



**מבוא למערכות לומדות (236756)**

**סמסטר חורף תשע"ט**

**מבחן מסכם מועד א', 11 בפברואר 2019**

--	--	--	--	--	--	--	--	--

מספר סטודנט:

**משך המבחן: 2.5 שעות. (150 דקות)**

**חומר עזר:** אין להשתמש בכל חומר עזר. בעמוד הבא לרשותכם דף נוסחאות והגדרות.

**הנחיות כלליות:**

- המבחן כתוב בלשון זכר ומיועד לנשים ולגברים כאחד.
- מלאו את הפרטים בראש דף זה ובדף השער המצורף, בעט בלבד.
- במבחן 18 דפים ממוספרים סהכ, כולל עמוד זה שמספרו 1. ודאו שיש לכם כל הדפים.
- במבחן 4 חלקים. יש לענות על כל השאלות.
- 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.

**Less is More**

**בהצלחה!**



דף נוסחאות

$$\binom{n}{k} \leq n^k \quad 1.$$

$$L_D^{01} = \text{true error} = \text{שגיאת הכללה} \quad 2.$$

$$L_S^{01} = \text{training error} = \text{empirical error} = \text{שגיאת אימון} = \text{שגיאה אמפירית} \quad 3.$$

(השגיאות על מדגם)

$$L_D^{01} - L_S^{01} = \text{estimation error} = \text{שגיאת הערכה} \quad 4.$$

$$e \approx 2.72 \quad 5.$$

$$\text{התפלגות גאומטרית עם פרמטר } p \quad 6.$$

$$P(x = n) = p (1 - p)^{n-1}$$

$$7. \text{ תהא } H \text{ מחלקת היפוטזות של בעיית למידה כלשהי, ו- } S \text{ קבוצת אימון שנבחרת באקראי. נסמן}$$

$$\hat{h} = \operatorname{argmin}_{h \in H} L_D^{01}(h)$$

$$h = \operatorname{argmin}_{h \in H} L_S^{01}(h) \text{ אזי, לכל } \delta > 0: \text{ בהסתברות של לפחות } 1 - \delta \text{ מתקיים:}$$

$$L_D^{01}(\hat{h}) \leq L_D^{01}(h) + O\left(\sqrt{\frac{\operatorname{VCDIM}(H) + \frac{1}{\log(\delta)}}{|S|}}\right)$$

$$8. \text{ מאפיין } = \text{feature}$$

$$9. \text{ סיווג } = \text{label}$$

$$10. [x]_+ = \max(x, 0)$$

$$11. \text{ מונום (monomial) במשתנים } x[1] \dots x[k] \text{ מדרגה } d \text{ הוא מכפלה מהצורה}$$

$$x[i_1] \cdot x[i_2] \cdots x[i_d] \text{ כאשר } i_1 \dots i_d \text{ הם בתחום } \{1..k\} \text{ (עם חזרות).}$$

$$12. \llbracket \text{תנאי} \rrbracket \text{ מוגדר כ-1 אם התנאי מתקיים, אחרת 0.}$$

$$13. \text{ Decision stumps על ייצוג מאפיינים כלשהו } x[1]..x[k] \text{ הוא מחלקת היפותזות הכוללת}$$

$$\text{פונקציות מהצורה } h_{i,a} = \llbracket x[i] \geq a \rrbracket \text{ עבור } i=1..k \text{ ו- } a \text{ מספר ממשי כלשהו.}$$



**חלק א : שאלות קצרות (30 נק')**

1. נתונה בעיית סיווג עם 3 קלאסים כאשר לכל דוגמא יש 5 מאפיינים בינאריים. כמה פרמטרים חופשיים יש לשערך אם ברצוננו להשתמש באלגוריתם Naive Bayes?  
\* הסבר: פרמטר הוא חופשי אם הוא לא תלוי בפרמטר אחר. לדוגמא, שני פרמטרים  $a, b$  עם האילוץ  $a+b=1$  ניספרים כפרמטר יחיד.

