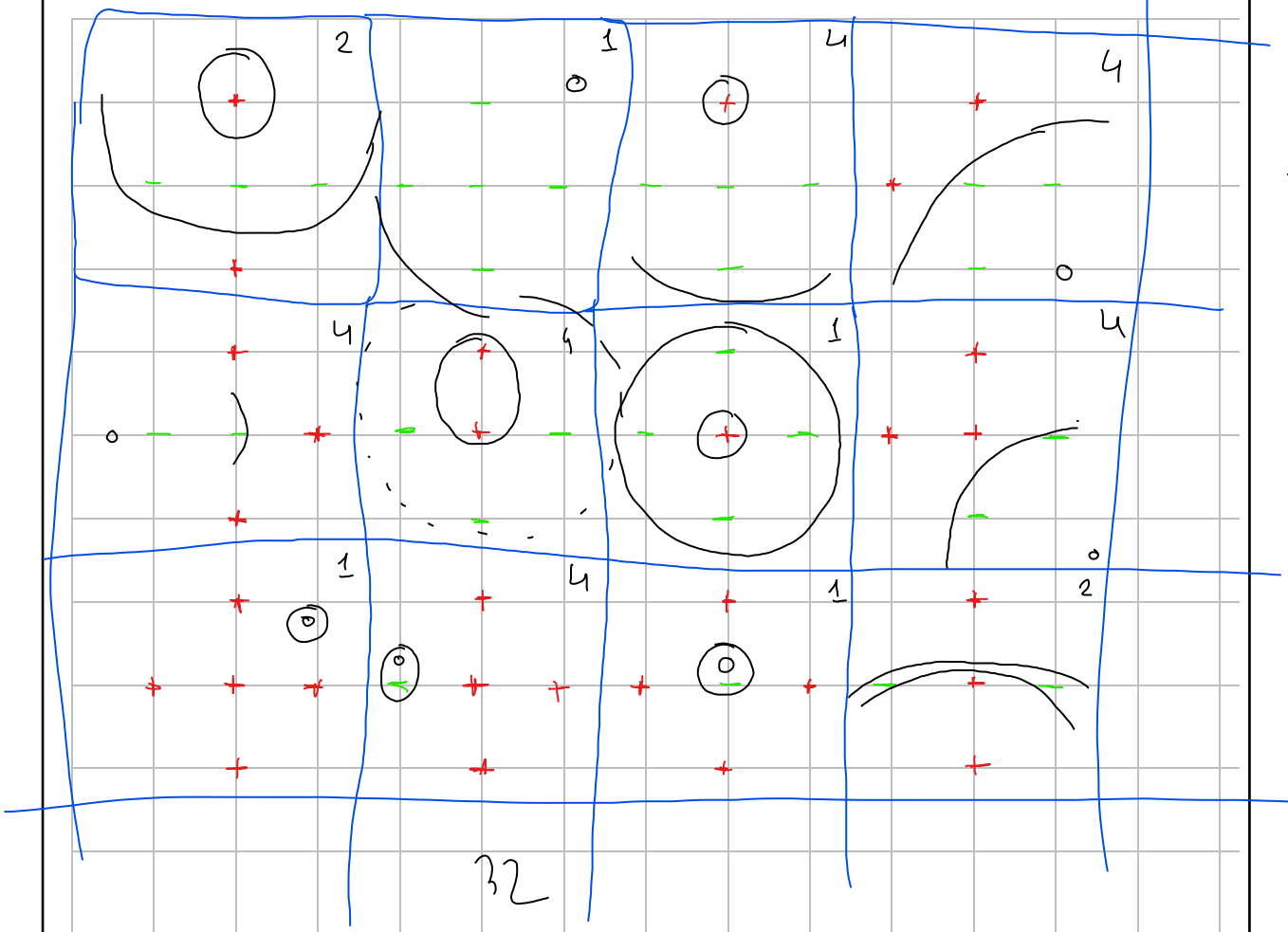
השלימו את שלושת הכמתים החסרים. בכל מקום כתבו האם חסר בהגדרה ∀ או ∃.

$$\mathcal{H} \text{ shatters } C \iff \underbrace{\forall}_{\text{השלימו}} y_1, \dots, y_{|C|} \in \mathcal{Y}: \underbrace{\exists}_{\text{השלימו}} h \in \mathcal{H}: \underbrace{\forall}_{\text{השלימו}} x_i \in C: h(x_i) = y_i$$

ב. [13 נק'] כתבו את החסם התחתון **ההדוק ביותר** שתוכלו למצוא לממד ה-VC: $\boxed{5}$ $\leq \mathrm{VCdim}(\mathcal{H})$.

