10.11.2022 ***פרטיות מידע***

הרצאה 3: כלים בסיסיים

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

מרצה: אורי שטמר

חישוב חציון:

נתון דטהבייס $X = (x_1, \dots, x_n)$ כאשר לכל i מתקיים $x_i \in [0,1]$.
אנחנו מעוניינים לחשב/להעריך את $\text{Median}(X)$.

<u>שאלה:</u> מה הרגישות הגלובלית של $\text{Median}(X)$?

<u>תשובה:</u> $\text{GS}_{\text{Median}} = 1$, כאשר 1 הוא <u>אורך הקטע</u>!

דוגמה:

- $X = (0,0,0,1,1,1,1)$ החציון הוא 1
- $X = (0,0,0,0,1,1,1)$ החציון הוא 0

ברור שזה לא יעזור לנו להוסיף רעש לפלאסי מסדר גודל של אורך הקטע, כי אז אנחנו נפסיד את כל המידע...

המכניזם האקספוננציאלי:

עכשיו נראה אלגוריתם אחר שבעזרתו נוכל להעריך את החציון של הדטה. האלגוריתם הזה לא בדיוק מחשב איזשהו מספר ומוסיף רעש לתוצאת החישוב, אלא בוחר באקראי תשובה מתוך אוסף של תשובות אפשריות.

<u>קלט:</u> $X \in D^n$

אנחנו מעוניינים לבחור תשובה מטווח של תשובות אפשריות R.

נניח שקיימת פונקציה $q: \underbrace{D^n}_{\text{דטהבייס}} \times \underbrace{R}_{\substack{\text{תשובה}\\ \text{אפשרית}}} \to \underbrace{\mathbb{R}}_{\substack{\text{אמדן לכמה התשובה האפשרית}\\ \text{הזאת טובה}}}$

כאשר $q(X,r)$ "גדול" אם r היא תשובה "טובה" עבור הדטהבייס X. אנחנו חושבים על q כעל פונקציית "ציון" או "איכות".
דוגמה: נניח שהדטהבייס מכיל הצבעות בבחירות (כלומר כל שורה מכילה הצבעה של אדם אחד) ואנחנו מעוניינים לזהות בצורה פרטית את המועמד שקיבל הכי הרבה קולות. במקרה זה, הפונק' q צריכה להיות גדולה יותר עבור מועמד r ככל שיש לו יותר קולות כי אז אינטואיטיבית הוא מועמד "טוב יותר".

הרגישות של פונק' הציון q מוגדרת כך:

$$\Delta q = \max_{r \in R} \max_{\substack{X,X' \\ \text{שכנים}}} |q(X,r) - q(X',r)|$$

<u>"האלגוריתם":</u> בקלט X, בחר $r \in R$ בהסתברות פרופורציונית ל- $\exp\left(\frac{\varepsilon \cdot q(X,r)}{2 \cdot \Delta q}\right)$

* כאשר הטווח R דיסקרטי, ההסתברות לבחירת $r \in R$ היא $\frac{\exp\left(\frac{\varepsilon \cdot q(X,r)}{2 \cdot \Delta q}\right)}{\sum_{f \in R} \exp\left(\frac{\varepsilon \cdot q(X,f)}{2 \cdot \Delta q}\right)}$

<u>טענה:</u> המכניזם האקספוננציאלי משמר ε-פ"ד.

<u>הוכחה:</u>
נסמן את המ.אקספ' ב A. עלינו להוכיח כי לכל X, X' שכנים ולכל פלט אפשרי $r \in R$ מתקיים
$$\Pr[A(X) = r] \leq e^{\varepsilon} \cdot \Pr[A(X') = r]$$

נזכור שאמרנו ש-
$$\Pr[A(X) = r] = \frac{\exp\left(\frac{\varepsilon \cdot q(X,r)}{2 \cdot \Delta q}\right)}{\sum_{f \in R} \exp\left(\frac{\varepsilon \cdot q(X,f)}{2 \cdot \Delta q}\right)}$$

מה הקשר בין $q(X,r)$ לבין $q(X',r)$?
לפי הגדרת Δq, ההבדל ביניהם הוא לכל היותר Δq. לכן,

$$\Pr[A(X) = r] \leq \frac{\exp\left[\frac{\varepsilon}{2 \cdot \Delta q} \cdot (q(X',r) + \Delta q)\right]}{\sum_{f \in R} \exp\left[\frac{\varepsilon}{2 \cdot \Delta q} \cdot (q(X',f) - \Delta q)\right]}$$

$$= \frac{e^{\frac{\varepsilon}{2}} \cdot \exp\left(\frac{\varepsilon \cdot q(X',r)}{2 \cdot \Delta q}\right)}{e^{-\frac{\varepsilon}{2}} \cdot \sum_{f \in R} \exp\left(\frac{\varepsilon \cdot q(X',f)}{2 \cdot \Delta q}\right)} = e^{\varepsilon} \cdot \Pr[A(X') = r]$$

מ.ש.ל.

