22.12.2022 ***פרטיות מידע***

Lecture 9: Private Empirical Risk Minimization

Textbook: Cynthia Dwork and Aaron Roth. The Algorithmic Foundations of Differential Privacy

מרצה: אורי שטמר

היום נתחיל לדבר על בעיות אופטימיזציה. זוהי משפחה חשובה של בעיות שמופיעות בהרבה מקומות בספרות של *ML*. בצורה לא פורמלית, המשימה שלנו תהייה כזאת:

בהינתן דטהבייס, נגדיר פונקציית *loss* שנרצה להביא למינימום

כדי להבין יותר טוב מה זה אומר, נסתכל על הדוגמאות הבאות:

<u>בעיה 1 – בעיית הממוצע:</u> בהינתן דטהבייס $S = (x_1, \dots, x_n) \in \mathbb{R}$, חשב את הממוצע של S.

כלומר מה שאנחנו רוצים לחשב כאן זה את $\frac{1}{n}\sum_{i=1}^{n} x_i$.

<u>טענה:</u> מסתבר שאפשר להציג את הממוצע גם בצורה הבאה:

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \operatorname*{argmin}_{w\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}(w - x_i)^2$$

<u>הוכחה:</u>
נגזור את הביטוי $\frac{1}{n}\sum_{i=1}^{n}(x_i - w)^2$ לפי w ונקבל

$$\frac{\mathrm{d}}{\mathrm{d}w}\left(\frac{1}{n}\sum_{i=1}^{n}(w - x_i)^2\right) = \frac{1}{n}\sum_{i=1}^{n} 2(w - x_i) = 2w - \left(\frac{2}{n}\sum_{i=1}^{n} x_i\right) = 2\big(\, w - \mathrm{AVG}(S)\,\big)$$

כלומר, הנגזרת שלילית עבור $w < \mathrm{AVG}(S)$, הנגזרת חיובית עבור $w > \mathrm{AVG}(S)$, והנגזרת שווה לאפס עבור $w = \mathrm{AVG}(S)$. כלומר הפונקציה שלנו יורדת עד לנקודה $w = \mathrm{AVG}(S)$ ואחריה הפונקציה עולה. לכן הנקודה שמביאה למינימום את הפונקציה הזאת היא בדיוק $w = \mathrm{AVG}(S)$ ולכן

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \operatorname*{argmin}_{w\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}(w - x_i)^2$$

המסקנה כאן היא שאת "בעיית הממוצע" אפשר להציג כך:
בהינתן דטהבייס $S = (x_1, \dots, x_n) \in \mathbb{R}$, מצא נקודה $w \in \mathbb{R}$ אשר מביאה למינימום את הפונקציה

$$L(w) = \frac{1}{n}\sum_{i=1}^{n}(w - x_i)^2$$

(אם כל מה שאני רוצה לעשות זה לחשב ממוצע אז אין באמת סיבה שאסתכל על הפונקציה $L(\cdot)$ הזאת. אבל המסר כאן הוא שבעיית הממוצע היא מקרה פרטי של הבעייה היותר כללית שנראה בהמשך)

<u>בעיה 2 – בעיית החציון:</u> בהינתן דטהבייס $S = (x_1, \dots, x_n) \in \mathbb{R}$, חשב חציון של S.

גם את הבעיה הזאת אפשר להציג כבעיית אופטימיזציה, באופן הבא:

בהינתן דטהבייס $S = (x_1, \dots, x_n) \in \mathbb{R}$, מצא נקודה $w \in \mathbb{R}$ אשר מביאה למינימום את הפונקציה

$$L(w) = \frac{1}{n}\sum_{i=1}^{n}|w - x_i|$$

כדי לראות שמתקיים

$$\text{median}(S) = \underset{w\in\mathbb{R}}{\text{argmin}}\frac{1}{n}\sum_{i=1}^{n}|w - x_i|$$

היזכרו כי מתקיים $\frac{\mathrm{d}}{\mathrm{d}x}|x| = \text{sign}(x)$

% למה? אם $x > 0$ אזי $|x| = x$ ואז הנגזרת היא 1 ואם $x < 0$ אז $|x| = -x$ ואז הנגזרת היא -1

לכן,

$$\frac{\mathrm{d}}{\mathrm{d}w}\left(\frac{1}{n}\sum_{i=1}^{n}|w - x_i|\right) = \frac{1}{n}\sum_{i=1}^{n}\text{sign}(w - x_i)$$

הנגזרת הזאת שווה לאפס בדיוק כאשר מספר המחוברים החיוביים שווה למספר המחוברים השליליים, מה שקורה כאשר $w = \text{median}(S)$.

