למידה חישובית ממידע: תרגול 9 – תיאוריה

אין ארוחות חינם – No Free Lunch

הגדרות:

- $Acc_G(L,c)$= הדיוק המוכלל של היפותזה (שנוצרה מלומד L) על קונספט c = הדיוק של L על דגימות שאינן מה-training data.
  הדיוק על ה-training, מבלי להתחשב בדיוק המוכלל אינו מעניין – בקלות נוכל להגיע אתו ל-100%.
- C = קבוצת כל הקונספטים האפשריים על מרחב האינסטנסים, y=c(x) (קונספט הוא מיפוי מדאטה ללייבל)

משפט: לכל learner, L, מתקיים:
$$\frac{1}{|C|}\sum_{c\in C} Acc_G(L,c)=\frac{1}{2}$$

כלומר: ממוצע הדיוק המוכלל מעל כל הקונספטים ב-C הוא 0.5 עבור כל התפלגות נתונה D על מרחב הדגימות X ו-training set בגודל n.

No Free Lunch Theorem

הרעיון שעומד מאחורי העיקרון הזה בא להראות לנו שעל מנת להצליח ללמד את המודל שלנו להכליל בצורה בעלת משמעות, אנחנו חייבים להגביל את מרחב ההיפותזות שלנו איכשהו. כלומר, אם כל ההיפותזות האפשריות הן בסבירות זהה להיות הקונספט, אז לא משנה איך נלמד, הדיוק הממוצע שלנו בהכללה (על כל הקונספטים) יהיה $\frac{1}{2}$. הסיבה לכך היא שעבור מודל מסוים L, אם הדיוק שלו על קונספט (דגימות שלא ראינו) c הוא $\frac{1}{2}+\delta$, קיים קונספט אחר (כזה שיחלק את הדגימות בצורה 'הפוכה') שעבורו הדיוק של L יהיה $\frac{1}{2}-\delta$. לכן הדיוק הממוצע עבור כל מודל L:

$$\frac{1}{|C|}\sum_{c\in C} Acc(L,c)=\frac{1}{2}$$

Proof: Given any training set S:

For every concept c where $Acc_G(L,c)=\frac{1}{2}+\delta$,

there is a concept c′ where $Acc_G(L,c')=\frac{1}{2}-\delta$

$$\forall x\in S, c'(x)=c(x)=y \quad \forall x\notin S, c'(x)=\neg c(x)$$

הכללת NFL: לכל שני לומדים L1, L2

אם קיים קונספט c כך שהדיוק המוכלל של L1 על c גדול מהדיוק המוכלל של L2 על c, אזי

קיים קונספט c’ כך שהדיוק המוכלל של L2 על c’ גדול מהדיוק המוכלל של L1 על c’.

For any two learner L1, L2

If ∃ learning problem c s.t $Acc_G(L_1,c)>Acc_G(L_2,c)$

Then ∃ learning problem c′ s.t $Acc_G(L_2,c')>Acc_G(L_1,c')$

דוגמה פשוטה עבור NFL: עם שני קונספטים

L1=

x1	x2	x3	ŷ
0	0	0	0
0	0	1	0
1	1	0	1
0	1	0	1
1	1	1	0
0	1	1	1
1	0	0	0
1	0	1	1

Training

Test

L2=

x1	x2	x3	ŷ
0	0	0	0
0	0	1	0
1	1	0	1
0	1	0	1
1	1	1	1
0	1	1	0
1	0	0	1
1	0	1	0

If the concept is (0,0,1,1,0,1,0,0) then L1 is more accurate with 75% and L2 has 25%

If the concept is (0,0,1,1,1,0,1,1) then L2 is more accurate with 75% and L1 has 25%

המסקנה מ-NFL:

- לא לצפות שהאלגוריתם הלומד האהוב עלינו תמיד יהיה הכי טוב
- אלגוריתם פשוט יכול להיות טוב יותר לפעמים (המורכבים יותר יובילו ל-overfit)
- מומלץ לנסות גישות שונות

