Machine Learning from Data – IDC
HW6 – Theory
Nitai Aharoni - 203626742

1 VC-Dimension

Compute the VC-dimension of the following hypothesis classes:

1. (10 pt.) Assume the instance space X satisfies |X| = ∞. The space of binary hypotheses, which given a training set, returns the target y of $\mathbf{x}$ if the pair $(\mathbf{x}, y)$ was observed in the training set, and +1 otherwise. Formally, compute VC(H) of
$\mathrm{H} = \{h : \mathrm{X} \to \{-1, +1\} : h \text{ equals } -1 \text{ on a finite subset of X and } +1 \text{ elsewhere}\}$.

- נתון מרחב X כך ש- $|X| = \infty$
- נגדיר את d להיות ה-training set כך ש-$d \subseteq X$ (תת קבוצה סופית)
- מהגדרת H, h תנתץ באופן מושלם את קבוצת האימון d
- היות ו- $d \subseteq X$ אז $|d| \leq |X| = \infty$ *ולכן עבור כל תת קבוצה* d *בכל גודל קיימת* h *שתנתץ אותה באופן מושלם*
- לכן $VC(H) = \infty$

2. (15 pt.) n-Interval classifiers of length ≥ 2. Let X = R,
$\mathrm{H} = \{x \to +1 \iff x \in [a_1, b_1] \cup [a_2, b_2] \cup \cdots \cup [a_n, b_n] : a_1 + 2 \leq b_1, \ldots, a_n + 2 \leq b_n\}$.

- נסביר: VC(H)=2n
  - *נראה* $VC(H) \geq 2n$
    - נתון מרחב X כך ש- $X = R$
    - נגדיר את d להיות ה-training set כך ש-$d \subseteq X$ (קבוצה סופית)
    - מוגדר לנו classifier המורכב מ-n אינטרוולים על הישר באורך גדול שווה ל-2
    - נחלק את הנקודות ב-d לזוגות - כך שעבור כל $x \epsilon d$ הזוגות (x_i, x_{i+1}) ממוקמים על הישר במרחק 1 ביניהם ובמרחק 10 מהזוג הבא.
    - עבור כל זוג ניתן לקבוע אינטרוול *[ai, bi]* - כך ש-*ai + 2 ≤ bi* , האינטרוול מנפץ את הזוג, וגם אינטרוול זה לא משפיע על ניפוץ הזוגות השכנים.
    - לכן בעזרת n אינטרוולים כאלה ניתן לנפץ 2n נקודות ולכן $VC(H) \geq 2n$
  - *נראה* $VC(H) < 2n + 1$
    - נוסיף נקודה אחת וננסה לנפץ
    - כעת, היות ויש לנו $2n + 1$ נקודות (כלומר מספר אי זוגי של נקודות) ננסה לנפץ שלישיית נקודות באמצעות אינטרוול אחד
    - נשים לב שאם נרצה לסווג את הנקודות באופן הבא: <u>+,-,+</u> - לא ניתן לעשות זאת באמצעות אינטרוול אחד
    - כמובן שבאותו אופן גם רביעיית נקודות וכך הלאה לא ניתן לנפץ באמצעות אינטרוול אחד.
  - לכן VC(H)=2n

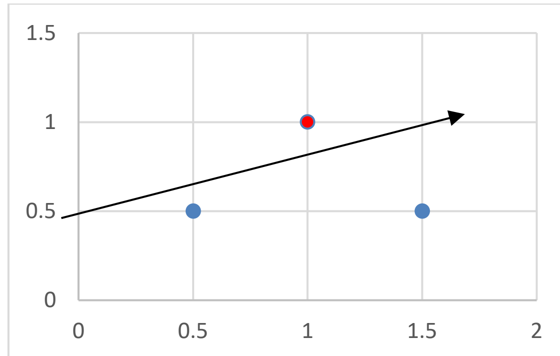
3. (20 pt.) Linear classifiers in the plain. Let $\mathrm{X} = \mathrm{R}^2$,
$\mathrm{H} = \{(x_1, x_2) \text{->} \quad +1 \quad w_1x_1 + w_2x_2 + b > 0 \quad -1 \quad w_1x_1 + w_2x_2 + b \leq 0$
{