<u>נחזור לבעיית החציון:</u>
נניח דיסקרטיזציה על הקטע *[0,1]*, נניח בקפיצות של $\frac{1}{1000}$.

בהינתן קלט של n נקודות מהגריד, נרצה לבחור נקודת גריד קרובה לחציון.
נרצה להשתמש במ.אקספ'.

נגדיר עבור דטהבייס X ונק' גריד ℓ ציון באופן הבא:

$$q(X,\ell) = -\left|\left|\{j \in X : j \geq \ell\}\right| - \left|\{j \in X : j \leq \ell\}\right|\right|$$

נשים לב ש- $\Delta q = 2$.

<u>הערה:</u> אם יתכנו חזרות בדטהבייס, אז צריך לשנות קצת את פונקציית הציון הזאת. למשל, אם הדטהבייס מכיל $3n/4$ עותקים של הנקודה 0 ועוד $n/4$ עותקים של הנקודה 1, אז לאף נקודה אין ציון אפס. בפרט, לנקודה 0 (שהיא הפתרון שהיינו רוצים לקבל) יש ציון $-n/4$.

אפשרות אחת לתיקון פונקציית הציון עבור מקרה כזה היא:

$$q(X,\ell) = -\left| \min\left\{\frac{n}{2}\ ,\ \ |\{j \in X : j \geq \ell\}|\right\} - \min\left\{\frac{n}{2}\ ,\ \ |\{j \in X : j \leq \ell\}|\right\} \right|$$

עכשיו תמיד קיים פתרון עם ציון 0. בנוסף, מבחינת *utility* הפונקציה הזאת עדיין משיגה את אותה מטרה. ספציפית, נניח נתון פתרון ℓ עם $q(X,\ell) \geq -\Delta$. באופן טריוויאלי, או שבדטהבייס X יש לפחות $n/2$ נקודות גדולות שוות מ ℓ או שיש לפחות $n/2$ נקודות קטנות שוות מ ℓ. נניח בה"כ שיש לפחות $n/2$ נקודות גדולות שוות מ ℓ. לכן, מכיוון ש $q(X,\ell) \geq -\Delta$ אנחנו מסיקים שישנם גם לפחות $\frac{n}{2} - \Delta$ נקודות קטנות שוות מ ℓ. כלומר הנקודה ℓ היא בערך מאוזנת.

לשם פשטות אנחנו נניח עכשיו שאין חזרות בדטהביס.

איזה פתרון נקבל מהמ.אקספ?

אנחנו "מקווים" לקבל פתרון ℓ עם ציון $q(X,\ell) = 0$.

נקבע פתרון "רע": ℓ עם ציון $q(X,\ell) = -w$. מה ההסתברות לקבל את ℓ ?

$$\Pr[A(X) = \ell] = \frac{\exp\left(-\frac{\varepsilon \cdot w}{4}\right)}{\sum_{f \in R} \exp\left(\frac{\varepsilon \cdot q(X,f)}{4}\right)} \leq \frac{\exp\left(-\frac{\varepsilon \cdot w}{4}\right)}{\exp\left(\frac{\varepsilon \cdot 0}{4}\right)} = \exp\left(-\frac{\varepsilon \cdot w}{4}\right) \underbrace{\leq}_{\substack{\text{נרצה} \\ \text{שיתקיים}}} \beta$$

וזה מתקיים עבור $w \geq \frac{4}{\varepsilon}\ln\left(\frac{1}{\beta}\right)$. כלומר, ההסתברות לקבל פתרון ספציפי עם ציון נמוך מ $-\frac{4}{\varepsilon}\ln\left(\frac{1}{\beta}\right)$ היא לכל היותר β.

לפי חסם האיחוד:

$$\Pr\left[\begin{array}{c}\text{המ.אקספ מחזיר}\\ \text{איזשהו פתרון עם}\\ \text{ציון} < -\frac{4}{\varepsilon}\ln\left(\frac{1}{\beta}\right)\end{array}\right] \leq 1000 \cdot \beta$$

לחילופין,

$$\Pr\left[\begin{array}{c}\text{המ.אקספ מחזיר}\\ \text{איזשהו פתרון עם}\\ \text{ציון} < -\frac{4}{\varepsilon}\ln\left(\frac{1000}{\beta}\right)\end{array}\right] \leq \beta$$

בעצם מה שעשינו כאן היה נכון לאו דווקא עבור בעיות החציון. באופן יותר כללי אנחנו מקבלים:

$$\mathbf{Pr}\begin{bmatrix}\text{המ.אקספ מחזיר}\\ \text{איזשהו פתרון עם}\\ \mathbf{OPT}-\boldsymbol{\lambda}>\text{ציון}\end{bmatrix}\leq |\boldsymbol{R}|\cdot\mathbf{exp}\left(-\frac{\boldsymbol{\varepsilon}\cdot\boldsymbol{\lambda}}{\mathbf{2}\cdot\boldsymbol{\Delta q}}\right)$$