כאמור, על מנת להיות מסוגלים להכליל בצורה משמעותית, נהיה חייבים להגביל (להקטין) את מרחב החיפוש שלנו - מרחב ההיפותזות, על ידי הנחות מסוימות.
מעבר לסיבה המקורית לכך, יתרון נוסף הוא שהחיפוש יהיה מהיר ופשוט יותר, אם אנחנו מצמצמים את האפשרויות שלנו. חסרון לכך יהיה העובדה שאם אנחנו מגבילים את מרחב ההיפותזות באופן כזה שלא משקף באמת את המציאות, ייתכן שלא נמצא היפותזה שתיתן לנו אפס טעויות על הקונספט.
דוגמאות להגבלה של מרחב ההיפותזות ראינו במהלך הקורס - חיפשנו רק פונקציות לינאריות ברגרסיה לינארית, בנינו עצי החלטה כשבכל קודקוד יש שאלה רק לגבי פיצ'ר בודד (זו הסיבה שה'מפרידים' שלנו בעצי החלטה מקבילים לצירים), וכו'.

Learning Complexity

חזרה להערכת הטעות

עד כה כדי לקבל מושג לגבי הטעות האמיתית שלנו, השתמשנו ב-test set - שאליו התייחסנו כדגימות שהמודל שלנו לא ראה, ולכן לא למד, ושמשקפות את המציאות (הקונספט). אם ה-test set שלנו הוא כל יותר הדגימות מ-X, אז הטעות שלנו על ה-test set היא בדיוק הטעות האמיתית שלנו. כמובן שזה לא מצב ריאלי, ולכן ככל שנגדיל את גודל ה-test set שלנו, נתקרב להערכה טובה יותר של הטעות האמיתית על הקונספט.

עבור test set בגודל $|S|$, ומספר טעויות r, שיעור הטעות הכללית היא (הדגימות בלתי תלויות):

$$p=\frac{r}{|S|}$$

בנוסף, נגדיר את שגיאת התקן להיות:

$$se=\sqrt{\frac{p(1-p)}{n}}$$

כתלות ב-$|S|$, נוכל להתחייב למרווח טעות מסוים: "בבטחון של $x\%$, הטעות האמיתית קטנה מ-$\left(\frac{r}{|S|}+\varepsilon\right)$".

מרווח הטעות הזה נקרא **רווח סמך** - CI (Confidence Interval):

$$CI=p\pm 2(se)$$

הטעות האמיתית של המודל - True Error

נעצור מלדבר על מודלים ואלגוריתמים לומדים, ונעבור לדבר על התיאוריה שמאחורי הלמידה. נושא זה נקרא PAC Learning - Probably Approximately Correct. כלומר, השאיפה הכללית שלנו היא להגיע **בהסתברות גבוהה** (probably) **להערכה טובה של המציאות** (approx. correct).
על מנת לעשות זאת נצטרך למדוד את הטעות האמיתית שלנו, ובנוסף להבין כמה דגימות דרושות לנו על מנת ללמוד בצורה איכותית מספיק. ניתן לכך הגדרות מדויקות בהמשך.

כמו שעשינו עד כה, אנחנו מנסים ללמוד קונספט (דיכוטומיה, פונקציה מסווגת) c. יש לנו דגימות (training set) במימד הדגימות X, ואנחנו מנסים למצוא היפותזה, במרחב ההיפותזות $h \in H$, שתסכים (תהיה קונסיסטנטית, עד כמה שאפשר) עם הקונספט שלנו:

$$samples\ x \in X \qquad concept\ c \in C \subseteq P(X) \qquad h \in H \subseteq P(X)$$