<u>בעיה 3 – רגרסיה לנארית:</u>

<u>קלט:</u> דטהבייס $S = \big((x_1, y_1), \dots, (x_n, y_n)\big)$ כאשר כל $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$

בבעייה הזאת אנחנו מניחים שישנה פונקציה לינארית מהצורה $f(x) = w_1 \cdot x_1 + \cdots + w_d \cdot x_d = \langle w, x\rangle$ אשר "מסבירה את הדטה" במובן הזה שלכל i מתקיים

$$y_i \approx f(x_i) = \langle w, x_i\rangle$$

תחת ההנחה שלנו, המטרה שלנו היא למצוא פונקציה כזאת (כלומר למצוא w כזה).

בהינתן פתרון אפשרי w, איך נכמת כמה הוא "טוב"? דרך סטנדרטית לעשות את זה היא באופן הבא:

אנחנו מחפשים ווקטור $w \in \mathbb{R}^d$ אשר מביא למינימום את הפונקציה הבאה

$$L(w) = \frac{1}{n}\sum_{i=1}^{n}\big(y_i - \langle w, x_i\rangle\big)^2$$

(דוגמה נוספת – רשתות נוירונים)

<u>נשים לב לכמה תכונות המשותפות לכל הדוגמאות האלה:</u>

1. פונקציית ה *loss* מוגדרת מתוך ה *data* (ברגע שקבענו את ה *data* זה מגדיר איזושהי פונקציית *loss* שאני רוצה להביא למינימום)
2. פונקציית ה *loss* מוגדרת כסכום עם מחובר אחד לכל נקודת קלט
3. מרחב הפתרונות איננו תלוי בדטה (בדוגמאות האלה מרחב הפתרונות היה פשוט $\mathbb{R}$ או $\mathbb{R}^d$). זה יהיה חשוב לנו בהמשך כי זה יפשט לנו את החיים כשננסה לתכנן אלגוריתם פרטי.

אז ראינו כמה דוגמאות לבעיות שניתן להציג אותן כמשימת אופטימיזציה. עכשיו אנחנו רוצים לעשות אבסטרקציה למשימה הזאת בצורה שתתפוס את הדוגמאות האלה (ועוד הרבה דוגמאות אחרות).

<u>הגדרה 1 – פונקציית הפסד פריקה (*decomposable loss function*):</u>

פונקציית הפסד <u>פריקה</u> היא פונקציית הפסד שניתנת להצגה כסכום של מחוברים – מחובר אחד לכל נקודת קלט. באופן יותר פורמלי, תהי $L(w;S)$ פונקציית הפסד הממפה דטהבייס $S=(x_1,\dots,x_n)$ ופתרון w למספר (ככל שהמספר הזה קטן יותר, כך w הוא פתרון טוב יותר ביחס לדטהבייס S).

נאמר ש $L(w;S)$ היא <u>פריקה</u> אם קיימת פונקצייה $\ell(w;x_i)$ כך שניתן לרשום:

$$L(w;S)=\frac{1}{n}\sum_{i=1}^{n}\ell(w;x_i)$$

לפעמים נרשה מחובר נוסף שתלוי רק ב w, שנקרא "רגולייזר" המאפשר לנו להגדיל את ה *loss* של פתרונות מסויימים. למשל, זה יאפשר לנו להעדיף פתרונות w "פשוטים" על פני פתרונות w "מסובכים". ספציפית,

$$L(w;S)=\frac{1}{n}\sum_{i=1}^{n}\underbrace{\ell(w;x_i)}_{\text{"individual losses"}}+\underbrace{\Lambda(w)}_{\text{"regularizer"}}$$

<u>The empirical risk minimization (ERM) problem for decomposable loss functions</u>

בהינתן:

- מרחב פתרונות $C\subseteq\mathbb{R}^d$ (נקרא *feasible set*)
- דטהבייס $S=(x_1,\dots,x_n)\in D^n$
- פונקצייה $\ell: C\times D\to\mathbb{R}$

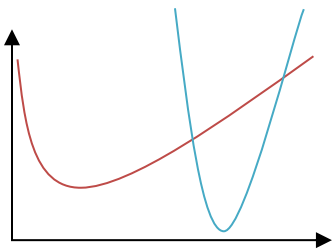
המשימה היא למצוא ווקטור $\widehat{w}\in C$ עם *loss* קטן ככל האפשר. ספציפית, נגדיר את השגיאה שלנו באופן הבא:

Excess empirical risk at $\widehat{w}$: $\quad L(\widehat{w};S)-\min_{w\in C}L(w;S)$

<u>איך נתכנן אלגוריתם פרטי לבעיית ה *ERM*?</u>

<u>אבחנה ראשונה:</u> אנחנו נהייה חייבים להגביל באיזושהי צורה את ההשפעה של נקודת קלט בודדת על פונקציית ההפסד.

למשל, אם שינוי של נקודת קלט בודדת משנה את פונקציית ההפסד שלנו מהכחול לאדום, אז אין לנו כל כך סיכויי לפתור את זה עם פרטיות כי הפתרון שאנחנו מחזירים לא אמור להיות רגיש לשינויי של נקודת קלט בודדת, מה שאומר שבשני המקרים אנחנו חייבים להחזיר את אותם פתרונות, אבל אין אף פתרון שהוא טוב לשני המקרים...



לצורך כך, נסתכל על ההגדרות הבאות:

<u>הגדרה 2: פונקציית הפסד חסומה *(bounded loss function)*:</u>
פונקיית הפסד $\ell: C \times D \to \mathbb{R}$ היא Δ-חסומה אם היא מקבלת ערכים בקטע $[0, \Delta]$ בלבד.

<u>אבחנה:</u> נשים לב שעבור פונקית הפסד חסומה כנ"ל מתקיים:
עבור דטהבייסים שכנים S, S' מתקיים $|L(w;S) - L(w;S')| \leq \frac{\Delta}{n}$

<u>הגדרה 3: פונקציית הפסד ליפשיץ *(Lipschitz)*:</u>
פונקיית הפסד $\ell: C \times D \to \mathbb{R}$ היא G-ליפשיץ אם לכל $x \in D$ ולכל זוג ווקטורים $v, w \in C$ מתקיים
$$|\ell(v;x) - \ell(w;x)| \leq G \cdot \|v - w\|_2$$

מה זה אומר אינטואיטיבית? שהגרף של הפונקציה אף פעם לא משתנה "יותר מדי מהר". זה אומר שבכל נקודה השיפוע של גרף הפונקציה, $\frac{\ell(v;x)-\ell(w;x)}{v-w}$, לא יכול להיות יותר מ G (בערך מוחלט).

דוגמאות:

- עבור בעיית החציון שראינו: $\ell(w;x) = |w - x|$. הפונקצייה הזאת היא 1-ליפשיץ.
  למה? נסו להשתכנע שלכל נקודה $x \in \mathbb{R}$ ולכל זוג פתרונות $v, w \in \mathbb{R}$ מתקיים
  $$|\ell(v;x) - \ell(w;x)| = \big||v - x| - |w - x|\big| \leq 1 \cdot |v - w|$$
  רמז: נסו לחשוב על שני המקרים הבאים:
  * v, w מאותו צד של x, למשל $x \leq v \leq w$
  * v, w מצדדים שונים של x, למשל $v \leq x \leq w$

- עבור בעיית הממוצע שראינו: $\ell(w;x) = (w - x)^2$. פונקצייה זו היא ליפשיץ תחת ההנחה ש- D, C חסומים.
  למה?
  $$\begin{aligned}|\ell(v;x) - \ell(w;x)| &= |(v - x)^2 - (w - x)^2| = |v^2 - 2xv + x^2 - w^2 + 2xw - x^2| \\ &= |v^2 - w^2 - 2x(v - w)| = |(v - w)(v + w) - 2x(v - w)| \\ &= |(v - w)(v + w - 2x)| = |v - w| \cdot \text{BoundOn}\{v + w - 2x\}\end{aligned}$$