2. ענו על סעיף ב' כאשר מסירים את ההנחה של Naive Bayes

3. סכום הערכים העצמיים המתקבלים בביצוע PCA שווה לשונות של הדוגמאות המקוריות

☐ אמת

☐ שקר



חובה לספק הסבר עבור התשובה שנבחרה:

4. הרכבה של שתי פונקציות קמורות (ב  $R^d$ ) היא פונקציה קמורה

אמת ☐

שקר ☐ דוגמא נגדית (חובה לספק במקרה שסימנתם "שקר"):

$$\left. \begin{array}{l} f(x) = x^2 \\ g(x) = -x \end{array} \right\} \text{קמורות}$$
$$g(f(x)) = g(x^2) = -x^2 \leftarrow \text{שקמורה}$$

5. פונקציית המטרה של Soft-SVM היא  $L_S^{hinge}(w) + \lambda \|w\|_2^2$

ככל שמגדילים את  $\lambda$ , כך צפוי ששגיאת האימון (יש לסמן אפשרות אחת):

תרד ☐

תעלה ☒

תישאר ללא שינוי ☐

6. ככל שנגדיל את  $k$  באלגוריתם KNN ככה האלגוריתם ייטה לכיוון over-fitting

אמת ☐

שקר ☒

7. הסבירו מהו Feature Selection מסוג Filter



8. תיזכורת: אלגוריתם ה-EM מקבל כקלט נתונים ממודל הסתברותי ידוע, ונותן כפלט פרמטרים של המודל. משתמשים בו בעיקר כאשר המודל ההסתברותי מכיל הן משתנים גלויים (observables) ומשתנים נסתרים (latent variables). האלגוריתם עובד בסבבים (איטרציות).

סמנו את כל האפשרויות הנכונות (ניתן לסמן יותר מאפשרות אחת):

- ☐ האלגוריתם EM מוגדר רק עבור מודלים הסתברותיים שבהם המשתנים הנסתרים מקבלים ערכים מתוך קבוצה סופית. (כלומר, לדוגמא, האלגוריתם לא מוגדר אם משתנה נסתר כלשהו יכול לקבל ערך ממשי בתחום  $[0,1]$ ).
- ☐ נסמן את ערך פונקציית הנראות המתאימה לסבב  $t$  במשתנה  $L(t)$ . אז הסידרה  $L(1), L(2), L(3), \dots$  לא בהכרח מתכנסת.
- ☐ הסידרה  $L(1), L(2), L(3), \dots$  שהוגדרה לעיל הינה מונוטונית עולה.
- ☐ האלגוריתם עושה אופטימיזציה באופן ישיר לפונקציית הנראות על-ידי Stochastic Gradient Descent על פונקציית הלוגריתם של הנראות (Log Likelihood).

9. סמנו את האפשרויות הנכונה (ישנה אחת ויחידה), ופרטו בהתאם

- ☐ אלגוריתם החצייה (Halving) הינו Improper Learning Algorithm

נימוק (דוגמא למרחב דוגמאות, מחלקת היפותזות ואוסף דוגמאות שמדגימים זאת)



אלגוריתם החצייה הינו Proper Learning Algorithm ☐

10. סמנו את האפשרות הנכונה (ישנה אחת ויחידה):

אלגוריתם ERM (Empirical Risk Minimization) הינו Improper Learning Algorithm. ☐

נימוק (דוגמא למרחב דוגמאות, מחלקת היפותזות ואוסף דוגמאות שמדגימים זאת)

אלגוריתם ERM הינו Proper Learning Algorithm ☐

11. אם נריץ אלגוריתם AdaBoost עם מסווג חלש (weak classifier) שהוא לינארי במאפייני הנתונים שלנו, אז הפלט של AdaBoost אחרי כל מספר של איטרציות יהיה מסווג לינארי גם כן. טענה זו הינה (סמנו אפשרות אחת):

