למידה חישובית ממידע - שיעור 10

תאוריה של למידה חישובית

נדבר על איך אנו משערכים טעות של תוצאה של למידה, ובאופן כללי איך אנו מגדירים טעות הכללה באלגוריתם למידה ובלמידה באופן כללי.

בנוסף נדבר על סיבוכיות של למידה. סיבוכיות בלמידה חישובית אינה מדברת רק על זמני ריצה (אבל זמן הריצה אכן מעניין אותנו), אלא על כמות ואיכות הdata. בשביל ללמוד טוב אנו זקוקים להרבה data. השאלה של כמה data אני צריכה כדי ללמוד היטב נקראת sample complexity.

PAC - Probably Approximately Correct

אנו לומדים בצורה שבהסתברות גבוהה, יהיה בערך נכון. זה נשמע קצת כמו בדיחה, אך אנו באמת מבצעים ככה למידה. למידה חישובית נותנת לנו אלגוריתמים שבהסתברות גבוהה יתנו לנו חלוקה כלשהי של המידע שלי כך שבהסתברות גבוהה אוכל לסווג אותו.

<u>אנו מסתמכים על 2 מרכיבים:</u>

Probability- אחוז מסוים של certainty. תמיד תהיה טעות מסוימת לכן נעדיף להשתמש **בהסתברות לטעות.**

Approximation- מתייחס לerror bound. לרוב לא נדע את הטעות המדויקת על כל מרחב המופעים לכן נעדיף להשתמש **בהערכה של הטעות.** הApproximation כמה טוב ההיפותזה שלנו מתאימה לtraining data.

נוכל גם להעריך את הLearnability, כלומר האם בכלל ניתן לבצע את הלמידה.

ניזכר שהגדרנו בעבר את "האלגוריתם המבצע" להיות ההיפותזה של הdata שלנו:

$$L(D) = \underbrace{h}_{h \text{ ההיפותזה}}$$

נגדיר את $P(X)$ להיות קבוצת כל הטעויות שלנו.

$h \in H \subset P(X)$ - היפותזה, הפונקציה המסווגת. בעצם מה שאנו חוזים.

$c \in C \subset P(X)$ - כאשר C הוא ה concept האמיתי. מה שקורה בפועל.

נגדיר את המצב שבו תמיד יש היפותזה במרחב ההיפותזות שמתאר את הקונספט: $C \subseteq H$. כלומר במצב זה לכל c יש h שמייצג אותו. במצב זה אנחנו בטעות 0.

אנו לא יודעים מה הערך של c, אבל תמיד נרצה שc יהיה שווה ממש לh.

נגדיר:

- $C-$ קונספט האמיתי. פונקציה או קלסיפיקציה שאנו רוצים ללמוד.
- $X-$ המרחב שממנו מגיעים ה-instances.
- $x \in X-$ ה-training set שבאמצעותו נלמד את הקונספט.

Learning from Data

נתאים היפותזה (מ-H) לdata שאנו רואים (הtraining set)

לרוב, נזדקק ליותר data מאשר פרמטרים חופשיים במודל ההיפותזה.

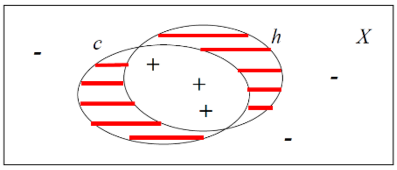
Approximation- מודד כמה טוב ההיפותזה מתאימה לtraining data. נקרא in- sample-error

Generalization- מודד כמה טוב ההיפותזה עתידה להתאים ל data חדש. נקרא out-of-sample-error.

בלמידה חישובית אנו מתעניינים בgeneralization ולכן למידה היא קשה. אנו רוצים להשיג generalization מה- sample data.

The True Error of h:

יש לי התפלגות כלשהי על X. נגדיר את ה-true error להיות ההסתברות שהערך קונספט C יתן על איבר שהגרלתי מ-x יהיה שונה מהערך שההיפותזה h תיתן עליו:



$$TrueErr(h) = error_D(h) = Prob_{X\sim D}\big(c(X) \neq h(X)\big)$$

<u>הטעות האמיתית</u> היא ההסתברות של הנקודה שהגרלתי ליפול בשטח האדום.

נראה כעת דוגמה למרחב שאינו אוקלידי:

דוגמה ללמידה של פונקציה בוליאנית:

Example	x_1	x_2	x_3	x_4	y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

נרצה ללמוד פונקציה בוליאנית (קלסיפיקציה בינארית) עבור 4 משתנים בוליאנים:

$$f(x_1, x_2, x_3, x_4) = t \in \{0,1\}$$

נשים ♥ ש: $X = \{0,1\}^4$ ולכן $|X| = 16$.

אם כל הפונקציות הבוליאניות במרחב ההיפותזה H, אז $|H| = 2^{16}$ (כי יש 16 פונקציות שונות ולכל אחת מהן 2 אפשרויות).