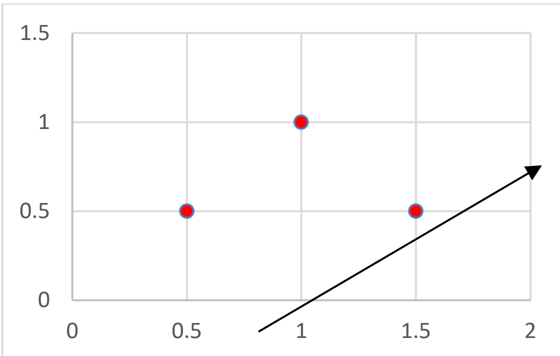
$: w_1, w_2, b \in \mathbb{R}\}$

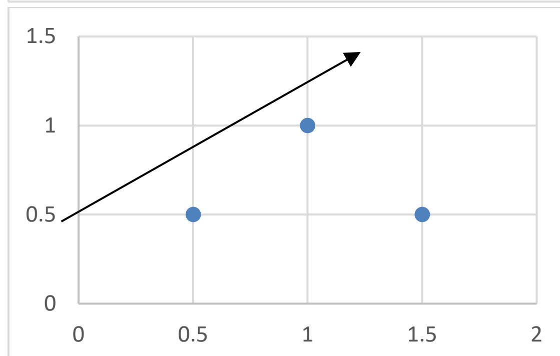
Show that VCdim(H) = 3:

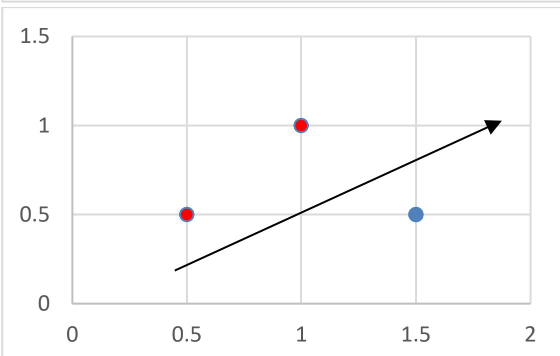
(a) Find a set of size 3 that H shatters.

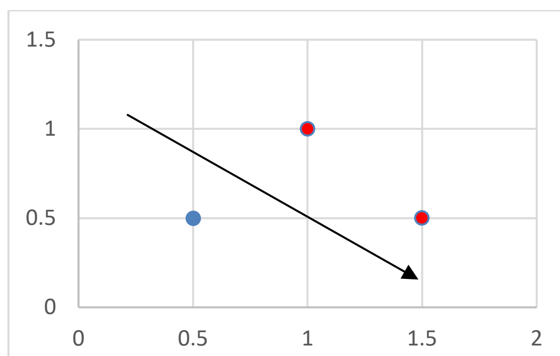
- נראה פיזור של 3 נקודות כאלה כך ש-H מנפץ אותן

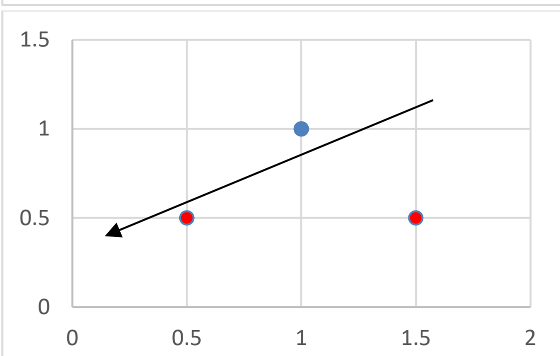
- 

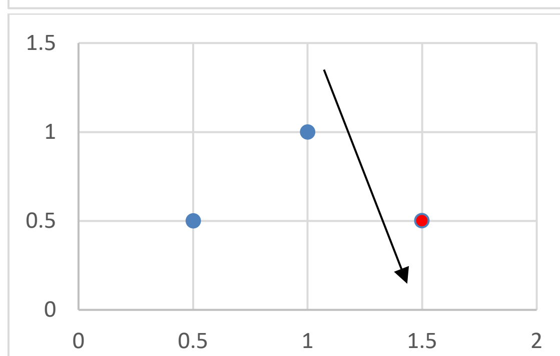
- 

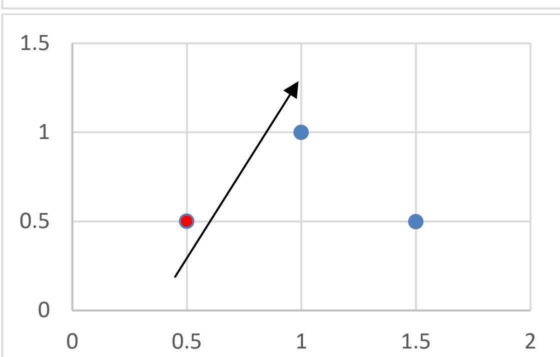
- 

- 

- 

- 

- 

- 

(b) Show that no set of size 4, $A = (\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4)$, $\mathbf{z}_i \in \mathrm{R}^2$ can be shattered by H.
Guidance: First prove the following lemma:
Lemma 1. *Suppose a linear classifier h obtains prediction $y \in \{-1, +1\}$ on a set of points $\mathbf{z}, \mathbf{z'} \in \mathrm{R}^2$ $(h(\mathbf{z}) = h(\mathbf{z'}) = y)$. Then it also obtains the same prediction on any intermediate point. Namely,*
$\forall \alpha \in [0,1] \qquad h((1-\alpha)\mathbf{z} + \alpha\mathbf{z'}) = y.$
And use it in each of the following 3 possible cases:

- The convex hull of A forms a line.
- The convex hull of A forms a triangle.
- The convex hull of A forms a quadrilateral.

