מבוא ללמידה חישובית (0368-3235)

נכתב ע"י רון גולדמן
על פי הרצאות של פרופ' נדב כהן

26 באוקטובר 2025

תוכן העניינים

חלק I

מבוא

פרק 1

מה זה למידה?

- שינוי התנהגות עם הזמן במטרה להשיג איזשהו יעד.
- דוגמאות:
  - ללמוד ללכת.
  - ללמוד לנהוג.
  - ללמוד שפה.
  - ללמוד מתמטיקה.

מה זה למידה חישובית?

- בניית מכונות שיכולות ללמוד.
- בניית תאורייה מתמטית ללמידה.
- בעיקרון, מכונות יכולות ללמוד יותר טוב מבני אדם!

למה למידה חישובית?

- האם למידה היא הדרך היחידה לבנות "מכונות חכמות"?
- כנראה שלא.
- למה לא לבנות מכונות שיבצעו פעולות.

דוגמה 1.1 [מסנן ספאם]. בגישה של המומחה (לא מלמידת חישובית): נרשום חוקים שקובעים מהו ספאם ומה לא.
בעיות עם הגישה של המומחה:

- דורשת בן אדם שיעבוד קשה.
- דורשת עדכון בכל רגע.
- בני אדם לא בהכרח ימצאו את הפתרון האופטימלי.

סינון באמצעות למידת מכונה:

- **נלמד** מכונה לסנן ספאם מתוך **דוגמאות**.
- אלגוריתם למידה ישתמש בדוגמאות הללו כדי ללמוד איך נראה ספאם.

מבנה הקורס

- למידה מפוקחת - תאוריה
- למידה מפוקחת - אלגוריתמים
- למידה לא מפוקחת

ייצוג פרדיקטור

הגדרה 1.2 פרדיקטור זו פונקציה $h : \mathcal{X} \to \mathcal{Y}$ כאשר $\mathcal{X}$ **הקלטים** ו-$\mathcal{Y}$ **התוויות**.

- בדרך כלל $\mathcal{X} = \mathbb{R}^d$.
- בבעיית קלאסיפיקציה $\mathcal{Y} = \{0, 1\}$.

למידה מפוקחת

- מטרה: **ללמוד פרדיקטורים מדוגמאות.**
- נושאים שנכסה:
  - **מודלים: מסווגים לינאריים, שיטות אנסמבל, עצי החלטה, למידה עמוקה.**
  - **אלגוריתמים: SVM, SGD, boosting.**
  - **תאוריה: למידת PAC, מימד VC.**

למידה לא מפוקחת

- תוויות קשה לאסוף.
- מידע לא מתויג קל לאסוף.
- מה אפשר לעשות עם מידע לא מתויג.
- אתגרים:
  - האם אפשר למצוא או לתאר מבנה במידע.
- נושאים שנכסה:
  - ניתוח גורמים עיקריים (PCA).
  - קיבוץ (clustering).
  - מודלים יוצרים (Generative models).
  - אלגוריתם מקסום תוחלת (EM).

פרק 2

חזרה על הסתברות וצעדים ראשונים בהכללה

הערה 2.1. אני הולך לספק הגדרות יותר מדויקות ממה שנדב נתן לשם שלמות, אין מה להיבהל.

2.1 הגדרות בסיסיות בהסתברות

הגדרה 2.2 [מרחב הסתברות]. **מרחב הסתברות** הוא שלשה $(\Omega, \mathcal{A}, \mathbb{P})$ כאשר:

1. Ω היא **קבוצת התוצאות**

2. $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ הוא **מרחב המאורעות**, $\mathcal{A}$ הוא σ-אלגברה מעל Ω, כלומר:

   (א) לכל $A \in \mathcal{A}$ גם $A^c \in \mathcal{A}$.

   (ב) לכל $A_1, A_2, \ldots \in \mathcal{A}$ מתקיים כי $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$.

3. $\mathbb{P} : \mathcal{A} \to [0,1]$ היא **פונקציית ההסתברות**, המקיימת:

   (א) $\mathbb{P}(\Omega) = 1$.

   (ב) לכל $A_1, A_2, \ldots \in \mathcal{A}$ זרים בזוגות מתקיים

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

הגדרה 2.3 [משתנה מקרי]. **משתנה מקרי** הוא פונקציה $X : \Omega \to \mathbb{R}$ כך שלכל $a < b \in \mathbb{R}$ מתקיים $\{\omega \in \Omega :\ a < X(\omega) < b\} \in \mathcal{A}$.

