מבוא למערכות לומדות (236756)

סמסטר אביב תשע"ד

מבחן מסכם מועד ב' , 2 אוקטובר 2014

מספר סטודנט:

משך המבחן: 3 שעות.

חומר עזר: אין להשתמש בכל חומר עזר.

הנחיות כלליות:

- המבחן כתוב בלשון זכר ומיועד לנשים ולגברים כאחד.
- מלאו את הפרטים בראש דף זה ובדף השער המצורף, בעט בלבד.
- במבחן 15 דפי בחינה וכן 2 דפים לטיוטה. נא ודאו כי כל הדפים נמצאים ברשותכם.
- במבחן 5 שאלות. יש לענות על **כל** השאלות.
- משך המבחן 3 שעות (180 דקות).
- כל התשובות יכתבו על טופס הבחינה, **ויש להחזירו בתום הבחינה**.
- אנא כתבו בכתב יד קריא וברור. **תשובה בכתב יד שאינו קריא לא תיבדק.**
- נא לא לתלוש עמודים ממחברת הבחינה.
- נא לכתוב רק את מה שהתבקשתם ולצרף הסברים קצרים רק כפי שמבוקש בשאלה—אין צורך בהסברים או פרטים נוספים על אלו שהתבקשתם במפורש.

כל המוסיף גורע

1. לשימושכם, דף הגדרות מצורף בסוף המבחן.
2. המילה העברית ל-feature היא תכונה או מאפיין. המילה העברית ל-label היא תיוג.

בהצלחה!

חלק א : אמת או שקר (12 נקודות)

עבור כל אחת מהשאלות הבאות, אנא סימנו **אמת** או **שקר**. במידה שסימנתם **שקר**, הוסיפו הסבר קצר או דוגמא נגדית. אין צורך בהסבר או בנימוק אם סימנתם **אמת**.

1. נניח שעבור משפחת מחלקות $\mathcal{H}_n$ מעל $\{\pm 1\}^n$ **לא** קיים אלגוריתם שרץ בזמן פולינומיאלי, שעבור קלט של דוגמאות מסומנות $(x_1, y_1), \dots, (x_m, y_m)$ $x_i \in \{\pm 1\}^n, y_i \in \pm 1$, מחזיר היפותזה עקבית ($h \in \mathcal{H}_n$ שמקיימת $\forall i\ h(x_i) = y_i$) אם קיימת כזו, או מצהיר שלא קיימת במידה שלא קיימת. אז משפחת מחלקות ההיפותזות $\mathcal{H}_n$ לא ניתנת ללמידת-PAC יעילה.
($\mathcal{H}_n$ is **not** efficiently PAC learnable)

☐ **אמת**

☐ **שקר** הסבר/דוגמא נגדית:

2. נניח שעבור משפחת מחלקות $\mathcal{H}_n$ מעל $\{\pm 1\}^n$ קיים אלגוריתם שרץ בזמן פולינומיאלי A שבהינתן קלט דוגמאות מסווגות $(x_1, y_1) \dots (x_m, y_m)$ מחזיר היפותזה ב $\mathcal{H}$ שמביאה למינימום את מספר השגיאות, כלומר מחזיר את $\arg\min_{h \in \mathcal{H}_n} |\{i = 1..m : h(x_i) \neq y_i\}|$. אז משפחת מחלקות ההיפותזות $\mathcal{H}_n$ ניתנת ללמידת-PAC.
($\mathcal{H}_n$ is efficiently PAC learnable)

☐ **אמת**

☐ **שקר** הסבר/דוגמא נגדית:

3. אם קבוצת אימון $(x_1, y_1), \dots, (x_m, y_m), x_i \in \mathbb{R}^d, y_i \in \{\pm 1\}$ ניתנת להפרדה לינארית, אז שימוש ב SVM עם גרעין (Kernel) $K(x, x') = \langle x, x' \rangle$ ושימוש ב- AdaBoost עם כל אחד מ- d התכונות כמסווג חלש (weak predictor) הם שקולים, כלומר הפלט של שני האלגוריתמים (לאחר מספר מספיק של איטרציות AdaBoost) זהה.