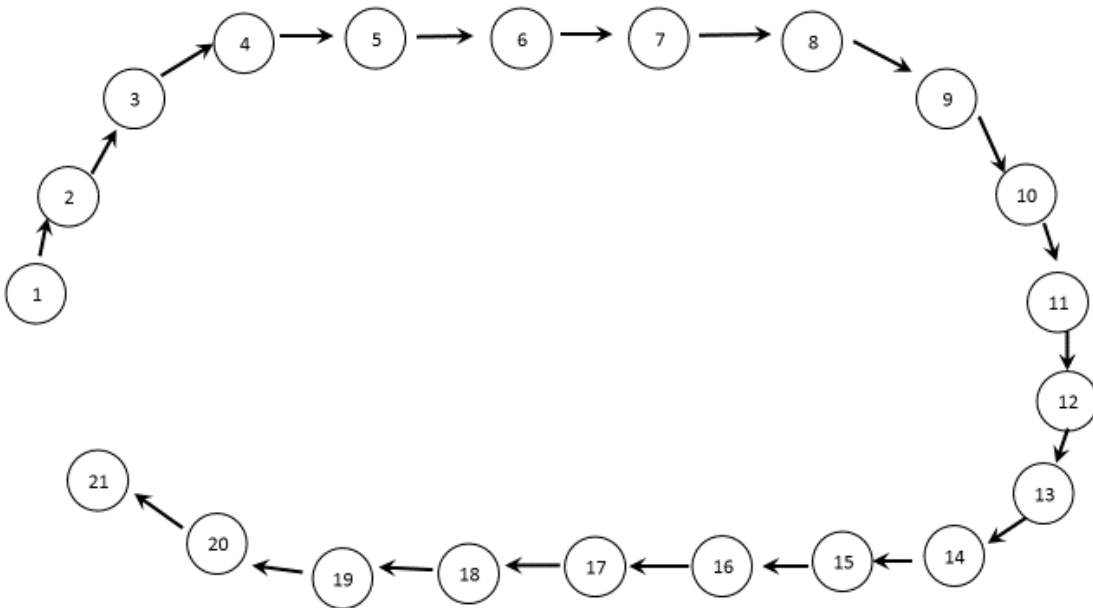
נכונה ☐

לא נכונה ☒



**חלק ב: מימד VC של פונקציות מונוטוניות (28 נקודות)**

א. לפניכם ציור של 21 עיגולים ממוספרים, כאשר חלק מזוגות העיגולים מחוברים בחץ (למשל, עיגול 16 מחובר בחץ שהולך לעיגול 17).



בהינתן פונקציה בינארית על העיגולים  $h: \{1, 2, 3, \dots, 21\} \mapsto \{0, 1\}$ , נאמר שהיא מונוטונית עולה ביחס לציור אם לכל החיצים מתקיים שערך הפונקציה בעיגול שממנו החץ יוצא אינו גבוה מערך הפונקציה בעיגול שאליו החץ מגיע. לדוגמא, אם  $h(4) = 1, h(5) = 0$  אז הפונקציה  $h$  אינה מונוטונית עולה ביחס לציור, מכיוון שיש חץ שמחבר בין עיגול 4 לעיגול 5, ואסור שערך הפונקציה ירד מ-1 ל-0 לאורך החץ. לעומת זאת, הפונקציה  $h$  לפי הפירוט הבא דווקא כן מונוטונית ביחס לציור:

$$h(1) = h(2) = \dots h(15) = 0 \quad h(16) = \dots = h(21) = 1$$

תהי  $H$  מחלקת היפותזות של כלל הפונקציות הבינאריות המונוטוניות ביחס לציור.

$$h(1)=1 \quad h(2)=0$$

a. מהו מימד ה VC של  $H$ ?

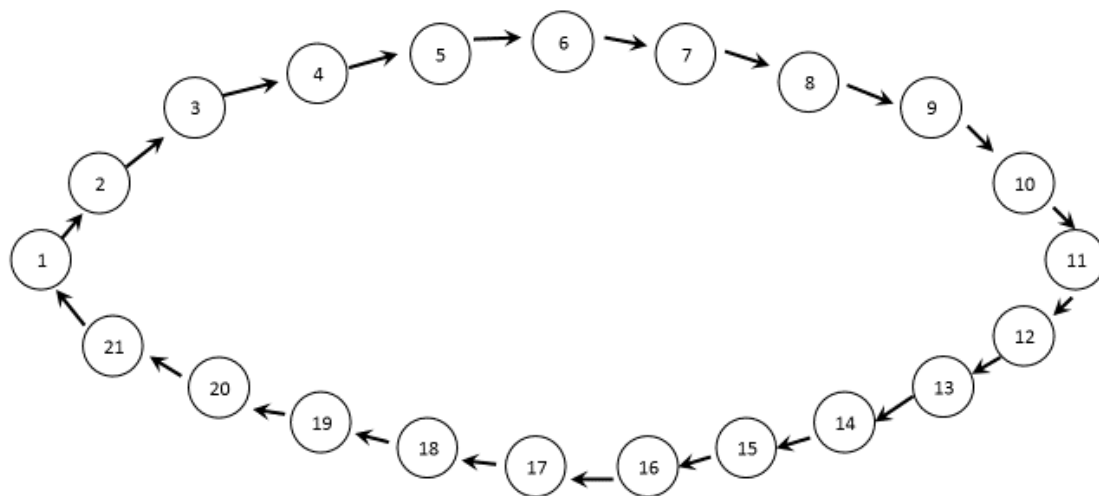
1

b. תארו קבוצה מקסימלית של עיגולים שאפשר לנתץ (to shatter) באמצעות  $H$ . (ניתן

פשוט לרשום רשימה של מספרים בין 1 ל 21)