סימון 2.4. יהי X משתנה מקרי, אז לכל $a \in \mathbb{R}$ נסמן

$$\{X = a\} \triangleq X^{-1}(a) = \{\omega \in \Omega :\ X(\omega) = a\}$$
$$\{X \leq a\} \triangleq X^{-1}((-\infty, a]) = \{\omega \in \Omega :\ X(\omega) \leq a\}$$

נרשום

$$\mathbb{P}(\{X = a\}) = \mathbb{P}(X = a)$$
$$\mathbb{P}(\{X \leq a\}) = \mathbb{P}(X \leq a)$$

2.2 הקשר בין למידה חישובית והסתברות

בלמידה מפוקחת נתון לנו משתנה אחד (למשל תמונה) ואנו מעוניינים במשתנה אחר (תגית). נמדל את הראשון כמשתנה מקרי X ואת השני כמשתנה מקרי Y.

מטרה: לקבוע את Y בהינתן X.

נניח התפלגות משותפת $\mathbb{P}(X=x,Y=y)$.

הגדרה 2.5 [התפלגות שולית]. ההתפלגות השולית של X מתוך (X,Y):

$$\mathbb{P}(X=x)=\sum_{y\in Y}\mathbb{P}(X=x,Y=y)$$

הגדרה 2.6 [התפלגות מותנה]. ההתפלגות המותנה של Y בהינתן X:

$$\mathbb{P}(Y=y|X=x)=\frac{\mathbb{P}(X=x,Y=y)}{\mathbb{P}(X=x)}$$

הנחה 2.7 [רק להיום! לא מציאותית]. אנחנו יודעים את ההתפלגות המשותפת $\mathbb{P}(X=x,Y=y)$.

נרצה למצוא את הפרדיקטור האופטימלי $h:\mathcal{X}\to\mathcal{Y}$.

הערה 2.8. $\hat{Y}=h(X)$ הוא משתנה מקרי.

הגדרה 2.9 [פונקצית הפסד]. פונקצית הפסד היא פונקציה $\ell:\mathcal{Y}\times\mathcal{Y}\to\mathbb{R}_{\geq 0}$ כך ש-$\ell(y,\hat{y})=$ המחיר לנחש $\hat{y}$ כאשר התגית האמיתית היא y.

דוגמה 2.10 [פונקציות הפסד נפוצות].

- פונקצית הפסד $0-1$:

$$\ell(y,\hat{y})=\mathbb{1}\{y=\hat{y}\}$$

- פונקצית הפסד ריבועית (ℓ_2):

$$\ell(y,\hat{y})=\|y-\hat{y}\|_2^2$$

בהינתן פונקצית הפסד מסוימת, המטרה הטבעית היא למזער את תוחלת ההפסד:

$$L(h)\triangleq\mathbb{E}[\ell(Y,h(X))]=\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}}\mathbb{P}(X=x,Y=y)\ell(y,h(x)) \tag{2.1}$$

הגדרה 2.11 [פרדיקטור אופטימלי]. הפרדיקטור האופטימלי הוא $h=\arg\min_h L(h)$.

דוגמה 2.12 [קלאסיפיקציה בינארית]. נתבונן במקרה הפשוט של קלאסיפיקציה בינארית, $\mathcal{Y}=\{0,1\}$ עם הפסד $0-1$.

בהינתן x, מה הפרדיקציה האופטימלית $h(x)$?

נחשב על הגורמים ב-$L(h)$ (2.1) שתלויים ב-x:

$$\begin{aligned}L(h)=&\mathbb{P}(X=x,Y=1)\ell(1,h(x))+\mathbb{P}(X=x,Y=0)\ell(0,h(x))\\=&\mathbb{P}(X=x)(\mathbb{P}(Y=1|X=x)\ell(1,h(x))+\mathbb{P}(Y=0|X=x)\ell(0,h(x)))\end{aligned} \tag{2.2}$$

אם $h(x)=1$ אז (2.2) = $\mathbb{P}(X=x)\mathbb{P}(Y=1|X=x)$, אם $h(x)=0$ אז (2.2) = $\mathbb{P}(X=x)\mathbb{P}(Y=0|X=x)$.