☐ **אמת**

☐ **שקר** הסבר/דוגמא נגדית

4. נניח שקבוצת אימון $(x_1, y_1), \dots, (x_m, y_m), x_i \in \mathbb{R}^d, y_i \in \{\pm 1\}$ היא ניתנת להפרדה לינארית, ואנו מריצים stochastic gradient descent על משתנה המסווג $w \in \mathbb{R}^d$ באופן שבכל איטרציה בוחרים נקודה $i \in \{1..m\}$ באופן אקראי ואוניפורמי ומשתמשים בהערכת הגרדיאנט הסטוכסטי (stochastic gradient estimate) הבא:

$$-[0.5 - y_i \langle w, x_i \rangle]_+ \cdot y_i \cdot x_i$$

אז בסיכוי גבוה, לאחר מספר איטרציות סופי נמצא מסווג שמפריד את קבוצת האימון באופן לינארי.

Commented [NS1]: I made a correction to the subgradient here (my typo in the original draft)

☐ **אמת**

☐ **שקר** הסבר/דוגמא נגדית

חלק ב: זוגות של קטעים (40 נקודות)

בשאלה זו נתייחס למרחב הדוגמאות (instance space) $\mathcal{X} = [0,1]$, ומרחב ההיפותזות של זוגות של קטעים: $\mathcal{H} = \{x \mapsto (2 \cdot [\![a \le x \le b \text{ or } c \le x \le d]\!] - 1) : a,b,c,d \in \mathbb{R}\}$

(כאשר $[\![P]\!]$ מוגדר כ- 1 אם ערך האמת של הפרדיקט P הוא true, אחרת 0.)

1. מהו מימד ה VC של מחלקת ההיפותזות $\mathcal{H}$?

D = 4

2. מצאו דוגמא ל D+1 נקודות שלא ניתן לנתץ, והראו סיווג של הנקודות שמצאתם שלא ניתן לממש באמצעות $\mathcal{H}$.

3. לאבי ולבסאם גישה לתוכנית POLYHINGEFIT אשר מקבלת כקלט דרגה (degree) r, וקבוצת אימון $(x_1, y_1), \dots (x_m, y_m), x_i \in \mathbb{R}, y_i \in \mathbb{R}$ ומחזירה פולינום $p(x)$ מדרגה r שמביא למינימום (מבין כל הפולינומים מדרגה r) את **שגיאת המפרק (hinge loss)**, כלומר את הביטוי $\sum_{i=1}^{m}(1 - y_i p(x_i))_+$

אבי ובסאם מעוניינים להשתמש ב- POLYHINGEFIT כדי ללמוד את $\mathcal{H}$. בכוונתם להשתמש במסווג $x \mapsto sign(p(x))$ בתור מנבא (predictor), כאשר $p()$ הוא הפלט של POLYHINGEFIT. (לצורך השאלה, הניחו ש $sign(0) = 1$). אבי מציע להשתמש בפרמטר $r = 3$ בעוד שבסאם מציע $r = 4$.

a. מהו מימד ה- VC של מחלקת ההיפותזות של מרחב המסווגים של אבי?

תשובה: 4

b. מהו מימד ה- VC של מחלקת ההיפותזות של מרחב המסווגים של בסאם?

תשובה: 5

4. בסעיף זה שני שלבים.

בשלב ראשון, נניח את המקרה נטול הרעש (לא-אגנוסטי, כלומר הסיווגים תמיד עקביים (Consistent) עם היפותזה ב- $\mathcal{H}$) ואת מודל ה- PAC. עבור כל אחת מהשאלות הבאות, סמנו **כן** (ללא הסברים) או **לא** (עם הסבר קצר במקום המיועד).

a. האם כלל הלמידה של אבי מהווה לומד PAC (PAC learner) ביחס ל $\mathcal{H}$?

☐ **כן**

☒ **לא** הסבר:

b. האם כלל הלמידה של אבי מהווה לומד PAC **נאות** (**proper** PAC learner) ביחס ל $\mathcal{H}$?