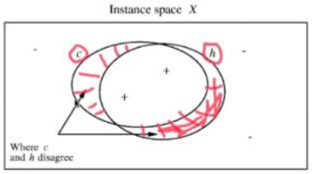
נרצה למצוא את ההיפותזה המתאימה ביותר ל-training data, מתוך הנחה שהוא מייצג טוב את הדאטא במציאות (הדאטא שלא ראינו, הקונספט). הטעות שאנחנו מייצרים על הדאטא נקראת **in-sample error**. כמובן שנשאף להביא אותה למינימום, אך בפועל לא ה-training data הוא מה שמעניין אותנו - אלא הדאטא שלא ראינו. לטעות שהמודל שלנו מייצר על הדאטא החדש, אנחנו קוראים **out-of-sample error**.
באופן אינטואיטיבי, טבעי להגיד שאין לנו דרך לדעת את ה-out-of-sample error - כי הוא על דאטא שלא ראינו עדיין, איך בכלל נוכל למדוד אותו - ופה נמצאת הפואנטה של תיאוריית הלמידה - אנחנו נוכל להגיד דברים על הטעות הזו, על בסיס ה-training data.

נסתכל על דוגמא כדי להבין את הרעיון הכללי. נטיל מטבע הוגן 100 פעמים. נניח שרק 8 פעמים יצא פאלי, והשאר עץ. אנחנו כמובן לא יודעים את תוצאת ההטלה הבאה (לצורך ההמחשה, ההטלה הבאה היא הקונספט - דגימה שלא ראינו). אך בגלל שאנחנו יודעים שרק 8 מתוך ה100 יצאו פאלי, יש לנו תחושה שאנחנו כן יכולים להעריך מה תהיה התוצאה של ההטלה הבאה. אם נעלה את מספר ההטעלות ל-1000, ומתוכם רק 80 יצאו פאלי, התחושה הזאת תתחזק - כלומר ככל שיעלה מספר הדגימות שנראה ב-training, אנחנו נדע יותר על הדגימות שלא ראינו.

נרצה כעת להגדיר את הטעות האמיתית של ההיפותזה על הקונספט (הדאטא שלא ראינו). נניח שקיימת פונקציית התפלגות כלשהי $\mathcal{D}$ על הדגימות X (הדגימות יכולות להתפלג נורמלית, לדוגמא, אך הן לא בהכרח חיות במרחב אוקלידי - לכן לא ממש משנה איזו התפלגות זו, אך הפונקציה מקיימת את התנאים הבסיסיים של פונקציות התפלגות). נגדיר את הטעות להיות **ההסתברות שההיפותזה לא הסכימה עם הקונספט, בנוגע לדגימה רנדומלית כלשהי** $x \in X$:

$$TrueError(h) = error_{\mathcal{D}}(h) = P(c(x) \neq h(x))$$

True Error of a Hypothesis

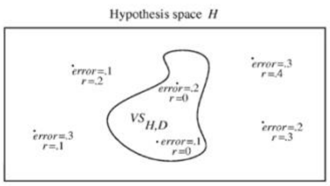


Definition: The **true error** (denoted $error_{\mathcal{D}}(h)$) of hypothesis h with respect to target concept c and distribution $\mathcal{D}$ is the probability that h will misclassify an instance drawn at random according to $\mathcal{D}$.

$$error_{\mathcal{D}}(h) \equiv \Pr_{x \in \mathcal{D}}[c(x) \neq h(x)]$$

Version Spaces

Version Space $VS_{H,D}$**:**
Subset of hypotheses in H consistent with training data D



(r = training error, $error$ = true error)

משפט: אם מרחב היפותזות H הוא סופי, ו-D היא סדרה של m (כאשר m גדול שווה מ-1) דגימות רנדומליות בלתי תלויות של קונספט מטרה c, אזי לכל אפסילון בין 0 ל-1, ההסתברות שקיימת היפותזה ששייכת למרחב ה-VS עם H,D (הגדרה לעיל: תת קבוצה ממרחב ההיפותזות H שקונסיסטנטית עם ה-training data, D), עם טעות אמתית: $error_D(h) > \varepsilon$ היא קטנה מ- $|H|e^{-\varepsilon m}$.

כמה דגימות יספיקו? להלן הוכחת החסם.