ב. ענו על סעיפים a,b לעיל ביחס לציור הבא:



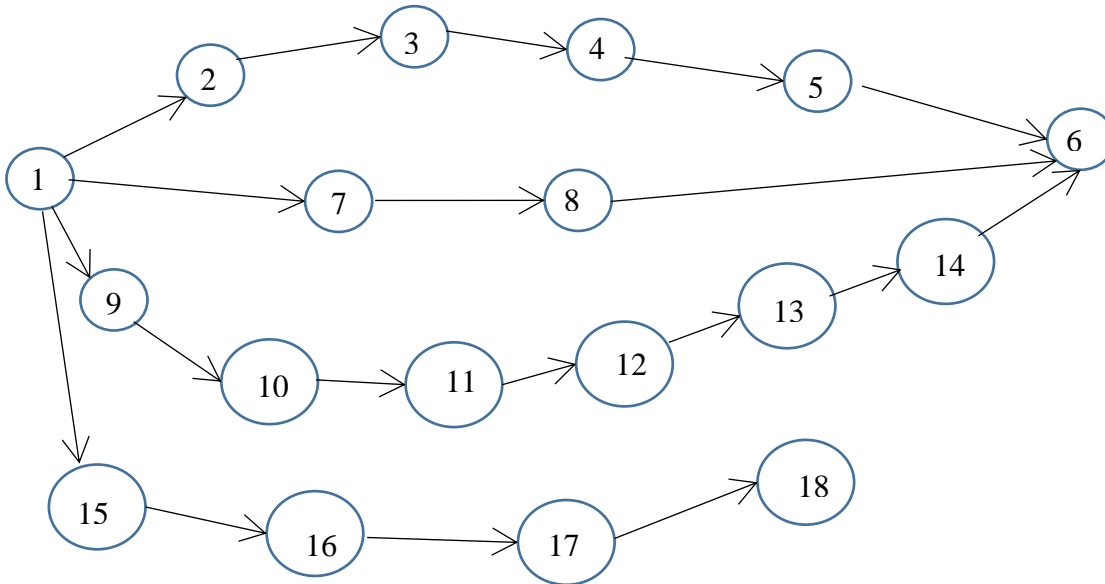
a. מימד ה VC:

b. ניתוח קבוצה מקסימלית:





ג. ענו על סעיפים a,b לעיל ביחס לציור הבא, המכיל 18 עיגולים:

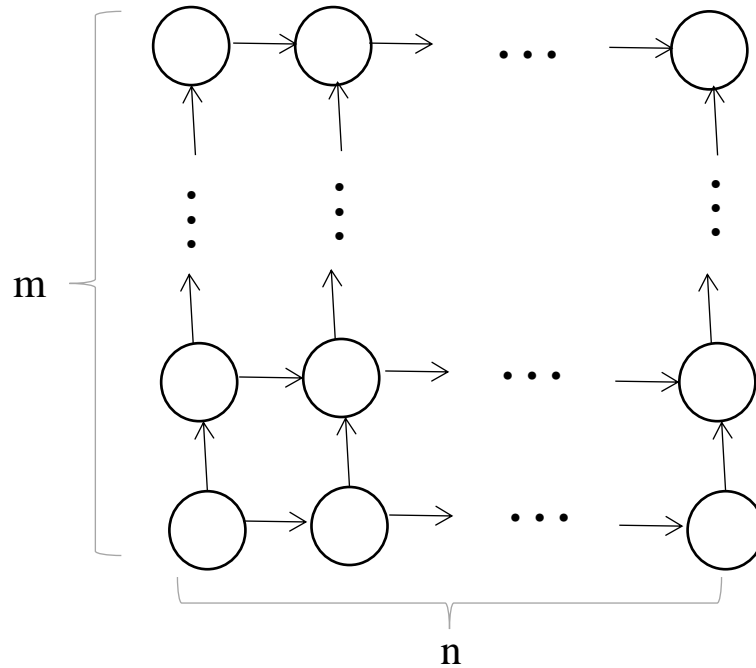


a. מימד ה VC:

b. ניתוח קבוצה מקסימלית:



ד. ועתה, ביחס לציור הבא, המכיל  $m \cdot n$  עיגולים, כאשר  $m, n$  הינם פרמטרים  $1 \leq$



a. מימד ה VC:

b. ניתן קבוצה מקסימלית (יש לתאר את הקבוצה במילים או בציור):



**תל"ג: מודלים הסתברותיים (27 נקודות)**

1. חשבו את אומד הנראות המקסימלי (MLE) עבור התפלגות גיאומטרית. להזכירכם התפלגות

גיאומטרית הינה התפלגות בדידה שפונקציית ההתפלגות שלה נתונה ע"י הנוסחא הבאה:

$$P(X = k) = (1 - p)^k p$$

הביטוי שקיבלתם צריך להיות פונקצייה של  $m$  הגרלות בת"ל  $x_1 \dots x_m$

2. נתון prior על הפרמטר  $p$ , שלפיו  $\Pr[p = 0.25] = \Pr[p = 0.75] = \frac{1}{2}$ . רישמו כיצד

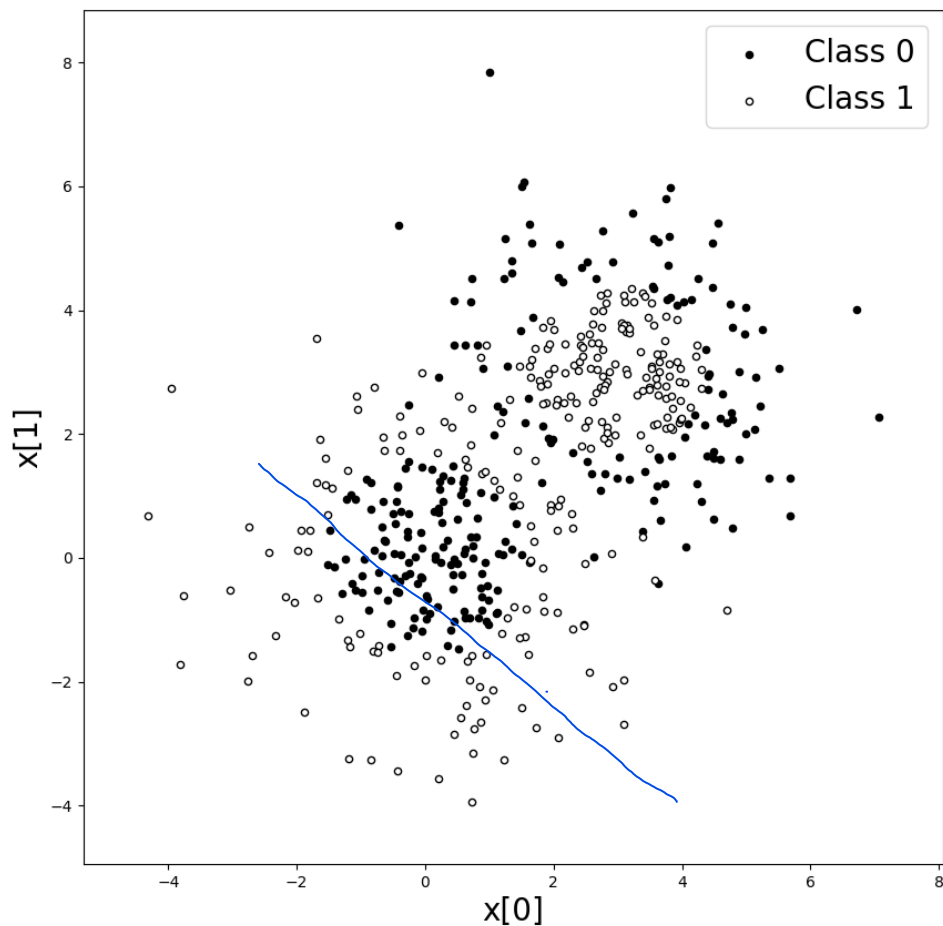
לבצע שיערוך MAP של  $p$ .



