HW5 – Theory + SVM

Nir Shoham 322657073

Ron Aharonson 211741921

1. PAC Learning and VC dimension (30 pts)

   Let $X = \mathbb{R}^2$. Let

   $$C = H = \left\{ h(r_1, r_2) = \left\{ (x_1, x_2) \middle| \begin{matrix} x_1^2 + x_2^2 \geq r_1 \\ x_1^2 + x_2^2 \leq r_2 \end{matrix} \right\} \right\}, \text{ for } 0 \leq r_1 \leq r_2,$$

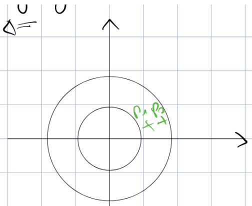
   the set of all origin-centered rings.

   a. (8 pts) What is the $VC(H)$? Prove your answer.

נוכיח כי $VC(H) \geq 2$. נראה כי עבור כל 2 נקודות קיימת טבעת H אשר מנתצת אותם בהתאם לקונספט:

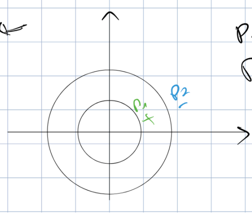
יהיו $p_1 = (x_1, x_2), p_2 = (y_1, y_2)$ נקודות שונות, ומתקיים כי $y_1^2 + y_2^2 > x_1^2 + x_2^2$:

$r_1 \leq x_1^2 + x_2^2$
$r_2 \geq y_1^2 + y_2^2$ ⇐



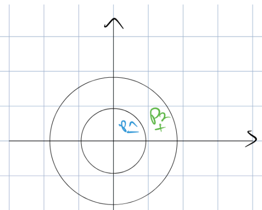
$p_1 - (+)$
$p_2 - (+)$

$r_1 \leq x_1^2 + x_2^2 \leq r_2$ ⇐
$x_1^2 + x_2^2 \leq r_2 \leq y_1^2 + y_2^2$



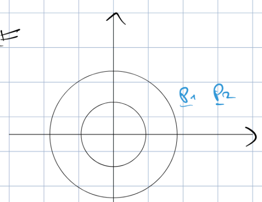
$p_1 - (+)$
$p_2 - (-)$

$r_1 > x_1^2 + x_2^2$ ⇐
$r_1 \leq y_1^2 + y_2^2 \leq r_2$



$p_1 - (-)$
$p_2 - (+)$

$r_1 \leq r_2 < x_1^2 + x_2^2$ ⇐
או
$y_1^2 + y_2^2 < r_1 \leq r_2$



$p_1 - (-)$
$p_2 - (-)$

נשים לב כי $VC(H) < 3$: נניח כי קיימת נקודה $p_3 = (z_1, z_2)$. ונניח בה"כ כי

$$z_1^2 + z_2^2 > y_1^2 + y_2^2 > x_1^2 + x_2^2$$

עבור קונספט שמקיים $p_2 = -$, $p_1, p_3 = +$, לא קיימת היפותזה המנתצת את הקבוצה הזו.

לכן $VC(H) = 2$.

b. (14 pts) Describe a polynomial sample complexity algorithm L that learns C using H. State the time complexity and the sample complexity of your suggested algorithm. Prove all your steps.

נביט באלגוריתם הבא:

עבור כל $instance$ שמסווג חיובי, נמדוד את מרחקו מראשית הצירים (ניתן לביצוע בעזרת פיתגורס). את r_1 נקבע להיות המרחק המינימלי שמצאנו וr_2 כמרחק הגדול ביותר. נשים לב כי המודל consistent מהיות ולא יכולות להיות נקודות שלילות במרחב שבין r_1 ל r_2. ובנוסף ברגע שמצאנו נקודה במרחב נהיה חייבים לכלול אותה בתוך השטח המהווה סיווג חיובי אחרת הדאטה לא יהיה עקבי, לכן ניאלץ לקחת בחשבון את הנקודות הקיצוניות ביותר בכל מצב.