- יהי $z=(x,y),\ z'=(x',y')$
- נתון $h(z)=h(z')=y$
- נראה כי לכל $\alpha \in [0,1]$ מתקיים $h((1-\alpha)\mathbf{z} + \alpha\mathbf{z'}) = y$
- $h\big((1-\alpha)\mathbf{z} + \alpha\mathbf{z}'\big) = h\big((1-\alpha)(\mathrm{x},\mathrm{y}) + \alpha(\mathrm{x}',\mathrm{y}')\big) = \mathrm{h}\big((1-\alpha)\mathrm{x} + (1-\alpha)\mathrm{y} + \mathrm{ax}' + \mathrm{ay}'\big) = \mathrm{w}_1\big((1-\alpha)\mathrm{x} + \mathrm{ax}'\big) + \mathrm{w}_2\big((1-\alpha)\mathrm{y} + \mathrm{ay}'\big) + b = \mathrm{w}_1\big((1-\alpha)\mathrm{x} + \mathrm{ax}'\big) + \mathrm{w}_2\big((1-\alpha)\mathrm{y} + \mathrm{ay}'\big) + (1-\alpha)b + ab = (1-\alpha)(w_1x + w_2y + b) + a(w_1x' + w_2y' + b) = (1-\alpha)h(z) + ah(z') = (1-\alpha)y + ay = y$

- יהיו 4 נקודות: $z_{1=}(x_1,y_1), z_{2=}(x_2,y_2), z_{3=}(x_3,y_3), z_{4=}(x_4,y_4)$
- נשתמש בכך עבור שלושת המקרים:
  - קו ישר:
    - מהיות קו ישר אזי לכל i, $y_i = 0$
    - ונניח $x_1 < x_2 < x_3 < x_4$
    - ו- $h(z_1) = 1, h(z_2) = -1, h(z_3) = 1, h(z_4) = 1$
    - לפי הלמה שהוכחנו אם $h(z_1) = 1$ וגם $h(z_3) = 1$ אזי בהכרח $h(z_2) = 1$ בסתירה
    - לכן הקבוצה A לא יכולה להיות מנופצת ע"י H
  - משולש:
    - כאשר z1,z2,z3 יוצרות משולש ו-z4 מוכלת במשולש
    - ו- $h(z_1) = 1, h(z_2) = 1, h(z_3) = 1, h(z_4) = -1$
    - אם z4 נמצא על אחת מצלעות המשולש - בה"כ על הצלע שבין z1 ל-z3:
      - לפי הלמה שהוכחנו אם $h(z_1) = 1$ וגם $h(z_3) = 1$ אזי בהכרח $h(z_4) = 1$ בסתירה
      - לכן הקבוצה A לא יכולה להיות מנופצת ע"י H
    - אם z4 נמצא בתוך המשולש:
      - אזי קיימת נקודה z' כך ש-z' שנמצאת על אחת מצלעות המשולש, בה"כ בין z1 ל-z3, כך ש z4 נמצאת על הישר שבין z' לבין z2.
      - לפי הלמה $h(z_1) = 1$ וגם $h(z_3) = 1$ אזי בהכרח $h(z') = 1$
      - ולכן לפי הלמה שהוכחנו אם $h(z') = 1$ וגם $h(z_2) = 1$ אזי בהכרח $h(z_4) = 1$ בסתירה
      - לכן הקבוצה A לא יכולה להיות מנופצת ע"י H

- מרובע:
  - כאשר z1,z2,z3,z4 יוצרות מרובע
  - מדובר בבעיית XOR אותה לא ניתן לנפץ בעזרת מפריד לינארי כפי שהראינו בכיתה

2 Learning Conjunctions of Literals

(30 pt.) Let X = $\{0,1\}^n$ (all Boolean strings of length *n*), let *C* = H = the set of all conjunctions on X (e.g. $x_1 \wedge \neg x_3 \wedge x_n$ is in *C* and H). Define an algorithm *L* so that *C* is PAC-learnable by *L* using H. Prove all your steps.