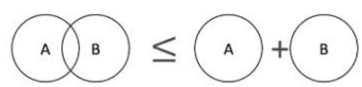
נרצה לחסום את הסיכוי לקבל היפותזה עם שגיאה אמתית גדולה מאפסילון לכן נחסום אותה עם דלתא:

P(1 hyp. w/ error > ε consistent w/ 1 ex.) $< 1 - \varepsilon \leq e^{-\varepsilon}$

P(1 hyp. w/ error > ε consistent w/ m ex.) $< (e^{-\varepsilon})^m = e^{-m\varepsilon}$

P(1 of $|H|$ hyps. w/ error > ε consistent w/ m ex.) $\leq |H|e^{-m\varepsilon}$

* Because of Union Bound



This bounds the probability that any consistent learner will output a hypothesis h with $error_D(h) \geq \varepsilon$

We want this probability to be at most δ

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

דוגמות לשימוש בחסם הזה: (דוגמה מימין ודוגמה נוספת משמאל)

- 1 attribute with 3 values
- 9 attributes with 2 values

$$|X| = 3 \times 2^9$$

- H contains conjunctions of attributes, then

$$|H| = 4 \times 3^9 = 78733$$

- We want to ensure in 95% that our hypothesis will have error < 10%

$$m \geq \frac{1}{0.1}\left(\ln(78733) + \ln\left(\frac{1}{0.05}\right)\right) = 143$$

- Suppose H contains conjunctions of constraints on up to $n=13$ Boolean attributes. Then $|H| = 3^{13} = \boxed{1594323}$
- We want to ensure in 95% that our hypothesis will have error < 5% ε

$1-\delta$
$$m \geq \frac{1}{0.05}\left(\ln(1594323) + \ln\left(\frac{1}{0.05}\right)\right) = 346$$

VC dimension

עד כה דיברנו על מרחב היפותזות סופי, כעת נדבר על מרחב היפותזות אינסופי = למשל מרחב ההיפותזות של מפרידים לינאריים במימד דו מימדי.
מימד ה-VC (Vapnik-Chervonenekis dimension) הוא מדד של קיבולת (סיבוכיות, expressive power, עושר, או גמישות) של אלגוריתם סיווג סטטיסטי, המוגדר **כקרדינליות (הגודל) של הקבוצה הגדולה ביותר של נקודות שהאלגוריתם יכול לנפץ (shatter)**.

נגדיר ניפוץ Shattering: מרחב היפותזות H מנפץ קבוצת נקודות X = { x1, x2, …, xm } (ששייך למרחב הדגימות) אם ורק אם לכל השמה Y = { y1, y2, …, ym } כך ש yi = 1 או yi = -1, קיימת היפותזה h (ממרחב ההיפותזות H) שמקיימת לכל i: h(xi) = yi.

- Let $S(H,X) = \begin{cases} T & H \text{ Shatters } X \\ F & H \text{ Can't shatter } X \end{cases}$
- If $S(H,X) = F$ this means there is a specific assignment $y_1, y_2, \dots, y_m$ for which $\forall h \in H \; \exists i \; h(x_i) \neq y_i$

אם H לא מנפצת את X, אזי קיימת השמה y1,…,ym כך שלכל היפותזה h ממרחב ההיפותזות, קיים i בין 1 ל-m מסוים עבורו לא מתקיים: h(xi)=yi.

דוגמה:

Let U be some universe and let $X = \{x_1, x_2\}$. how many possible assignments Y does X have?

Y_1	Y_2	Y_3	Y_4
$X_1 = -1$	$X_1 = 1$	$X_1 = -1$	$X_1 = 1$
$X_2 = -1$	$X_2 = -1$	$X_2 = 1$	$X_2 = 1$

Let H by some hypothesis space.
- Can $S(H,X) = True$ if $|\boldsymbol{H}| < \boldsymbol{4}$? No.
- Can $S(H,X) = True$ if $\boldsymbol{h}(\boldsymbol{x_2}) = -\boldsymbol{1} \; \forall \boldsymbol{h} \in \boldsymbol{H}$? No.
- Can $S(H,X) = True$ if $\boldsymbol{h}(\boldsymbol{x_1}) = \boldsymbol{h}(\boldsymbol{x_2}) \forall \boldsymbol{h} \in \boldsymbol{H}$? No.