הוכיחו את החסם התחתון שכתבתם (אין להוכיח שוויון). יש לכתוב הסבר מילולי תמציתי ולצרף תרשימים נדרשים.

הוכחה (לרשותכם דפי טיוטה בסוף הגיליון):



32

בעזרת הפסתי בין ה+ ל- אב אתו עוקרין

דמיינו שבמקום + כתבתי -

ובמקום - כתבתי +

32 הומב סה"כ

כיושר 12 מהכ "שונות"

5
10
10
5
1

$\binom{5}{2}=10$

24
28
3

1 ✓
2 ✓
3 ✓

$5\cdot4\cdot3 = 10$
$3\cdot2$

המשך סעיף ב':

$\frac{1}{2}$

ג. [5 נק'] חוקרת וחוקר רוצים לאמן מודל סיווג בינארי.

$\frac{5}{10}$ $\frac{10}{20}$

<u>החוקרת</u> משתמשת במחלקת ההיפותזות $\mathcal{H}$ שהגדרנו.

<u>החוקר</u> משתמש במחלקת היפותזות של donuts שמרכזם בראשית הצירים, משמע:

$$\mathcal{H}' = \{h_\theta \mid \theta = (0, 0, r_{\text{out}}, r_{\text{in}}),\ r_{\text{out}} > r_{\text{in}} \geq 0\} \subset \mathcal{H}.$$

מי צפוי להזדקק לפחות דוגמאות בתהליך הלמידה ע"מ להבטיח (במונחי PAC) שגיאת הכללה $\epsilon = 0.1$? נמקו בקצרה.

תשובה תמציתית:

$VC(H') \leq VC(H)$

$m_H(\varepsilon,\delta) = O\left(\frac{VC(H)\log(\frac{1}{\varepsilon}) + \log(\frac{1}{\delta})}{\varepsilon}\right)$

$H' \subseteq H \Rightarrow H \supseteq H'$

$H' \subseteq H$

החוקר יצטרך לדגום דוגמאות ע

שאלה 3 – רגרסיה ליניארית ו-Generative models [21 נק']

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

<u>שימו לב</u>: הדוגמאות $\boldsymbol{x}_i \in \mathbb{R}^d$ והתיוגים $y_i \in \mathbb{R}$ נתונים. וְקְטור המשקלים $\boldsymbol{w} \in \mathbb{R}^d$ לא ידוע ואותו אנו רוצים ללמוד.

<u>תזכורת</u>: הוכחנו שתחת הנחות אלה ה-likelihood שווה ל:

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

א. [5 נק'] הוכיחו שתחת הנחות השאלה, בעיית ה-LS ללא רגולריזציה, משמע $\underset{\boldsymbol{w}}{\operatorname{argmin}} \frac{1}{m}\sum_{i=1}^m (\boldsymbol{w}^\top \boldsymbol{x} - y_i)^2$, שקולה לבעיית ה-MLE הבאה: $\underset{\boldsymbol{w}}{\operatorname{argmax}}\, L(\boldsymbol{w}|\{\boldsymbol{x}_i, y_i\}_i)$.

הוכחה ע"י פיתוח תמציתי מנומק:

כעת, נניח בנוסף **שווקטור המשקלים** הלא ידוע $\boldsymbol{w}$ הגיע מהתפלגות לפלאס.

משמע, כל משקל נדגם i.i.d באופן הבא: $\forall k = 1, \dots, d: w_k \sim \text{Laplace}(0, b)$, עבור $b > 0$ נתון (משותף לכל המשקלים).

שימו לב: עדיין מניחים שהרעש ϵ_i מתפלג גאוסיאנית ככתוב בתחילת השאלה.

פונקציית הצפיפות של התפלגות לפלאס הינה: $Z \sim \text{Laplace}(\mu, b) \Longrightarrow f(z) = \frac{1}{2b} \exp\left\{-\frac{1}{b}|z-\mu|\right\}$.