### חלק ד: הבנת תוצאות ריצה (15 נקודות)

הערה: השאלה הבאה בנויה על פלטי קוד שמבוסס במקורו על הקישור הבא:  
[https://scikit-learn.org/stable/auto\\_examples/ensemble/plot\\_adaboost\\_twoclass.html](https://scikit-learn.org/stable/auto_examples/ensemble/plot_adaboost_twoclass.html)

שרלוק רצה להתנסות ב adaboost וב kernel svm, ולצורך זה התקין סביבת פיתוח python ובה התקין סיפריות ML סטנדרטיות. בשלב הראשון הוא הגריל מערך נתונים הכולל 500 דוגמאות  $x_1, \dots, x_{500}$  ו-500 סיווגים בינאריים  $y_1, \dots, y_{500}$ , כאשר כל דוגמא  $x_i$  היא נקודה בעלת 2 קואורדינטות  $x_i[1], x_i[2]$ . את 500 הנקודות והסיווגים הוא צייר בדיאגרמה הבאה:



את הנתונים הוא הכניס למסווג מסוג AdaBoost, שעובד בדיוק כפי שנלמד בכיתה. בתור מסווג חלש (weak classifier) הוא השתמש ב decision stump.

הוא קרא פעם בפוסט בפייסבוק, שמספר האיטרציות של AdaBoost הוא פרמטר שחשוב לקבוע בזהירות. לכן הוא עצר את האלגוריתם אחרי 7 איטרציות, אחרי 10 איטרציות ואחרי 100 איטרציות.



לאחר התבוננות בפלטים הוא החליט להשתמש גם עם feature generation, וייצר שני ייצוגים חדשים לנתונים:

$$\begin{aligned}\phi(x) &= (x[0], x[1], x[0] - x[1]) & \chi(0) &= \chi(1) \\ \psi(x) &= (x[0], x[1], x[0] + x[1]) & \chi(0) &= -\chi(1)\end{aligned}$$

במילים אחרות הייצוג  $\phi(x)$  מעשיר את הייצוג של  $x$  על-ידי הוספת מאפיין ששווה להפרש המאפיינים הקיימים, והייצוג  $\psi(x)$  מעשיר את הייצוג על-ידי הוספת מאפיין הסכום. גם עבור כל אחד משני הייצוגים  $\phi, \psi$  הוא הריץ AdaBoost/Decision-Stump עם מספר איטרציות 7, 10, 100.

לאחר כל הרצה, הוא ייצר תרשים של פונקציית ה prediction (הפלט של AdaBoost) על כל הנקודות שבריבוע  $(x[0], x[1]) \in [-4, 8] \times [-4, 8]$ .

את שלוש התמונות המתאימות ל 100 איטרציות הוא שיתף באינסטגרם\*, אבל לא ציין איזו תמונה מתאימה לאיזה feature vector.

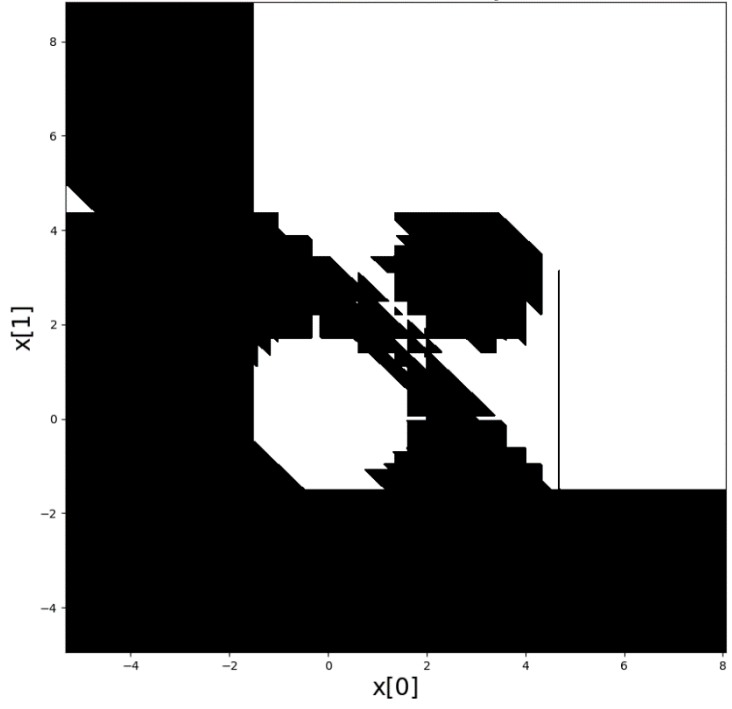
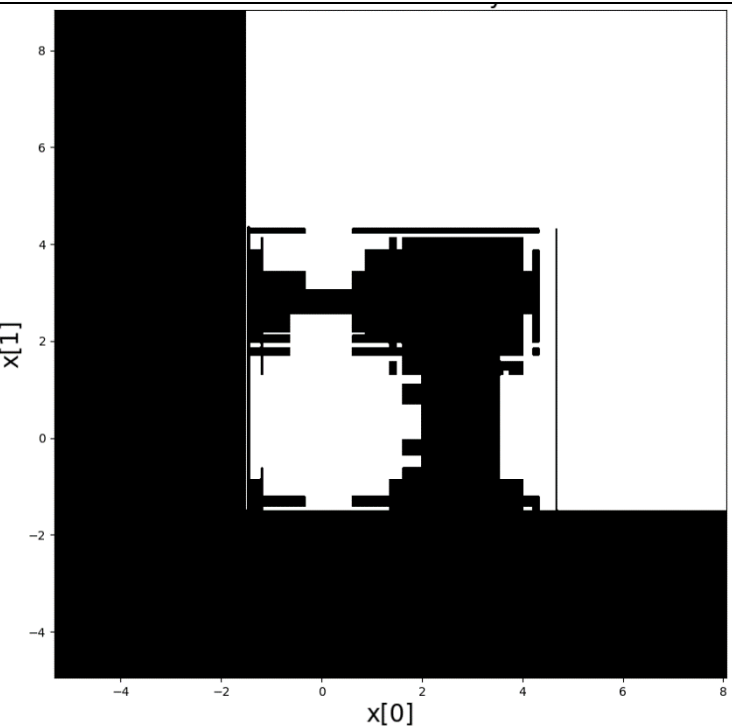
עליכם לעזור לאלפי העוקבים של שרלוק באינסטוש\*\* להתאים את התמונות ל feature vector המתאים בכל מקרה.