- ברור שאם הגודל של מרחב ההיפותזות קטן ממספר ההשמות, אזי מרחב ההיפותזות לא יכול לנפץ את קבוצת הנקודות (משום שתהיה השמה שלא תהיה מסופקת ע״י המרחב)
- מרחב היפותזות שממפה את כל הנקודות x2 לערך -1 גם אינו מנפץ את X מכיוון שקיימות השמות שלא מסכימות אתו (למשל Y3,Y4).
- מרחב היפותזות שממפה כל x1 באותו אופן שהיא ממפה את x2 גם אינה מנפצת את X מהיות שקיימות השמות שלא מסכימות אתו (למשל Y3,Y2)

VC Dimension – VC(H), VC dimension של מרחב היפותזות H, המוגדר מעל מרחב אינסטנסים U, הוא הגודל של תת הקבוצה הסופית הגדולה ביותר של X שמנופצת ע״י מרחב ההיפותזות H.

- נשים לב כי מספיק למצוא תת קבוצה אחת בעלת גודל נתון ש-H יכולה לנפץ!
- אם קבוצות שרירותיות גדולות סופיות של U יכולות להתנפץ על ידי H, אזי VC(H) = infinity.
- זהו מדד עבור מרחב ההיפותזות H

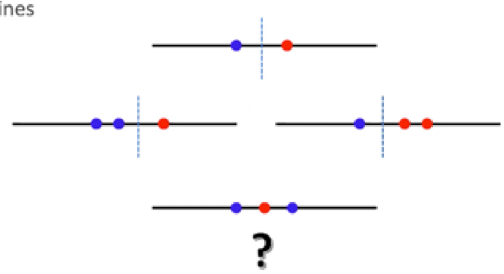
ה-**VC Dimension** של מרחב היפותזות H על מימד דגימות X, מוגדר להיות העוצמה של תת הקבוצה הגדולה ביותר של X ש-H יכולה לנפץ.
מספיק למצוא קבוצה אחת בגודל m ש-H יכולה לנפץ, כדי להוכיח ש- $VC(H) \geq m$.
צריך להראות ש-H לא יכולה לנפץ **אף קבוצה** בגודל $m+1$ על מנת להוכיח ש- $VC(H) < m+1$.

דוגמאות:

גודל מרחב ההיפותזות של מפרידים לינאריים חד ממדיים הוא 2. הראינו **קבוצת נקודות אחת** שכל השמה שלה שניתן לנפץ ולכן גודל H גדול שווה מ-2, ויש להוכיח שלכל קבוצה של 3 נקודות לא קיימת השמה שמנפצת ולכן H קטן ממש מ-3. (בהמשך נראה כיצד מוכיחים פורמלית) ולכן VC(1-dimensional linear separators) = 2.

Shattering – example I

- 1-dimension space
- H – linear lines



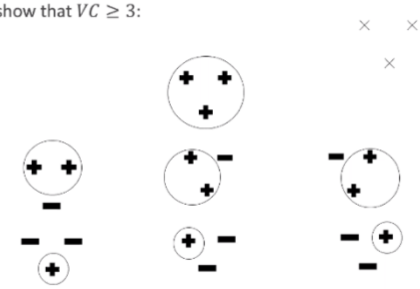
נראה הוכחה פורמלית:

יהי U, מרחב הדגימות של כל הנקודות במישור דו-ממדי כלומר (x,y) ששייך ל-R^2. מצאו את ה-VC dimension שבו מרחב ההיפותזות הוא כל המעגלים (החלק הפנימי של כל מעגל מסווג כחיובי), והוכיחו אותו.

הוכחה: נוכיח כי VC(H) = 3.

- **הצעד הראשון הוא להראות כי VC גדול שווה 3.** לכן נראה **קבוצה אחת** של 3 נקודות (קבוצה של משולש הפוך שווה שוקיים), ונראה את כל ההשמות שמנפצות אותה (אכן מקיפות את הפלוסים במעגל).