☐ **כן**

☐ **לא** הסבר:

c. האם כלל הלמידה של בסאם מהווה לומד PAC ביחס ל $\mathcal{H}$?

☑ **כן**

☐ **לא** הסבר:

d. האם כלל הלמידה של בסאם מהווה לומד PAC **נאות** ביחס ל $\mathcal{H}$?

☐ **כן**

☐ **לא** הסבר:

בשלב שני, הניחו למידה אגנוסטית.

e. האם כלל הלמידה של אבי מהווה לומד PAC **אגנוסטי** (**agnostic** PAC learner) ביחס ל $\mathcal{H}$?

☐ **כן**

☐ **לא** הסבר:

f. האם כלל הלמידה של בסאם מהווה לומד PAC אגנוסטי ביחס ל $\mathcal{H}$?

☐ **כן**

☐ **לא** הסבר:

5. אבי ובסאם החליטו שבמקום שגיאת המפרק הם מעוניינים לעבוד עם שגיאה ריבועית. כלומר, הם מעוניינים בתוכנית שמביאה למינימום את השגיאה הריבועית: $\sum_{i=1}^{m}\left(y_i - p(x_i)\right)^2$

הקיפו בעיגול את סעיפי שאלה 4 שהתשובה עליהן תשתנה בעיקבות המעבר מ-POLYHINGEFIT ל POLYSQUAREFIT: a b c d e f

6. השלימו את הקוד (pseudocode) של POLYHINGEFIT **או** של POLYSQUAREFIT.

שימו לב: עליכם לסמן בצורה ברורה איזה קוד בחרתם להשלים, ובמידה והשלמתם את שניהם איזה ברצונכם שיבדק. בכל מקרה ייבדק **רק קוד של אלגוריתם אחד.**

ניתן להניח גישה לפונקצית ספריה מטריציונית pinv(X) המחשבת את הpseudo-inverse של X, המוגדרת כ- $pinv(X) = (X^TX)^{-1}X^T$ כאשר (X^TX) הפיך, ובצורה מתאימה גם כשאיננו הפיך.

```
POLYSQUAREFIT(r, (x1,y1),…,(xm,ym))
Y = transpose([y1,y2,…,ym])
D = [                    ]
Z = zeros(m, D)
for i=1..m
     Z[I,:] = [                                  ]
endfor
return [                                                    ]
```

```
POLYHINGEFIT(r, (x1,y1),…,(xm,ym))
D = [                    ]
Z = zeros(m, D)
for i=1..m
     Z[I,:] = [                                  ]
endfor
use an LP solver to solve the following LP:
     variables: w∈ ℝ^D, e∈ [        ]
     objective: minimize [                                  ]
     constraints:
       forall i=[     ]to[     ]: [                                  ]
                                        ei ≥ 0
output w
```

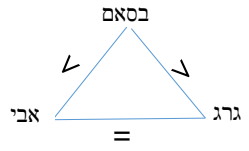
7. במהלך החיפושים באינטרנט, נתקלו אבי ובסאם ב POLY01FIT שעובד באותה צורה כמו POLYHINGEFIT, רק מביא למינימום את שגיאת ה- 0/1, כלומר את הביטוי $\sum_{i=1}^{m}[\![y_i p_i(x) \leq 0]\!]$

הקיפו בעיגול את סעיפי שאלה 4 שהתשובה עליהן תשתנה בעיקבות המעבר מ- POLYHINGEFIT ל- POLY01FIT: a b c d e f

8. בנוסף לאבי ולבסאם, גרג לומד ע"י מציאת היפותזה ב- $\mathcal{H}$ ישירות (כלומר ללא שימוש בפולינומים) שמביאה למינימום את שגיאת ה- 0/1. האם קיים אלגוריתם יעיל (הרץ בזמן פולינומיאלי) לביצוע המשימה של גרג? אם לא, הסבירו בקצרה מדוע. אם כן, הסבירו בשורה או שתיים את הרעיון או האסטרטגיה העיקריים שאלגוריתם כנ"ל ינצל. רישמו את זמן הריצה במקום המיועד לכך.