- אלגוריתם:
  - נגדיר $h = x_1 \wedge \overline{x_1} \wedge x_2 \wedge \overline{x_2} \wedge \ldots\ \ x_n \wedge \overline{x_n}$
  - נעבור על כל ה-instances:
  - נתבונן ב-instance ה-i.
    - אם $C(instance_i)$=1, וכל האלמנטים ב-h הנוכחי קיימים ב-attributes של ה-instance ← אז h הוא קונסיסטנטי עם ה-instance הנוכחי.
    - אם $C(instance_i)$=1, ואחד או יותר מהאלמנטים ב-h הנוכחי לא מופיעים בו ← אז נסיר מ-h את כל ה-attributes שקיימים ב-h ולא ב-instance.
  - אחרי מעבר על כל ה-instances נקבל h שהוא קונסיסטנטי עם C
  - ואם קיבלנו h כזה אז C הוא PAC-learnable באמצעות L.

- נכונות:
- לומד קונסיסטנטי:
  - עבור $C(instance_i)$=1 היות ומדובר בביטוי AND, כל אחד מה-attributes של אותו instance אמורים לחזות את ערכו האמיתי – ולכן אם קיימים ביטויים ב-h שלא מסכימים עם ה-attributes של ה-instance, הם יוסרו מ-h
  - עבור $C(instance_i)$=0 היות ומדובר בביטוי AND, אם כל הביטויים ב-h יופיע ב-attributes של ה-instance אז h יהיה קונסיסטנטי עם אותו instance.
    - נניח בשלילה שעבור instance מסוים כל הביטויים ב-h קיימים ב-attributes של ה-instance אזי h לא קונסיסטנטי עם C. אבל נתון ש-H=C ולכן קיים $h\epsilon H$ כך ש-h קונסיסטנטי עם C בסתירה.
    - אם עברנו על כל ה-instances ו-h קונסיסטנטי עם כולם אז h קונסיסטנטי עם C.
- sample complexity:
  - $|H| = 3^n$ - ה-attribute קיים בביטוי כ-true, קיים כ-false או לא קיים. n=attribute num
  - ולכן $m \geq \frac{n}{3} ln3 + \frac{1}{\varepsilon}\ln(\frac{1}{\delta})$ כלומר sample complexity פולינומיאלי
- בדיקת זמן ריצה:
  - עבור כל instance נשווה את ה-attributes שלו אל מול האלמנטים ב-h – ניתן לבדוק ב-O(n)
  - נעבור על כל ה-instances כלומר m
  - סה"כ זמן ריצה O(n*m) – פולינומיאלי
- לכן C הוא PAC-learnable באמצעות L

3 (Almost) PAC-learnability

(25 pt.) Let C denote the class of all possible target concepts defined over a set of instances X. Suppose that H is a space of binary hypotheses containing the constant concept $c1$ defined by $c1(x) = +1$ for all $x \in$ X, and having the property that C \ {c1} is PAC-learnable by an algorithm L using H with sample complexity m(δ, ε). Provide a learning algorithm L’ that uses L, so that C (including c1) is PAC-learnable by L’ using H with sample complexity max{m(δ, ε), $\left\lceil \frac{\log(\frac{1}{\delta})}{\varepsilon} \right\rceil$ }. Prove all your steps.

- אלגוריתם L’:
  - נעבור על כל ה-instances ב-D ונבדוק:
  - אם קיים $x\epsilon D$ כך ש-$c(x) = 0$ ← אז נריץ את L הידוע בתור PAC-learnable ונחזיר את h
  - אם לכל $x\epsilon D$ מתקיים $c(x) = 1$ ← אז נחזיר h=1

- נראה L’ הינו PAC-learnable:
  - ידוע ש-L הינו PAC-learnable עם sample complexity m(δ, ε) ולכן אין צורך להראות במקרה בו קיים $x\epsilon D$ כך ש-$c(x) = 0$.
  - נראה עבור המקרה שלכל $x\epsilon D$ מתקיים $c(x) = 1$
    - לומד קונסיסטנטי:
      - ידוע שלכל $x\epsilon D$ מתקיים $c(x) = 1$ ולכן h=1 קונסיסטנטי עם c
    - sample complexity:
      - $|H| = 1$
      - ולכן $m \geq \frac{1}{\varepsilon}(\ln|H| + ln\left(\frac{1}{\delta}\right)) \geq \frac{1}{\varepsilon}\ln(\frac{1}{\delta})$ כלומר sample complexity פולינומיאלי
    - בדיקת זמן ריצה:
      - עבור כל $x\epsilon D$ נבדוק את $c(x)$ לכן מדובר ב-m בדיקות O(m)
- לכן C הוא PAC-learnable באמצעות L’

- ידוע שבמקרה של c_1 מתקיים $m \geq \frac{1}{\varepsilon}\ln(\frac{1}{\delta})$
- לכן ה-sample complexity הוא $\max\{m(\delta, \varepsilon), \left\lceil \frac{\log(\frac{1}{\delta})}{\varepsilon} \right\rceil\}$