ה-training data שלי יהיה 7 הדוגמאות שבאיור. כלומר יש לי $2^{16-7} = 2^9$ אפשרויות שונות להשלים את הטבלה הנתונה. לכאורה יש לנו 2^9 אפשרויות שיכולות להיות הפלט של הלמידה. כלומר בהינתן מצב בו אין לי הגבלה כלשהי על H אני לא באמת יכולה להכליל על ה-data הנתון.

נלמד בתרגול את תאוריית No Free Lunch שמרחיבה על הנושא.

נרצה כעת לייצר הגבלות על ה-data שלי

Rule	Counterexample
$\Rightarrow y$	1
$x_1 \Rightarrow y$	3
$x_2 \Rightarrow y$	2
$x_3 \Rightarrow y$	1
$x_4 \Rightarrow y$	7
$x_1 \wedge x_2 \Rightarrow y$	3
$x_1 \wedge x_3 \Rightarrow y$	3
$x_1 \wedge x_4 \Rightarrow y$	3
$x_2 \wedge x_3 \Rightarrow y$	3
$x_2 \wedge x_4 \Rightarrow y$	3
$x_3 \wedge x_4 \Rightarrow y$	4
$x_1 \wedge x_2 \wedge x_3 \Rightarrow y$	3
$x_1 \wedge x_2 \wedge x_4 \Rightarrow y$	3
$x_1 \wedge x_3 \wedge x_4 \Rightarrow y$	3
$x_2 \wedge x_3 \wedge x_4 \Rightarrow y$	3
$x_1 \wedge x_2 \wedge x_3 \wedge x_4 \Rightarrow y$	3

ההגבלה היא כל הפונקציות שניתנות לכתיבה כקומבינציה של $\wedge$ על 4 הפרמטרים שלי.

יש 2^4 פונקציות כאלה (כי זאת שאלה בינארית האם לשים או לא את הסימן $\wedge$).

Example	x_1	x_2	x_3	x_4	y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

כדי לתת מענה להגבלה שקיבלנו נחפש דוגמאות נגדיות בתוך הtraining data, כלומר ננסה לחפש עבור כל פונקציה אפשרית האם היא מקיימת את הגבלות של ה- training data.

נביט לדוגמה בשורה השנייה של הטבלה השמאלית. בשורה זו הוגדר ש: $x_1 = y$, אבל ניתן לראות שקיימת דוגמה נגדית לטענה זו בשורה השלישית של הטבלה הימנית שמראה ש $x_1 \neq y$.

הגבלת הHypothesis space

בדוגמאות שהראנו הגבלנו את מרחב ההיפותזות. כמעט תמיד נוכל לבחור טיפוס פונקציה מסוים ובכך להגביל את הhypothesis space.

<u>יתרונות:</u>

סביר יותר שנצליח ללמוד. בנוסף אם יש לי סיבה להניח שההגבלה הזו מייצגת מציאות, למרות שעלולות להיות טעויות - אנו מוכנים לשלם אותן, ועדיין ללמוד מתוך ההגבלה הזו.

<u>חסרון:</u>

לא בהכרח נוכל למצוא היפותזה שהטעות שלה תהיה 0.

נרצה לשערך את הטעות שלנו באמצעות בדיקת ההיפותזה שלנו על ה-test. נשתמש ב-test set בגודל $|S|$ ונניח שמספר הטעויות הינו r, אזי נוכל להראות כי $\frac{r}{|S|}$ הוא שערוך לטעות generalization.

בנוסף, אנו יודעים ממשפט הגבול המרכזי כי אם יש לנו n גדול דיו של דגימות בלתי תלויות מהאוכלוסייה, אני יכולה לומר ב-95% וודאות כי ה-true error תהיה קטנה מ:

$$\frac{r}{|S|} + \epsilon$$

אם נרצה לדעת מה הפרופורציה של תכונה כלשהי באוכלוסייה, וה-data שלי גדול דיו, נגדיר:

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

- p הוא הפרופורציה של התכונה ב-sample.
- $p(1-p)$ היא השונות של כל instance.
- ה-se הולך וקטן עם n.

רווח סֶמֶך (confidence interval)- אינטרוול המחושב מתוך תוצאות מדגם. עונה על השאלה כמה אנו יכולים להיות בטוחים שפרמטר כלשהו נמצא בתוך אינטרוול נתון.

$$CI\ :\ p +/- 2(se)$$

למידה חישובית 19.05.2019
שיעור 10

נחזור לדיון של Pac Learning.

sample complexity

מספר הדגימות להן אנו זקוקים ב-training על מנת לאפשר לאלגוריתם להחזיר היפותזה שתיתן טעות קטנה על data לא מוכר.

אחת הסיבות ל-overfitting היא data קטן ולכן נרצה לדעת לענות על השאלה: כמה דוגמאות ב-data מספיקות בשביל ללמוד.