כאשר

$$\mathbf{OPT}=\max_{\boldsymbol{r}\in\boldsymbol{R}}\{\boldsymbol{q}(\boldsymbol{X},\boldsymbol{r})\}$$

<u>עוד דוגמה לשימוש במ.אקפס:</u>
נתונה רשימת סרטים $r_1,\dots,r_t$.
הדטהבייס שלנו $X=(x_1,\dots,x_n)$ מכיל לכל משתמש את רשימת הסרטים שהוא אוהב.
מטרה: לבחור סרט פופולרי.

איך נגדיר את $q(\cdot,\cdot)$?

$q(X,r_j)=$ מספר המשתמשים שאוהבים את הסרט r_j.

מה הרגישות?

$\Delta q=1$ כי שינוי של ההעדפות של משתמש אחד יכול לשנות בלכל היותר 1 את הציון של כל סרט.

דוגמה נוספת לשימוש במכניזם האקספוננציאלי: מענה על שאילתות ספירה

נתון דטהבייס $X\in D^n$ ונתון אוסף C של שאילתות ספירה.

כל שאילתת ספירה $c\in C$ מוגדרת כך:

$$c\colon D\to\{0,1\}$$

וכדי לחשב את c על X נבצע:

$$c(X)=\frac{1}{n}\sum_{i=1}^{n}c(x_i)$$

כלומר, $c(X)=$ החלק של השורות ב- X שמקיימות את התכונה c.
<u>המטרה:</u> לתכנן אלג' פרטי שמחזיר תשובה מקורבת לכל $c\in C$

<u>השאלה:</u> מה צריך להיות n כפונקציה של פרמטרי הפרטיות שלנו, של הדיוק הרצוי, ושל מספר השאילתות $|C|$?

<u>חימום:</u> מה קורה אם $|C|=1$? כמה אפשר למתוח את זה?

<u>משפט:</u> לכל דטהבייס X קיים דטהביס $\hat{X}$ בגודל $m=O\left(\frac{\log|C|}{\alpha^2}\right)$ כך שמתקיים

$$\max_{h\in C}\left|h(X)-h\left(\hat{X}\right)\right|\leq\frac{\alpha}{2}$$

איך נשתמש במשפט הזה כדי לתכנן אלגוריתם?

נרצה להשתמש במ.אקספ. לצורך כך עלינו להגדיר מהי קבוצת הפתרונות ומהי q.

קבוצת הפתרונות תכיל את כל הדטהבייסים בגודל m:

$$R = D^m$$

פונק' הציון עבור דטהבייס $X \in D^n$ ופתרון (גם דטהבייס, אבל קטן יותר) $\hat{X} \in D^m$:

$$q(X,\hat{X}) = -\max_{h\in C}\left|h(X) - h(\hat{X})\right|$$

מהי הרגישות? $\Delta q \le \frac{1}{n}$

אז מה המ.אקספ מבטיח?

$$\Pr\left[\begin{matrix}\text{המ.אקספ בוחר } \hat{X} \text{ כך ש}\\ q(X,\hat{X}) < \mathrm{OPT} - \frac{\alpha}{2}\end{matrix}\right] \le |R|\cdot\exp\left(-\frac{\varepsilon\cdot\alpha n}{4}\right) \underbrace{\le}_{\text{נרצה שיתקיים}} \beta$$

מה שמתקיים כאשר

$$n \ge \frac{4}{\alpha\varepsilon}\ln\left(\frac{|R|}{\beta}\right) = \frac{4}{\alpha\varepsilon}\ln\left(\frac{|D|^m}{\beta}\right) = O\left(\frac{m}{\alpha\varepsilon}\ln\left(\frac{|D|}{\beta}\right)\right) = O\left(\frac{\log|C|\cdot\log\left(\frac{|D|}{\beta}\right)}{\alpha^3}\right)$$

אבל אנחנו יודעים שמתקיים $\mathrm{OPT} \ge -\frac{\alpha}{2}$ ולכן בהסתברות לפחות $1-\beta$ נקבל פתרון (=דטהבייס) $\hat{X}$ כך ש-

$$q(X,\hat{X}) \ge -\alpha$$

כלומר

$$\max_{h\in C}\left|h(X) - h(\hat{X})\right| \le \alpha$$

* דוגמה מעניינת: נניח שהדטהבייס X מכיל נקודות מגריד סופי כלשהו $D \subseteq \mathbb{R}$ ונניח ש-
$C = \{c_d : d \in D\}, \quad c_d(x) = 1 \iff x \le d$