זמן הריצה: O(_____)

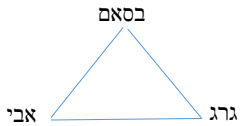
9. לצורך שאלה זו, הניחו שאבי ובסאם משתמשים במסווג $x \mapsto sign(p(x))$ כאשר $p()$ הוא הפלט של POLY01FIT עם הפרמטרים לעיל, כלומר r=3 עבור בסאם ו- r=4 עבור אבי. גרג ממשיך לפי התנאים של שאלה 8. כמו כן הניחו שהסיווגים אכן מתאימים להיפותזה ב- $\mathcal{H}$ (כלומר המקרה הלא-אגנוסטי). השתמשו בסימני השוואה $=, <, >, \leq, \geq$ כדי לתאר את הקשר בין שלושת הלומדים ביחס למדדים הבאים:



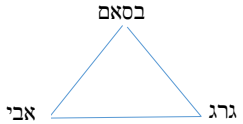
דוגמא: מספר האותיות בשם המשתמש בעיברית:

(המשך הסעיף בעמוד הבא)

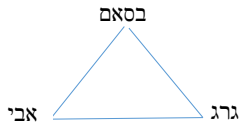
שגיאת סיווג (training error):



שגיאת קירוב (approximation error):



שגיאת שיערוך (estimation error):



חלק ג: חיזוי של סוכרת (30 נקודות)

ד"ר מרגלית מעוניין לנבא, בהינתן אדם אקראי בן 30, האם הוא צפוי לסבול מסוכרת בגיל 60, בהינתן 20 מאפיינים בינאריים שנראים לו רלוונטיים.

לצורך משימה זו, ד"ר מרגלית אסף נתונים על 10,000 חולי סוכרת אקראיים בני 60 ו- 10,000 אנשים אקראיים שאינם סובלים מסוכרת (להלן: "בריאים") בני 60. עבור כל אחת מ- 20,000 הדוגמאות אסף ד"ר מרגלית מארכיון קופת חולים את 20 המאפיינים מהתיק הרפואי שנשמר כשהיו בני 30. ד"ר מרגלית בדק ומצא (ע"י חיפוש באינטרנט) ש- 9% מכלל האוכלוסיה של בני ה- 60 סובלים מסוכרת.

לצורך השאלה, ניתן להניח שכל אדם בן 30 יזכה לחיות עד גיל 60 לפחות.

נסמן את וקטורי המאפיינים של 10,000 החולים ב- $x_1, \dots, x_{10000} \in \{0,1\}^{20}$ ואת וקטורי המאפיינים של הבריאים ב- $x_{10001}, x_{10002}, \dots, x_{20000} \in \{0,1\}^{20}$.

הסיווג $+1$ יסמן חולה סוכרת, בעוד ש- -1 יסמן בריא.

1. עבור מסווג $h: \{0,1\}^{20} \mapsto \{\pm 1\}$ כלשהו, הציעו משערך בלתי-מוטה (unbiased estimator) $\tilde{L}(h)$, לתוחלת השגיאה האמיתית של h כמנבא סוכרת עתידית בגיל 60 לאדם בן 30.

$$\tilde{L}(h) = \frac{1}{10000}\sum_{i=1}^{10000}\left[h(x_i) \neq -1\right] \cdot \frac{99}{100} + \frac{1}{10000}\sum_{10001}^{20000}\left[h(x_i) \neq 1\right] \cdot \frac{9}{100}$$

False Positive False Negative

2. בניסיון הראשון של ד"ר מרגלית, הוא החליט לחשב, מבין כל הפונקציות הבינאריות האפשריות, פונקציה $h \in \{0,1\}^{20} \mapsto \{\pm 1\}$ המביאה למינימום את $\tilde{L}(h)$.

~~a~~. האם מדובר בגישה דיסקרימינטיבית (discriminative) או גנרטיבית (generative)?