ב. [11 נק'] הוכיחו שתחת כלל ההנחות, בעיית LS עם רגולריזציית ℓ^1, משמע $\underset{\boldsymbol{w}}{\text{argmin}} \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{w}^\top \boldsymbol{x} - y_i)^2 + \lambda\|\boldsymbol{w}\|_1$,

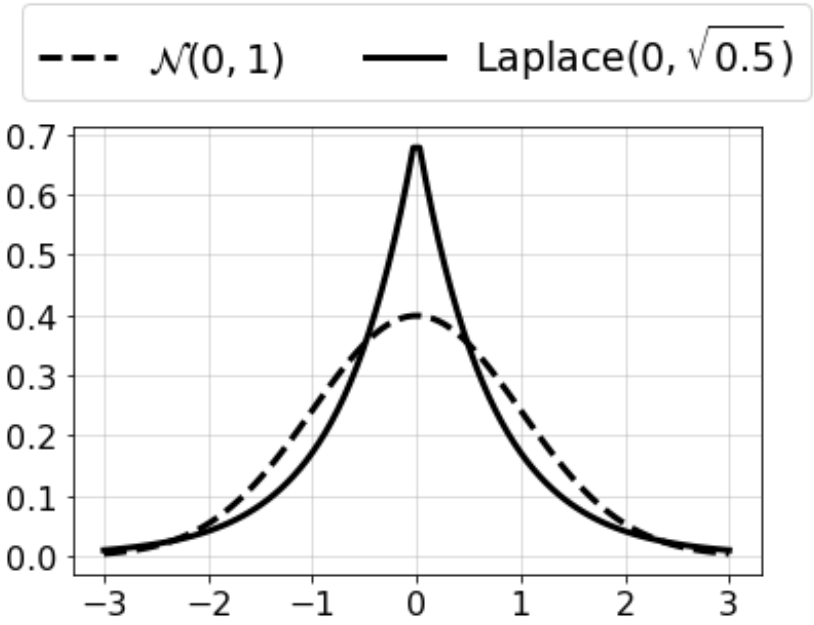
שקולה לבעיית MAP תחת ה-prior על המשקלים, משמע $\underset{\boldsymbol{w}}{\text{argmax}} \Pr\left(\boldsymbol{w} \,\middle|\, \{\boldsymbol{x}_i, y_i\}_i,\ \mu = 0,\ b\right)$.

ניתן להשתמש בחישובים שכבר הוכחו בסעיף הקודם.

הוכחה ע"י פיתוח תמציתי מנומק:

השונות של התפלגות לפלאס נתונה ע"י $\mathrm{Var}[w_k] = 2b^2$.

התרשים משווה בין התפלגות לפלאס להתפלגות נורמלית שהשונות שלהן היא 1.



<u>תזכורת</u>: הוכחנו בתרגול שאם מניחים $w_k \sim \mathcal{N}(0,1)$, בעיית ה-MAP שקולה לבעיית LS עם רגולריזציית ℓ^2.

~~ג.~~ [5 נק'] מתוך הסתכלות בתרשים, מתוך התזכורת ומתוך מה שהוכחתם בסעיף הקודם, הסבירו בקצרה ובאופן אינטואיטיבי (לא פורמלי) הבדל שלמדנו בין אופי הפיתרונות שמתקבלים ע"י רגולריזציית ℓ^2 לאלה המתקבלים ע"י רגולריזציית ℓ^1.

הסבר קצר:

שאלה 4 – Kernel SVM [17 נק']

עבור פרמטר נתון $\gamma > 0$, נגדיר את ה-Gaussian kernel לקלט חד-ממדי באופן הבא:

$$K: \mathbb{R} \times \mathbb{R} \to \mathbb{R}, \qquad K(a,b) = \exp(-\gamma(a-b)^2)$$

א. [12 נק'] הציעו פונקציית מיפוי $\phi: \mathbb{R} \to \mathbb{R}^p$ והוכיחו בעזרתה שהפונקציה K מהווה קרנל חוקי (בחד ממד).

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

תזכורת: פירוק טור טיילור של e^x לכל x נתון ע"י: $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$.