אז, האם אנחנו צריכים לבחור ש-$h(x)=1$ או $h(x)=0$? נבחר לפי הערך הקטן יותר של (2.2).
כלומר:

$$\begin{aligned}h(x)=&\mathbb{1}\{\mathbb{P}(X=x)\mathbb{P}(Y=1|X=x)\geq\mathbb{P}(X=x)\mathbb{P}(Y=0|X=x)\}\\=&\mathbb{1}\{\mathbb{P}(Y=1|X=x)\geq\mathbb{P}(Y=0|X=x)\}\end{aligned}$$

במילים אחרות,

$$h(x)=\arg\max_{y\in\{0,1\}}\mathbb{P}(Y=y|X=x)$$

זה נקרא מסווג א-פוסטריורי מקסימלי (MAP).

דוגמה 2.13 [סינון ספאם]. יהי X מספר האותיות הגדולות באימייל.
נניח ואנו יודעים את $\mathbb{P}(X=x,Y=y)$, וש-$\mathbb{P}(Y=1)=\mathbb{P}(Y=0)=\frac{1}{2}$.
נניח $\mathbb{P}(X=x|Y=y)$ מקורבת היטב על ידי גאוסיאן עם תוחלת c_y ושונות v, כאשר $c_1>c_0$.
נרצה לחזות את $h(x)=1$ אם ורק אם $\mathbb{P}(Y=1|X=x)\geq\mathbb{P}(Y=0|X=x)$:

$$\begin{aligned}&\iff\frac{\mathbb{P}(Y=1,X=x)}{\mathbb{P}(Y=1)}\geq\frac{\mathbb{P}(Y=0,X=x)}{\mathbb{P}(Y=0)}\\&\underset{\text{Baye's}}{\iff}\mathbb{P}(X=x|Y=1)\geq\mathbb{P}(X=x|Y=0)\\&\iff x\geq\frac{c_0+c_1}{2}\end{aligned}$$

פונקציות הפסד שונות נותנות פרדיקטור אופטימלי שונה. נניח $\ell(y,\hat{y})=(y-\hat{y})^2$, אז:

$$\mathbb{E}[\ell(Y,\hat{y})|X=x]=\mathbb{E}\Big[(Y-\hat{y})^2|X=x\Big]=\mathbb{E}\big[Y^2|X=x\big]-2\hat{y}\mathbb{E}[Y|X=x]+\hat{y}^2$$

גודל זה ממוזער עבור $\hat{y}=\mathbb{E}[Y|X=x]=\mathbb{P}(Y=1|X=x)$.

שאלה 2.14. מה קורה אם אנחנו לא יודעים את ההתפלגות המשותפת? זה מה שקורה במציאות.

כעת נדבר בקצרה על למידת תכונות של התפלגות מתוך נתונים.

דוגמה 2.15. נניח ונרצה למצוא את התוחלת של הטלת מטבע.
יש לנו משתנה ברנולי עם פרמטר $p\in[0,1]$:

$$\mathbb{P}(X=1)=1-\mathbb{P}(X=0)=p$$

נרצה למצוא את p, על סמך מ"מ $X_1,\ldots,X_m\overset{\text{iid}}{\sim}\text{Bernoulli}(p)$.
נתבונן בממוצע:

$$\overline{X}_m\triangleq\frac{1}{m}\sum_{i=1}^{m}X_i$$

מתקיימים:

$$\mathbb{E}\left[\overline{X}_m\right] = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[X_i] = p$$
$$\mathrm{Var}\left(\overline{X}_m\right) = \mathrm{Var}\left(\frac{1}{m}\sum_{i=1}^{m}X_i\right) = \frac{1}{m^2}\sum_{i=1}^{m}\mathrm{Var}(X_i) = \frac{mp(1-p)}{m^2} = \frac{p(1-p)}{m} \xrightarrow{m\to\infty} 0$$

אז, $\overline{X}_m$ "מתכנס" לערך הנכון (p) כאשר $m \to \infty$.

מה קורה עבור m סופי?

משפט 2.16 [אי-שוויון הופדינג].

$$\forall \varepsilon > 0. \qquad \mathbb{P}\left(\left|\overline{X}_m - p\right| \geq \varepsilon\right) \leq 2\exp\left(-2m\varepsilon^2\right)$$

כעת, לכל $\varepsilon > 0$ ו-$\delta \in [0,1]$, כך ש-$\overline{X}_m$ במרחק לכל היותר ε מ-p ("approximately correct") בהסתברות $\leq 1-\delta$ ("probably").

לכן, כל מה שאנחנו צריכים, זה ש-$2\exp\left(-2m\varepsilon^2\right) \leq \delta$, זה קורה אם ורק אם:

$$m \geq \frac{1}{2\varepsilon^2}\log\frac{2}{\delta}$$

אנחנו מצאנו m שמבטיח משערך probably approximately correct (PAC).