תשובה:

~~b~~. האם פיתרון זה מומלץ? הסבירו את תשובתכם בקצרה:

3. בניסיון השני, ד"ר מרגלית מחליט להשתמש בשיטת ה- logistic regression כדי לחשב מפריד לינארי לניבוי סוכרת על סמך 20 המאפיינים.

a. האם מדובר בגישה דיסקרימינטיבית או גנרטיבית?

תשובה: _____

b. מה היתרון **העיקרי** של גישה זו על פני הגישה של שאלה 2 לעיל?

תשובה:

4. ד"ר מרגלית מחליט לנקוט בגישה גנרטיבית ללא שום הנחות על המודל המייצר את הנתונים.

כמה פרמטרים על ד"ר מרגלית לשערך? תשובה: _____

(הערה: עליכם לכתוב מספר מדויק. תשובה שרחוקה בלכל היותר ± 1 מהתשובה הנכונה תזכה בכל הנקודות.)

5. בניסיון האחרון, מחליט ד"ר מרגלית לנסות את גישת Binary Naïve Bayes

a. אנא סמנו ב- X את **כל** הטענות **הנכונות** מבין הבאות:

- ☐ ד"ר מרגלית נוקט בגישה דיסקרימינטיבית, והפלט זהה לפלט של logistic regression
- ☐ ד"ר מרגלית נוקט בגישה גנרטיבית
- ☐ הפלט ניתן לביטוי כמסווג לינארי
- ☐ הגישה מניחה שבהינתן הארוע "האדם יסבול מסוכרת בגיל 60", שני הארועים "$x[1] = 1$" ו- "$x[2] = 0$" הינם בלתי-תלויים.
- ☐ הגישה מניחה:

$$\Pr[x[1] = 1 \mid healthy\ at\ 60] = \Pr[x[2] = 1 \mid healthy\ at\ 60]$$

b. כמה פרמטרים על ד"ר מרגלית לשערך עתה?

תשובה: _____

(הערה: עליכם לכתוב מספר מדויק)

ה. אחד הפרמטרים המשוערכים לצורך מודל זה הוא

$$\Theta_{1,+} = \Pr[x[1] = 1 \mid diabetes\ at\ 60]$$

רשמו את משערך הנראות המקסימלית (Maximum Likelihood Estimator)

עבור פרמטר זה:

$\widehat{\Theta_{1,+}} =$

חלק ד: גרעינים (Kernels) (6 נקודות)

יהיו K_1, K_2 ו- K_3 גרעינים תקינים. הגרעין K_2 מתאים למיפוי מאפיינים (feature map) $\phi(x) \in \mathbb{R}^{100}$. הגרעין K_3 מתאים למיפוי מאפיינים 20 ממדי $\psi(x) \in \mathbb{R}^{20}$ המקיים:

$$\forall i = 1..20, \forall x:\ \psi_i(x) = \phi_i(x)$$

כלומר, ψ שווה ל- 20 הקואורדינטות הראשונות של ϕ.

סמנו את התשובה הנכונה בעיגול:

האם $K_1(x,x') + K_2(x,x')$ גרעין תקין באופן כללי?	כן	לא
האם $K_1(x,x') + 10 \cdot K_2(x,x')$ גרעין תקין באופן כללי?	כן	לא
האם $K_1(x,x') - K_2(x,x')$ גרעין תקין באופן כללי?	כן	לא
האם $K_2(x,x') - K_3(x,x')$ גרעין תקין באופן כללי?	כן	לא
האם $K_3(x,x') - K_2(x,x')$ גרעין תקין באופן כללי?	כן	לא

חלק ה: שכנים קרובים (Nearest Neighbors) (12 נקודות)

אנסטסיה כתבה תוכנית שמציירת את גבול ההחלטה (decision boundary) של מסווג k-nearest neighbor עבור חמישה תסריטים שונים, המתאימים למספר שונה של נקודות אימון ובחירות שונות של הפרמטר k (מספר השכנים הקרובים שקובעים את הסיווג). בכל המקרים, נקודות האימון הוגרלו מאותה ההתפלגות.