<u>אלגוריתם *ERM* פרטי – נסיון ראשון – שימוש במכניזם האקספוננציאלי:</u>

- נתון דטהבייס $S = (x_1, \dots, x_n) \in D^n$
- נתונה פונקציית הפסד חסומה $\ell: C \times D \to [0, \Delta]$
- אנחנו מסמנים $L(w;S) = \frac{1}{n}\sum_{i=1}^{n} \ell(w;x_i)$
- נדגום ונחזיר ווקטור $\widehat{w} \in C$ מהתפלגות המקיימת $\Pr[\widehat{w} = w] \propto \exp\left(-\frac{\varepsilon n}{2\Delta} \cdot L(w;S)\right)$

<u>משפט:</u> האלג' הנ"ל מקיים ε-פ"ד.

הסבר: נובע מהתכונות של המכניזם האקספוננציאלי, מכיוון שהרגישות הגלובלית של L היא לכל היותר $\frac{\Delta}{n}$.

אוקיי. אז האלגוריתם הזה משמר פרטיות. מתי הוא מועיל?

<u>משפט:</u> נניח ש- $C \subseteq \mathbb{R}^d$ הוא כדור היחידה עם רדיוס R. כלומר,
$$C = \{w : \|w\|_2 \leq R\}$$
בנוסף, נניח כי ℓ היא G-ליפשיץ וגם Δ-חסומה, עבור $\Delta \leq G \cdot R$.
נקבע פרמטר $\beta > 0$. אזי, לכל דטהבייס $S \in D^n$, האלגוריתם הנ"ל מחזיר ווקטור $\widehat{w}$ כך שבהסתברות לפחות $1 - \beta$ מתקיים
$$\underbrace{L(\widehat{w}; S) - \min_{w \in C} L(w; S)}_{\text{Excess empirical risk}} = O\left(\frac{dGR}{\varepsilon n} \log\left(\frac{\varepsilon n}{d\beta}\right)\right)$$

הפאנץ' כאן זה שאם $n \gg d/\varepsilon$, אז הביטויי הזה קטן (בהנחה ש G, R קבועים). בהמשך נראה אלגוריתמים שמשיגים תוצאות טובות יותר (יספיק לנו ש $n \gg \sqrt{d}$). אבל בתור התחלה זה לא רע.

<u>הערה:</u> האלגוריתם הנ"ל איננו יעיל חישובית באופן כללי, כי אנחנו צריכים לדגום מההתפלגות "המוזרה" שמוגדרת באלגוריתם. (תחת ההנחה שפונקציית ההפסד שלנו היא קמורה, כן ניתן לממש את האלגוריתם הזה בזמן פולינומי, אבל לא בצורה פרקטית. בהמשך נראה אלגוריתמים יעילים הרבה יותר).

<u>הוכחת המשפט:</u>
נקבע דטהבייס $S = (x_1, \ldots, x_n) \in D^n$. נניח לשם פשטות שמתקיים $G = R = 1$.

נסמן
$$w^* = \operatorname*{argmin}_{w \in C} L(w; S)$$

נקבע $\beta > 0$ (הסתברות כישלון) ונסמן פרמטרים
$$r = \frac{2d}{n\varepsilon}$$
$$t = \frac{2}{\varepsilon n}\left(d \ln\left(\frac{2}{r}\right) + \ln\left(\frac{1}{\beta}\right)\right)$$

כעת נסתכל על שתי תתי הקבוצות הבאות של C:
$$\text{GOOD} = \{w \in C : L(w; S) \leq L(w^*; S) + r\}$$
$$\text{BAD} = \{w \in C : L(w; S) \geq L(w^*; S) + r + t\}$$

אנחנו רוצים להראות עכשיו שהנפח של הקבוצה GOOD לא יכול להיות יותר מדי קטן.