- זמן ריצה: ראשית נשים לב כי כל פעולה מתבצעת בסיבוכיות של $O(1)$ וסה"כ ישנן $O(m)$ (m כמות הinstances) פעולות. לכן סה"כ סיבוכיות של $O(m)$. נשים לב כי לא נוכל לבצע זאת בסיבוכיות טובה יותר כי בכל מקרה נצטרך לעבור על כל הדאטה.
- סיבוכיות הדגימה: יהי $\varepsilon, \delta > 0$. עבור קונספט $c \in C$, נגדיר r_1, r_2 כרדיוסים של הקונספט, הקובעים את גבולות הגזרה של הדגימות החיוביות עבור קונספט זה. כעת, נגדיר c^ε, ועבורו נגדיר $r_1^\varepsilon, r_2^\varepsilon$. נגדירם כך ש r_1^ε הוא הרדיוס שנוצר כאשר נקצץ $\frac{\varepsilon}{2}$ מהשטח הנוצר בין המעגלים הנוצרים ע"י הרדיוסים מכל רדיוס. כלומר $\frac{\varepsilon}{2}$ לכיוון ראשית הצירים מr_2 ו$\frac{\varepsilon}{2}$ לכיוון ההפוך מr_1. נבהיר שבשטח אנו מתכוונים לשטח בהיבט הגאוסייני של התפלגות. פורמלית: נגדיר $r_1 \le a, b \le r_2$ כך ש:

$$S_1 = \{(x_1, x_2) | a \le \sqrt{x_1^2 + x_2^2} \le r_2\}$$

$$S_2 = \{(x_1, x_2) | r_1 \le \sqrt{x_1^2 + x_2^2} \le b\}$$

$$r_2^\varepsilon = \inf\left\{a \middle| \pi(S_1) \le \frac{\varepsilon}{2}\right\} \quad , \qquad r_1^\varepsilon = \sup\left\{b \middle| \pi(S_2) \le \frac{\varepsilon}{2}\right\}$$

כעת נפריד למקרים:

- קיימת נקודה בS_1 או S_2. במקרה זה הטעות בכל מקרה קטנה שווה לאפסילון לכן הטעות קטנה מספיק ואנו עומדים בתנאי.
- לא קיימת נקודה בS_1 או S_2. ידוע כי ההסתברות למקרה זה עבור m דגימות בלתי תלויות הוא $\left(1-\frac{\varepsilon}{2}\right)^m + \left(1-\frac{\varepsilon}{2}\right)^m$ ונראה כי:

$$2\left(1-\frac{\varepsilon}{2}\right)^m \leq 2\exp\left(-\frac{\varepsilon m}{2}\right)$$

כעת נרצה שההסתברות הזו תהיה קטנה מδ, לכן נחלץ את m:

$$2\exp\left(-\frac{\varepsilon m}{2}\right) \leq \delta \rightarrow \exp\left(-\frac{\varepsilon m}{2}\right) \leq \frac{\delta}{2} \rightarrow -\frac{\varepsilon m}{2} \leq \ln\left(\frac{\delta}{2}\right) \rightarrow m \geq 2 \cdot \frac{-\ln\left(\frac{\delta}{2}\right)}{\varepsilon} \rightarrow$$
$$m \geq 2 \cdot \frac{\ln\left(\frac{2}{\delta}\right)}{\varepsilon} \rightarrow$$

עבור כמות דגימות m העומדת בתנאי הנ"ל, ההסתברות שהטעות תהיה גדולה מאפסילון היא δ.

In class we saw a bound on the sample complexity when H is finite.

$$m \geq \frac{1}{\varepsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right)$$

When $|H|$ is infinite, we have a different bound:

$$m \geq \frac{1}{\varepsilon}\left(4\log_2\frac{2}{\delta} + 8VC(H)\log_2\frac{13}{\varepsilon}\right)$$

c. (8 pts) You want to get with 95% confidence a hypothesis with at most 5% error. Calculate the sample complexity with the bound that you found in b and the above bound for infinite $|H|$. In which one did you get a smaller m? Explain.

