تمرین سری ۵ SPML

امیرمحمد عزتی – ۴۰۲۲۱۲۲۶۹

سوال ۱:

یک مکانیزم ε-DP است اگر:

$$\frac{Pr[M(x) \in S]}{Pr[M(x') \in S]} \leq e^{\varepsilon}$$

برای پرسش شمارشی $f(X) = \sum_{i=1}^{n} X_i$, $X_i \in \{0,1\}$

فاصله دو دیتاست همسایه برابرات با $\Delta f = 1$

اگر $\tilde{f}(X)$ را مانند روبه رو تعریف کنیم: $\tilde{f}(X) = \sum_{i=1}^{n} X_i + Z$

$$Z \sim uniform\left(\left[\frac{-3}{\varepsilon}, \frac{3}{\varepsilon}\right]\right)$$

آنگاه تابع $\tilde{f}(X)$ نیز از یک توزیع یکنواخت با بازه زیر می‌آید: (خط شیفت خورده است)

$$X: \left[\sum_{i=1}^{n} X_i - \frac{3}{\varepsilon}, \sum_{i=1}^{n} X_i + \frac{3}{\varepsilon}\right]$$

همسایه

$$X': \left[\sum_{i=1}^{n} X_i + 1 - \frac{3}{\varepsilon}, \sum_{i=1}^{n} X_i + 1 + \frac{3}{\varepsilon}\right]$$

چون PDF توزیع یکنواخت Z برابرات با

$$P_Z(z) = \begin{cases} \frac{\varepsilon}{6} & z \in \left[\frac{-3}{\varepsilon}, \frac{3}{\varepsilon}\right] \\ 0 & \text{otherwise} \end{cases}$$

پس آنجایی که $\tilde{f}(X)$ با $\tilde{f}(X')$ اشتراک دارند نسبت احتمال برابرات با:

$$\frac{Pr[\tilde{f}(X) = y]}{Pr[\tilde{f}(X') = y]} = \frac{\frac{\varepsilon}{6}}{\frac{\varepsilon}{6}} = 1 \leq e^{\varepsilon} \quad \checkmark$$

اما آنجایی که اشتراک خروجی دو دیتاست تهی باشد، احتمال برای یک دیتاست ۰ می‌شود.

و نسبت احتمالات مکانیزم برای دو دیتاست بی‌نهایت (∞) می‌شود که $\infty \gg e^{\varepsilon}$ است.

ویژگی هایی که باعث می‌شود نویز لاپلاس ε-DP شود:

1. رنج دامنه آن نامحدود است
2. هیچ جایی مقدار احتمال صفر ندارد
3. احتمال بصورت یکنواخت و نمایی با دورشدن از مرکز (جمع واقعی) کاهش می‌یابد برای همین همواره محدود به e^{ε} می‌ماند.

پس نویز یکنواخت ε-DP نیست آما نویز لاپلاس بخاطر ویژگی های بالا ε-DP است.

سوال ۲:

دو دیتابیس همسایه x و x' بصورتی تعریف می‌شوند که یکی از سطرهای دیتابیس شان باهم متفاوت باشد.

$$x, x' \in \mathbb{X}^n$$

$$\begin{cases} x = (x_1, x_2, \dots, x_n) \\ x' = (x_1', x_2, \dots, x_n) \end{cases} \quad \text{where} \quad x_1 \neq x_1'$$

حال چون مکانیزم M بصورت رندوم m داده از دیتابیس انتخاب می‌کند و بعنوان خروجی می‌دهد پس احتمال انتخاب یک داده خاص برابر $\frac{m}{n}$ خواهد بود.

حال فرض می‌کنیم مجموعه S خروجی مکانیزم‌مان است درحالتی که یک پیشامد خاص رخ دهد.

حال پیشامد مدنظرمان را حالتی قرار می‌دهیم که مکانیزم داده‌ای از x را انتخاب کند که در x' موجود نیست. در آن صورت احتمال اینکه مکانیزم خروجی یکسان S را در دو دیتابیس بدهد را ارزیابی می‌کنیم.