השאלה הזו תלויה במספר דברים:

1. אופי מרחב הקונספטים C. מי הקונספטים שאני מנסה ללמוד.
2. מרחב ההיפותזות H.
3. מרחב ההסתברות (X, D)

הגדרה: נאמר שמרחב היפותזת H הוא **קונסיסטנטי** ביחס ל-C אם $C \subseteq H$

הגדרה: היפותזה H היא ***D*-קונסיסטנטית** ביחס לקונספט C ולכל training data D, אם:

$$\forall d \in D, h(d) = c(d)$$

הגדרה: נאמר שאלגוריתם L הוא **לומד קונסיסטנטי** אם מתקיים שמרחב ההיפותזות עמו הוא עובד הוא קונסיסטנטי, ושנית לכל $c \in C$ ולכל training data D שמיוצר על ידי C מתקיים:

$$L(D) \;\; is \; D - \; consistant \; with \; C$$

כלומר האלגוריתם שנלמד על ה-data (שהוא בעצם ההיפותזה שהתקבלה) הוא D- קונסיסטנטי

דוגמה:

מרחב ההיפותזות יכיל n ליטרלי $disjunction$ (ליטרל x או $\bar{x}$ מחובר בסימן V לליטרל הבא).
גודל מרחב ההיפותזות אם כן הוא: $|H| = 3^n$ (כי אני יכולה לשים ליטרל, את שלילתו או את שניהם).

1. Start: $x_1 \vee \overline{x_1} \vee x_2 \vee \overline{x_2} \vee x_3 \vee \overline{x_3} \vee x_4 \vee \overline{x_4}$
2. Instance 1: $x_1 \vee x_2 \vee \overline{x_3} \vee x_4$
3. Instance 2: $x_1 \vee x_4$

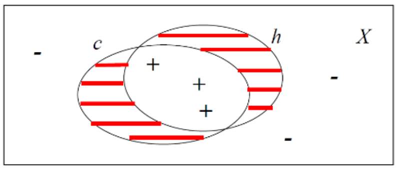
Example	x_1	x_2	x_3	x_4	y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1

ניתן אלגוריתם שלוקח את ה-training data הזה וימצא היפותזה קונסיסטנטית איתו.
נרצה להוכיח שהאלגוריתם מביא אותי לנוסחה שהיא קונסיסטנטית עם ה-training data.

נתחיל מהנוסחה המלאה שב-start ונעבור על כל הדוגמאות השליליות ב-data. לדוגמה, בשורה הראשונה של הטבלה (instance 1) $y = 0$, כלומר הוא שלילי, לכן נוכל לשלול את $\overline{x_1}, \overline{x_2}\, x_3, \overline{x_4}$ נחזור על התהליך. בסיום הנוסחה שתתקבל תהיה קונסיסטנטית עם הדוגמאות החיוביות.

<u>הגדרה</u>: היפותזה h תקרא $\epsilon - bad$ אם $error_D(h) > \epsilon$

שאלה: <u>מה ההסתברות שאלגוריתם קונסיסטנטי יחזיר היפותזה h שהיא $\epsilon - bad$?</u>



ידוע שהאלגוריתם קונסיסטנטי, לכן כל נקודות ה-data הנוספות בהכרח יהיו מחוץ לשטח האדום (אחרת לא קונסיסטנטי).

אם ההיפותזה h היא $\epsilon\, bad$, הסיכוי לסווג את אחד מהנקודות של ה-data בשטח האדום הינה ϵ (הגדרה).

לכן נסיק מעיקרון המשלים כי ההסתברות לסווג נכון את ה-data קטן מ - $1 - \epsilon$.

כלומר ההסתברות של כל ה-data points (m נקודות) להיות מחוץ לשטח האדום גדולה שווה ל:

$$(1-\epsilon)^m$$

אם h היא $\epsilon - bad$ אזי ההסתברות לפלוט h קונסיסטנטית (ללא טעויות על כל הנקודות) היא:

$$P(L(D) = h) \le (1-\epsilon)^m$$

נמצא חסם על ההסתברות למצוא היפותזה קונסיסטנטית המקיימת $\epsilon - bad$:

$$\Pr(\exists h\ s.t.\ \epsilon\ bad\ and\ consistent) = \sum_{h \in \epsilon\ bad} \mathrm{P}(h\ is\ consistent\ with\ D_m)$$
$$\le |\{h\ is\ \epsilon\ bad\}|(1-\epsilon)^m \le |H|(1-\epsilon)^m \le |H|e^{-\epsilon m}$$

כאשר:

- ניתן לחסום את $(1-\epsilon)$ ב- $e^{-\epsilon}$, כלומר $(1-\epsilon) < e^{\epsilon}$.

<u>משפט</u>

במרחב היפותזות H, עם קבוצות אימון D בעלי דגימות בת"ל של קונספט c, עבור כל $0 < \epsilon < 1$, ההסתברות שהמרחב H יכיל היפותזה עם שגיאה גדולה יותר מ- ϵ היא קטנה מ:

$$|H|e^{-\epsilon m}$$

אם נרצה שההסתברות להיפותזה h שהיא $\epsilon\, bad$ תהיה קטנה מ - δ, כלומר - $|H|e^{-\epsilon m} < \delta$, נצטרך כמות מידע:

$$m \ge \frac{1}{\epsilon}\ln\frac{|H|}{\delta} = \frac{1}{\epsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right)$$

זה ה-sample complexity שלנו.

ה-δ היא אחוז שאני קובעת, והיא מהווה חסם עליון על הטעות.