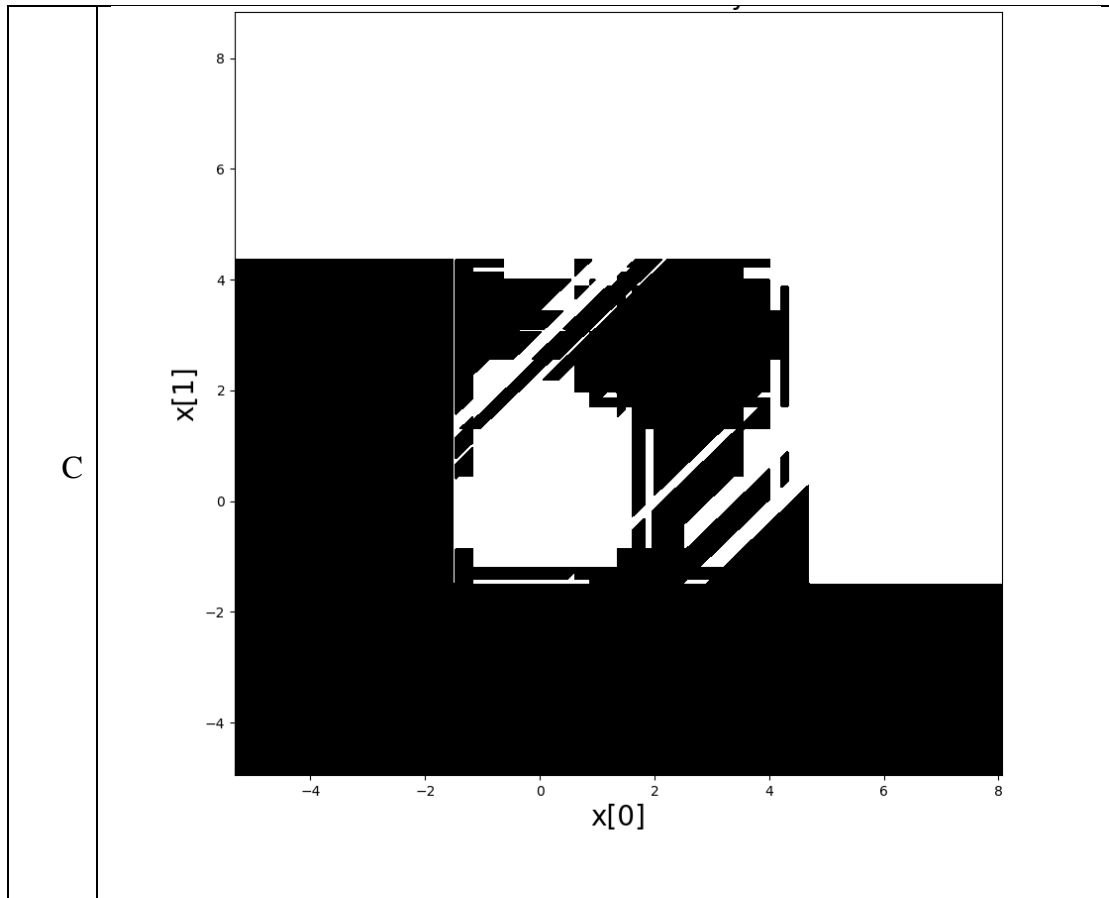
1. ליד כל שורה, כיתבו  $x$  או  $\phi(X)$  או  $\psi(x)$