First, we'll show that $VC \geq 3$:



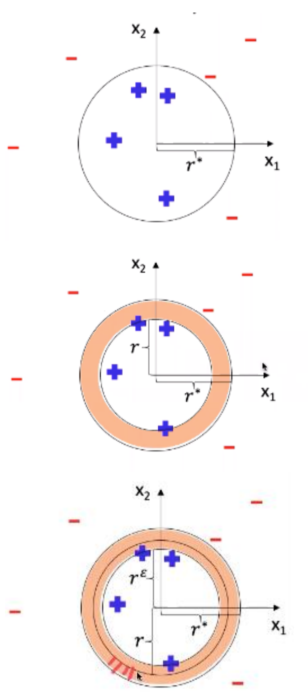
- **הצעד הבא הוא להוכיח כי VC < 4** להלן ההוכחה הפורמלית :

Second, we'll show that $VC < 4$:

We show this by constructing a counterexample in several cases

- If the four points are collinear, the labeling +-+- (going along the line) is impossible, among numerous others
- If the convex hull of the four points is a triangle, then the labeling with + (the three points of the triangle) and - (the interior point) is not possible
- If the convex hull of the four points is a quadrilateral, then let (a1, a2) be the points separated by the long diagonal and (b1, b2) be the points separated by the short diagonal. At least one of the labelings +(a1, a2), −(b1, b2) or +(b1, b2), −(a1, a2) must be impossible:
  - If they were both possible, then there would be some satisfying circle c1 for the first labeling and some other circle c2 satisfying the second labeling, and the symmetric difference of these circles ((c1 \ c2) ∪ (c2 \ c1)) would consist of four disjoint regions, which is impossible for circles

Since some set of 3 points is shattered by the class of circles, and no set of 4 points is, the VC dimension of the class of circles is 3



חישוב ישיר של סיבוכיות הדגימה (כמו דוגמת מרחב המלבנים)

דוגמה : לפנינו משחק ללמוד מעגל קונקרטי לא ידוע במישור האוקלידי עם 2 ממדים. יהי *r הרדיוס של מעגל המטרה. כל דגימה בדאטה אימון הוגרלה מתוך התפלגות לא ידועה D ומכילה 2 פיצ'רים (מיקום הדגימה על הצירים (x1,x2)) וערך מטרה (+1 אם היא בתוך המעגל ו- -1 אחרת). מרחב הקונספטים שלנו הוא עיגולים שמרכזם הוא ראשית הצירים. נבנה אלגוריתם שמוצא את המעגל הקטן ביותר שמכיל את הפלוסים – עבור כל הפלוסים שלנו נמצא את המיקום של הקיצוניים ביותר ועל פיהם נגדיר רדיוס היפותזה r.

- *r - הרדיוס של מעגל הקונספט
- r – היפותזה שמנחשת את המעגל
- r-epsilon – הטבעת (טבעת הטעות, הכתומה) הגדולה ביותר שנוצרת עם הסתברות (ליפול בתוכה) שהיא במקסימום אפסילון.

$$r^{\varepsilon} = \operatorname{arginf}_r \Pr[(x_1, x_2) \in A_r] \leq \varepsilon$$

- <u>מקרה 1</u> : אם רדיוס אפסילון קטן שווה מ-r אזי ההסתברות ליפול בטבעת קטנה מאפסילון (איור שלישי) – הטבעת מוכלת בתוך טבעת האפסילון.
- <u>מקרה 2</u> : נניח כי רדיוס אפסילון גדול מ- r

Case 2: Otherwise, what is the probability of missing the annulus of radii r^{ε}, r^{*} with m traning examples?

$$(1-\varepsilon)^m \leq \exp(-\varepsilon m)$$

With sample size m $\geq \frac{\ln\left(\frac{1}{\delta}\right)}{\epsilon}$, we get

$$\exp(-\varepsilon m) \leq \exp\left(-\ln\left(\frac{1}{\delta}\right)\right) = \exp(\ln(\delta)) = \delta$$

So if the probability of the annulus is very small, the error it incurs is also small

With enough examples, it is very unlikely to miss the annulus