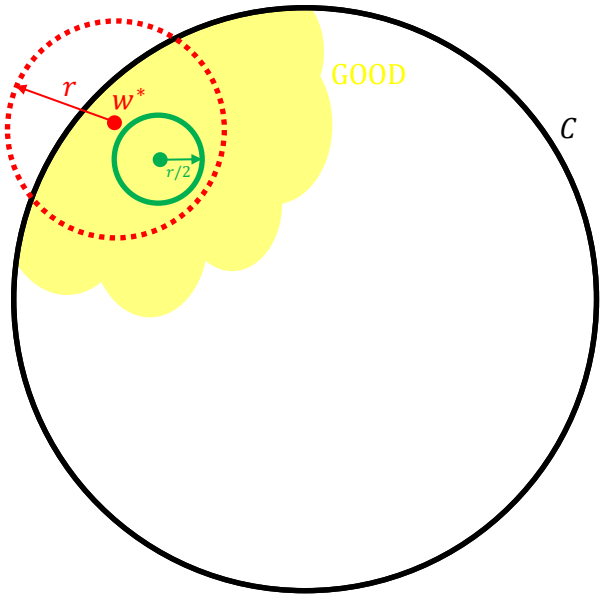
ראשית, נשים לב שמתקיים $w^* \in \text{GOOD}$.

בנוסף, לכל $w \in C$ כך ש- $\|w - w^*\|_2 \leq r$ מתקיים
$$\left| L(w; S) - L(w^*; S) \right| \underbrace{\leq}_{\text{ליפשיץ}} \|w - w^*\|_2 = r$$
ולכן
$$L(w; S) \leq L(w^*; S) + r$$
כלומר
$$w \in \text{GOOD}$$

כלומר, הקבוצה GOOD מכילה לא רק את w^*, אלא גם כל ווקטור ב- C שהמרחק שלו מ- w^* הוא לכל היותר r.

היינו רוצים להגיד שזה גורר שהקבוצה GOOD מכילה כדוק ברדיוס r סביב w^* ולכן הנפח שלה הוא לפחות כמו הנפח של כדור ברדיוס r. אבל זה רק כמעט נכון. המקרה הרע הוא שהווקטור w^* נמצא קרוב לקצה של C, כך שלא כל הכדור ברדיוס r סביבו מוכל ב C.

בציור:



בכל מקרה, הקבוצה *GOOD* מכילה כדור ברדיוס $r/2$.
לכן,

$$\Pr[\text{BAD}] \le \frac{\Pr[\text{BAD}]}{\Pr[\text{GOOD}]} \le \frac{\text{Vol}(\text{BAD}) \cdot \exp\left(-\frac{\varepsilon n}{2}(L(w^*;S)+r+t)\right)}{\text{Vol}(\text{GOOD}) \cdot \exp\left(-\frac{\varepsilon n}{2}(L(w^*;S)+r)\right)}$$

$$\le \frac{\text{Vol}(\text{Ball of radius } 1) \cdot \exp\left(-\frac{\varepsilon n}{2}(L(w^*;S)+r+t)\right)}{\text{Vol}(\text{Ball of radius } r/2) \cdot \exp\left(-\frac{\varepsilon n}{2}(L(w^*;S)+r)\right)} = ((1))$$

עובדה: נפח של כדור ברדיוס r במיימד d ניתן להצגה כ-

$$\text{Vol}(\text{Ball of radius } r) = f(d) \cdot r^d$$

כלומר, כפונקציה של d בלבד כפול r^d.
לכן,

$$((1)) = \frac{\exp\left(-\frac{\varepsilon n}{2}(L(w^*;S)+r+t)\right)}{\left(\frac{r}{2}\right)^d \cdot \exp\left(-\frac{\varepsilon n}{2}(L(w^*;S)+r)\right)} = \left(\frac{2}{r}\right)^d \cdot \exp\left(-\frac{\varepsilon n t}{2}\right) = \exp\left(d\ln\left(\frac{2}{r}\right) - \frac{\varepsilon n t}{2}\right) \underbrace{=}_{\text{הצבת } t} \beta$$

לכן, בהסתברות לפחות $\beta - 1$ נקבל ווקטור <u>שלא</u> נמצא ב *BAD*, כלומר ווקטור w עם *excess risk* לכל היותר

$$r + t \underbrace{=}_{\text{הצבת } r,t} \frac{2d}{n\varepsilon} + \frac{2}{\varepsilon n}\left(d\ln\left(\frac{n\varepsilon}{d}\right) + \ln\left(\frac{1}{\beta}\right)\right)$$