לפי הנוסחה בסעיף ב':

$$m \geq 2 \cdot \frac{\ln\left(\frac{2}{0.05}\right)}{0.05} \approx 148$$

עבור הנוסחה הנתונה:

$$m \geq \frac{1}{0.05}(4\log\left(\frac{2}{0.05}\right) + 8 \cdot 2\log\left(\frac{13}{0.05}\right) \approx 2993$$

נשים לב כי הנוסחה הראשונה הדוקה יותר כי היא כוללת בתוכה ידע מקדים על H. לעומת זו הנוסחה השנייה היא עבור כל H ללא שום מידע ומניחה רק את *VC(H)*.

2. VC dimension (20 pts)

Let $X = \mathbb{R}$ and $n \in \mathbb{N}$.

Define "x-node decision tree" for any $x = 2^n - 1$ to be a full binary decision tree with x nodes (including the leaves).

Let H_m be the hypothesis space of all "x-node decision tree" with $n \le m$.

a. (5 pts) What is the $VC(H_3)$? Prove your answer.

נשים לב כי H_3 הוא מרחב ההיפותזות עבור עצי ההחלטה המתוארים כאשר $n \le 3$ לכן

$$x \le 2^3 - 1 = 7$$

נשים לב כי בעץ המתואר (עץ בינארי מלא) ישנם ארבעה עלים.

כעת נוכיח כי $VC(H_3) \ge 4$:

נשים לב כי עבור כל קונספט, קיימת היפותזה אשר מנתצת אותו ועקבית איתו, מהיות וארבעת הדגימות מופרדות ל-4 עלים שונים, לכן נוכל לשנות את ההיפותזה כך שיתאימו עם כל קונספט.

נשים לב כי $VC(H_3) < 5$ מהיות וישנם ארבע עלים בעץ לכל היותר, מהיות וישנן 5 דגימות, יהיו 2 דגימות באותו העלה. במידה ו2 דגימות אלה לא יסכימו על אותו לייבל, הדבר אומר שלא ניתן לנתץ את הקבוצה.

ניתן דוגמה למצב שבו לא ניתן להפריד בין דגימות שלא מסכימות על אותו לייבל מאותו עלה:

$$x_1 = -1, x_2 = 0, x_3 = 3, x_4 = 4, x_5 = 6$$

$$C = \{x \in R | 0 \le x \le 2 \ or \ 4 \le x \le 5$$

אם שתי דגימות סמוכות נמצאות באותו עלה, הן אינן מסכימות ולכן לא ניתן לנתץ. אם הן אינן סמוכות (ומסכימות), גם הדגימה שבינהן נמצאת באותו טווח לכן תהיה באותו עלה ולא יהיה ניתן לנתץ.

b. (15 pts) What is the $VC(H_m)$? Prove your answer.

נשים לב כי $VC(H_m) \ge 2^{m-1}$. מהיות ובעץ בינארי מלא בעל $2^m - 1$ צמתים יש 2^{m-1} עלים. כפי שתיארנו מקודם, נוכל לבנות היפותזה שמפרידה בין כל הדגימות כאשר כל דגימה נמצאת בעלה ולכן נוכל לנתץ את הקבוצה מהיות ונוכל להכריז עבור כל עלה בנפרד על סיווגו לפי הדגימה שלו ולא יהיה חוסר הסכמה.

נשים לב כי י $VC(H_m) < 2^{m-1} + 1$, מהיות ותהיינה 2 דגימות באותו עלה, ועבור קונספט שבו כל שתי דגימות רצופות מקבלות סיווג שונה כפי שהראינו בסעיף הקודם, לא נוכל להפריד בין סיווגים. אם שתי דגימות סמוכות נמצאות באותו עלה, הן אינן מסכימות ולכן לא ניתן לנתץ. אם הן אינן סמוכות (ומסכימות), גם הדגימה שבינהן נמצאת באותו טווח לכן תהיה באותו עלה ולא יהיה ניתן לנתץ מהיות והיא לא מסכימה איתן.

לכן $VC(H_m) = 2^{m-1}$.

3. Kernels and mapping functions (25 pts)
    a. (20 pts) Let $K(x,y) = (x \cdot y + 1)^3$ be a function over $\mathbb{R}^2 \times \mathbb{R}^2$ (i.e., $x, y \in \mathbb{R}^2$).