תשובה :

$$e^{-\gamma(a-b)^2} = \underbrace{e^{-\gamma a^2}}_{\alpha} \cdot e^{+2\gamma ab} \cdot \underbrace{e^{-\gamma b^2}}_{\beta} =$$

$$= \alpha\beta \cdot \sum_{n=0}^{\infty} \frac{(2\gamma ab)^n}{n!} = \alpha \cdot \sum \frac{(2\gamma)^{\frac{n}{2}} a^n}{\sqrt{n!}} \cdot \frac{(2\gamma)^{\frac{n}{2}} b^n}{\sqrt{n!}} \cdot \beta$$

$\varphi: \mathbb{R} \to \mathbb{R}^\infty$

$$\varphi(a) = [\varphi_0(a), \varphi_1(a), \dots] \qquad \varphi_i(a) = e^{-\gamma a^2} \cdot \frac{(2\gamma)^{\frac{i}{2}} \cdot a^i}{\sqrt{i!}}$$

$$\varphi(a)^T \varphi(b) = e^{-\gamma(a-b)^2} = K(a,b)$$

המשך לסעיף ב':

ב. [5 נק'] נתון dataset עם $m = 1000$ דוגמאות חד-ממדיות. נרצה לפתור את הבעיה עם ה-Gaussian kernel שהגדרנו. מבחינת יעילות, האם עדיף לפתור את ה-primal problem עם ה-feature mapping שמצאתם, או שעדיף לפתור את ה-dual problem עם פונקציית ה-kernel שהוגדרה? ענו והסבירו בקצרה.

תשובה :

dual מכיוון שפו הדגם הימשלמ לפרמר מכיוון שאמשי הינו
למימד ∞ ורק באמצעות הקרנל זה אמל נשל לבפמ
בר. לוו ורמ צול יבק כתיר מ) בדיתל $\binom{\infty}{2}$

חלק ב' – שאלות אמריקאיות [25 נק']

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

א. [5 נק'] סמנו את כל התשובות הנכונות ביחס לאלגוריתמי One vs. All (1vA) ו-One vs. One (1v1)

(הניחו שיש 10 מחלקות ומעלה).

$\binom{10}{2}$ vs 10

a. ל-1vA סיבוכיות מקום נמוכה יותר מזו של 1v1 בזמן האימון.

b. ל-1vA סיבוכיות מקום נמוכה יותר מזו של 1v1 בזמן המבחן (לאחר שהאימון הושלם).

c. רק אחד משני האלגוריתמים ניתן למיקבול (parallelization).

d. 1v1 נוטה יותר ליצור בעיות לא מאוזנות (imbalanced).

ב. [5 נק'] סמנו את כל הטענות שמשלימות בצורה הגיונית את הטענה הבאה.

באופן כללי, כְּכֹל שה-complexity של מחלקת היפותזות עולה:

a. ה-bias עולה.

b. ה-variance עולה.

c. צריך פחות דאטה על מנת להכליל כראוי.

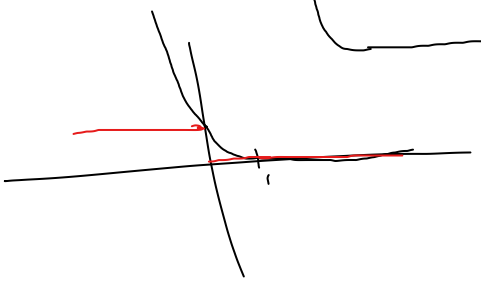
d. יש יותר נטייה ל-overfitting.

e. תהליך האימון של מסווג בודד דורש זמן רב יותר.

$z \le 1: z^2 - 2z + 1$

$z > 1: 0$

ג. [5 נק'] נגדיר את פונקציית ה-squared hinge loss: $\mathcal{L}(z) = (\max\{0,\ 1 - z\})^2$.



$\left(\frac{1}{2}\right)^2 = \frac{1}{4}$

סמנו את כל הטענות הנכונות ביחס לפונקציה זו.

a. הפונקציה קמורה ביחס ל-z.

b. הנגזרת של הפונקציה היא $\frac{\partial}{\partial z}\mathcal{L} = 2 - 2z$.