חמשת התסריטים הם כדלקמן:

A. 20 נקודות עם $k = 1$.

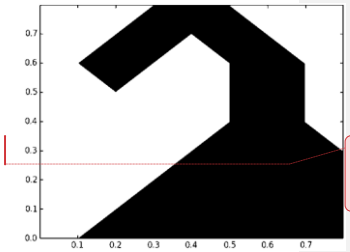
B. 20 נקודות עם $k = 20$.

C. 1000 נקודות עם $k = 1$.

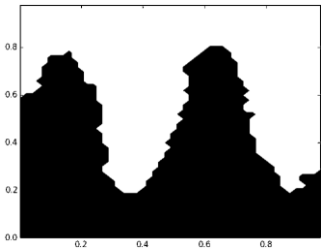
D. 1000 נקודות עם $k = 20$.

E. 1000 נקודות עם $k = 200$.

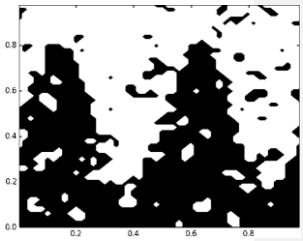
לרוע המזל, אנסטסיה שכחה איזה ציור מתאים לכל אחד מהתסריטים. עליכם לעזור לה לשחזר את ההתאמה ע"י ציון האותיות A,B,C,D,E (כל אחת בדיוק פעם אחת) מתחת לכל ציור.



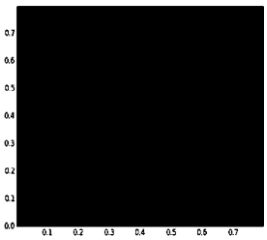
A



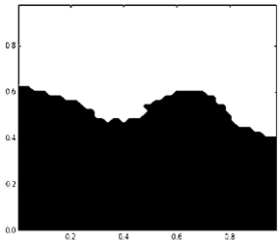
D



C



B



E

Commented [NS2]: I'd make the m=20,k=20 all black rather than all white, just so its clearer that this isn't a printing error or something very faint they can't see

תזכורת הגדרות:

For an instance space $\mathcal{X}$ and label space $\mathcal{Y} = \{-1, +1\}$, a learning algorithm takes as input a sequence of labeled examples $A = (x_1, y_1), \dots, (x_m, y_m)$ and outputs a predictor $A(S): \mathcal{X} \to \mathcal{Y}$.

We say that a learning algorithm A **PAC learns** a hypothesis class $\mathcal{H} \subseteq \mathcal{X}^{\mathcal{Y}}$, if for any $\epsilon, \delta > 0$, there exists m, such that for any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ where $\exists_{h \in \mathcal{H}} L_{\mathcal{D}}(h) = 0$, with probability at least $1 - \delta$ over the draw of m iid samples $S \sim \mathcal{D}^m$: $L_{\mathcal{D}}\big(A(S)\big) \leq \epsilon$.

We say that a learning algorithm A **agnostically PAC learns** $\mathcal{H}$ if for any $\epsilon, \delta > 0$, there exists m, such that for any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, with probability at least $1 - \delta$ over the draw of m iid samples $S \sim \mathcal{D}^m$: $L_{\mathcal{D}}\big(A(S)\big) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$.

We say that learning is **proper** if $A(S) \in \mathcal{H}$ for all training sets S.

For a family of hypothesis classes $\mathcal{H}_n$ over $\mathcal{X}_n = \{\pm 1\}^n$, we say that learning is **efficient** if the learning algorithm $A(\cdot)$ learns all $\mathcal{H}_n$ in time polynomial in $\frac{1}{\epsilon}, \frac{1}{\delta}$ and n and outputs a hypothesis that can be evaluated in time polynomial in these quantities.

We say a hypothesis class (or family of classes) is (properly/agnostically/efficiently) **learnable** if there exists a (perper/agnostic/efficient) learning algorithm for it.

עמוד לטיוטה

עמוד לטיוטה