       Find ψ for which $K$is a kernel. (It may help to first expand the above term on the right-hand side).

$$x = (x_1, x_2)\ , y = (y_1, y_2)$$

$$(x \cdot y + 1)^3 = (x_1 y_1 + x_2 y_2 + 1)^3 =$$
$$(x_1y_1)^3 + (x_2y_2)^3 + 1 + 3 \cdot (x_1y_1)^2 \cdot x_2y_2 + 3(x_1y_1)^2 + 3(x_2y_2)^2 \cdot x_1y_1$$
$$+3(x_2y_2)^2 + 3x_1y_1 + 3x_2y_2 + 6x_1y_1x_2y_2 \rightarrow$$
$$\psi(x) = (x_1^3, x_2^3, \sqrt{3}x_1^2x_2, \sqrt{3}x_1^2, \sqrt{3}x_2^2, \sqrt{3}x_2^2x_1, \sqrt{6}x_1x_2, \sqrt{3}x_1, \sqrt{3}x_2, 1)$$

       (2 pts) What did we call the function ψ in class if we remove all coefficients?

       Full rational variety of degree 3

    b. (3 pts) How many multiplication operations do we save by using $K(x,y)$ versus $\psi(x) \cdot \psi(y)$?

נשים לב כי:

$$(x \cdot y + 1)^3 = (x_1y_1 + x_2y_2 + 1)^3 \rightarrow$$

נכפיל בתוך הסוגריים פעמיים, ועוד 2 הכפלות של הביטוי בסוגריים – סה"כ 4.

לעומת:

בכל פונקציה ישנם פעולות כפל:

$$2 + 2 + 3 + 2 + 2 + 3 + 2 + 1 + 1 = 18$$

ישנם 10 איברים בכל וקטור לכן אם נכפול בין הוקטורים בנוסף לפעמיים התוצאה שראינו מקודם נקבל עוד 10 פעולות. סהכ 10+18+18=46

חסכנו 4-46=42 הכפלות.

4. <u>Lagrange multipliers (15 pts)</u>

Let $f(x,y) = 2x - y$. Find the minimum and the maximum points for f under the constraint $g(x,y) = \frac{x^2}{4} + y^2 = 1$.

$$L(x,y,\lambda) = 2x - y + \lambda\left(\frac{x^2}{4} + y^2 - 1\right) = 2x - y + \frac{\lambda x^2}{4} + \lambda y^2 - \lambda$$

$$\frac{\partial}{\partial x}L(x,y,\lambda) = 2 + \frac{\lambda x}{2} = 0 \rightarrow x = -\frac{4}{\lambda}$$

$$\frac{\partial}{\partial y}L(x,y,\lambda) = -1 + 2\lambda y = 0 \rightarrow y = \frac{1}{2\lambda}$$

$$\frac{\partial}{\partial \lambda}L(x,y,\lambda) = \frac{x^2}{4} + y^2 - 1 = 0 \rightarrow \frac{\frac{16}{\lambda^2}}{4} + \frac{1}{4\lambda^2} - 1 = \frac{4}{\lambda^2} + \frac{1}{4\lambda^2} - 1 = 0$$

$$\lambda^2 = 4.25 \rightarrow \lambda = \pm\frac{\sqrt{17}}{2} \rightarrow$$

$$if\ \lambda = \frac{\sqrt{17}}{2} \rightarrow x = -\frac{8}{\sqrt{17}}, y = \frac{1}{\sqrt{17}} \rightarrow f(x,y) = -\sqrt{17}$$

$$if\ \lambda = -\frac{\sqrt{17}}{2} \rightarrow x = \frac{8}{\sqrt{17}}, y = -\frac{1}{\sqrt{17}} \rightarrow f(x,y) = \sqrt{17}$$

לכן מתקיים כי נקודת המינימום היא $\left(-\frac{8}{\sqrt{17}}, \frac{1}{\sqrt{17}}\right)$ והמקס' - $\left(\frac{8}{\sqrt{17}}, -\frac{1}{\sqrt{17}}\right)$.

5. See notebook exercise (10 pts)

Done