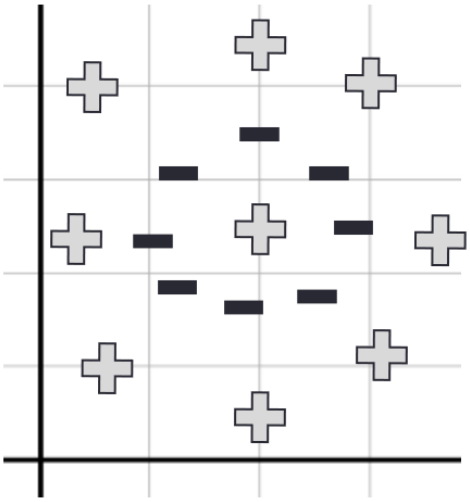
c. הפונקציה חוסמת מלמעלה את ה-0-1 loss בכל מקום.

d. הפונקציה חוסמת מלמעלה את ה-hinge loss בכל מקום.

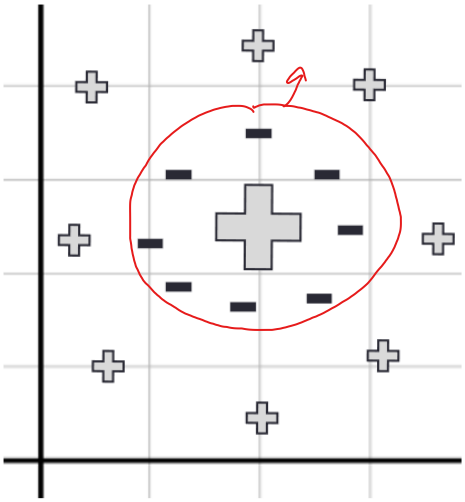
e. עבור בעיות סיווג תחת מודל ליניארי, משמע $z = y_i \boldsymbol{w}^\top \boldsymbol{x}_i$, הפונק' מעודדת margin מהמפריד.

ד. [5 נק'] נתון דאטה עם תיוגים בינאריים ("+" או "-"). מריצים AdaBoost עם מסווג בסיס לא ידוע.
גדלי הצורות בתרשימים מסמלים את ההסתברויות שהאלגוריתם מקצה (הסתברות גבוהה = צורה גדולה).
מריצים את האלגוריתם איטרציה אחת ומקבלים את התרשימים הבאים:

ההתפלגות האחידה ההתחלתית



ההתפלגות לאחר איטרציה אחת



(הנק' האמצעית גדלה והיתר קטנו)

איזה סוג של מסווג בסיס יכול להסביר את התרשים השמאלי שהתקבל? סמנו את התשובה הנכונה.

a. עץ החלטה בעומק 1 (decision stump). משמע, שורש ושני עלים.

b. עץ החלטה בעומק 2. משמע, שורש, רמת ביניים ועד ארבעה עלים.

c. מסווג שאומר על כל המרחב "שקר" או "אמת".

d. SVM עם קרנל פולינומיאלי ממעלה 2.

e. כל התשובות הקודמות לא נכונות.

$\beta \to 0 \to$ uniform

$\beta \to \infty \to$ onehot

ה. [5 נק'] היזכרו בפונקציית ה-Softmax שמשמשת כשכבה האחרונה של רשת נוירונים לסיווג ל-K מחלקות:

$$\text{softmax}(f_1(x), \dots, f_K(x);\ \beta\ ) = \left[\frac{\exp\{\beta f_1(x)\}}{\sum_{i\in[K]}\exp\{\beta f_i(x)\}}\ ,\ \dots\ ,\ \frac{\exp\{\beta f_K(x)\}}{\sum_{i\in[K]}\exp\{\beta f_i(x)\}}\right]^{\mathsf{T}}$$

בסעיף זה אנו לא מתייחסים כלל לאפשרות ש-$\beta = 0$ ומניחים של-β אותו סימן בזמן האימון ובזמן המבחן.
סמנו את כל הטענות הנכונות ביחס לפונקציה זו.

$\beta \to 0 \to e$

$e^0 = 1$

a. בזמן מבחן (לאחר האימון), כאשר $\beta \to \infty$, התפלגות הפלט הולכת להתפלגות אחידה.

b. כאשר משנים את β לאחר האימון בזמן המבחן, כל עוד β שומר על הסימן, אין לו השפעה על הדיוק של הרשת.

c. בזמן אימון, כל עוד הפרמטר β חיובי, אין לו השפעה על מהלך האימון.

d. בזמן אימון, אם הפרמטר β שלילי, לא ניתן ללמוד את הרשת בעזרת שיטות gradient.

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):