- תמונה A מתאימה ל  $\psi(x)$
- תמונה B מתאימה ל  $x$
- תמונה C מתאימה ל  $\phi(x)$

\*, \*\* אינסטגרם הינה פלטפורמה מקוונת לשיתוף תמונות, ואינסטוש הוא שם כינוי לאינסטגרם שבשימוש בעיקר בקרב צעירים.



A	
B	



לאחר ששיתף את התמונות המתאימות ל 100 איטרציות, הוא התבונן בציור המתאים ל 7 איטרציות  
ביחס ל feature vector  $\phi(x)$ . הציור, שאותו נסמן באות D, נראה כך:







לאחר שסיים להשתמש עם AdaBoost, עבר שרלוק ל Kernel soft-SVM של סיפריית התכנה sklearn ובחר לעבוד עם kernel פולינומיאלי ממעלה 3. כידוע גרעין (kernel) פולינומיאלי מוגדר ביחס לווקטור מאפיינים בסיסי ("vanilla features"). שרלוק השתמש בכל שלושת ווקטורי המאפיינים שהוגדרו לעיל  $\phi(x)$ ,  $\psi(x)$ , בתור מאפייני בסיס, ועבור כל אפשרות הריץ את סיפריית ה SVM, מבלי לשנות את הפרמטרים האחרים של SVM. שרלוק שמ לב לשלוש העובדות הבאות:

- המאפיין השלישי (הנוסף) של כל אחד הייצוגים  $\phi(x)$ ,  $\psi(x)$  הוא צרוף לינארי של שני המאפיינים הראשונים (להזכירכם: האחד הוא סכום, השני הוא הפרש)
  - אלגוריתם ה Kernel-soft-SVM בסיפריית sklearn הוא דטרמיניסטי (לא אקראי)
  - כל מונם מדרגה 3 במשתנים שהם צרופים לינארים של  $x[0]$ ,  $x[1]$  הוא צרוף לינארי של מונמים מדרגה 3 במשתנים  $x[0]$ ,  $x[1]$ .
- לאור עובדות אלה, הסיק שרלוק שהמסוג האופטימלי שפולט אלגוריתם ה soft-SVM חייב להיות זהה עבור כל אחד משלושת הייצוגים של הנתונים. להפתעתו, הוא גילה שלא כך המצב.
4. הסבירו בקצרה מדוע קיימים הבדלים בין פלטי שלוש ההרצות של ה kernel soft-SVM

הוספנו עוד מימד שאלו אפוא אפוא כיזון  
אך הברור יוכל אפוא אפוא אפוא  
כאלה margins ה



לאחר  $k$  ימים סיים שרלוק להריץ  $5k$  אלגוריתמי למידה שונים ומשונים על אותם הנתונים  $(x_1, y_1), (x_2, y_2) \dots (x_{500}, y_{500})$ . מבין  $5k$  המסווגים שצבר הוא רצה לבחור את הטוב ביותר, מבחינת שגיאת הכללה. לשם כך הוא החליט לעשות ולידציה (Validation) על נתונים חדשים.

5. כמה דוגמאות חדשות עליו להגריל לצורך הוולידציה, כדי למצוא מסווג ששגיאת ההכללה שלו היא לכל היותר  $\epsilon$  מעל שגיאת ההכללה של המסווג הטוב ביותר, בסיכוי לפחות  $1 - \delta$ . (מותר להשתמש בסימון  $O(\cdot)$ ):

PAC - Non Realizable case (Agnostic)

$$P(L_D(h_S) - L(h^*) < \epsilon) \geq 1 - |H| \cdot 2e^{-\frac{\epsilon^2 m}{2}} \geq 1 - \delta$$

$$m = \frac{2 \log(2 \cdot 5k) + \log \frac{1}{\delta}}{\epsilon^2}$$