همانطور که گفته شده احتمال انتخاب یک داده از دیتابیس x برابر $\frac{m}{n}$ است.

$$Pr\left[M(x) \in S\right] = \frac{m}{n}$$

امّا چون فرض کردیم داده‌ای انتخاب می‌شود که در x' نیست پس:

$$Pr\left[M(x') \in S\right] = 0$$

طبق شرط برقرار بودن (ε, δ) DP:

$$Pr\left[M(x) \in S\right] \leq e^{\varepsilon} Pr\left[M(x') \in S\right] + \delta$$

$$\frac{m}{n} \leq e^{\varepsilon} \times 0 + \delta \Rightarrow \boxed{\delta \geq \frac{m}{n}}$$

در صورت برقراری $(\varepsilon-\delta)$ DP

پس به نقض شرط $\delta < \frac{m}{n}$ رسیدیم.

سوال ۳:

یک مجموعه داده $X \in \mathcal{X}^n$ ورودی مکانیزم

یک مجموعه اشیاء H که هر $h \in H$ یک خروجی ممکن برای مکانیزم است

یک تابع امتیاز $s: \mathcal{X}^n \times H \to \mathbb{R}$

پس $s(X, h)$ یک تابع امتیاز برای ورودی X و خروجی h است.

$$\Delta s = \max_{h \in H} \max_{X, X'} \left| s(X,h) - s(X',h) \right|$$

X, X' دو دیتاست مجاور

$M(X) = h$

مکانیزم نمایی: $\exp\left(\frac{\varepsilon s(X,h)}{2\Delta s}\right)$

برای اینکه این مکانیزم ε-DP باشد باید ثابت کنیم:

$$\frac{Pr\left[M(X) = h\right]}{Pr\left[M(X') = h\right]} \leq e^{\varepsilon} \xrightarrow[\text{Symmetry}]{\text{also}} \frac{P\left[M(X') = h\right]}{P\left[M(X) = h\right]} \geq e^{-\varepsilon}$$

$$Pr\left[M(X) = h\right] = \frac{\exp\left(\frac{\varepsilon\, s(X,h)}{2\Delta s}\right)}{\sum_{h \in H} \exp\left(\frac{\varepsilon\, s(X,h)}{2\Delta s}\right)} \longrightarrow Z_X$$

$$\frac{Pr\left[M(X)=h\right]}{Pr\left[M(X')=h\right]} = \frac{\exp\left(\frac{\varepsilon S(X,h)}{2\Delta S}\right)}{\exp\left(\frac{\varepsilon S(X',h)}{2\Delta S}\right)} \times \frac{Z_{X'}}{Z_X}$$

$$\frac{Pr\left[M(X)=h\right]}{Pr\left[M(X')=h\right]} = \exp\left(\frac{\varepsilon\left[S(X,h)-S(X',h)\right]}{2\Delta S}\right) \times \frac{Z_{X'}}{Z_X}$$

ΔS

$$\exp\left(\frac{\varepsilon}{2}\right)$$

$$\begin{cases} Z_X = \sum_{h\in H} \exp\left(\frac{\varepsilon S(X,h)}{2\Delta S}\right) \\ Z_{X'} = \sum_{h\in H} \exp\left(\frac{\varepsilon S(X',h)}{2\Delta S}\right) \end{cases}$$

$$\exp\left(\frac{\varepsilon S(X,h)}{2\Delta S}\right) \leq \exp\left(\frac{\varepsilon}{2}\right) \times \exp\left(\frac{\varepsilon S(X',h)}{2\Delta S}\right)$$

$$\exp\left(\frac{\varepsilon S(X',h)}{2\Delta S}\right) \leq \exp\left(\frac{\varepsilon}{2}\right) \times \exp\left(\frac{\varepsilon S(X,h)}{2\Delta S}\right)$$

Sum over $h\in H$ $\Rightarrow$ $\boxed{Z_{X'} \leq Z_X \exp\left(\frac{\varepsilon}{2}\right)}$ $\Rightarrow$ $\boxed{\frac{Z_{X'}}{Z_X} \leq \exp\left(\frac{\varepsilon}{2}\right)}$

$$\Rightarrow \frac{Pr\left[M(X)=h\right]}{Pr\left[M(X')=h\right]} \leq \exp\frac{\varepsilon}{2} \cdot \exp\frac{\varepsilon}{2} = \exp(\varepsilon) = e^{\varepsilon}$$

سوال ۴:

۱) مفاهیم اولیه

الف) حمله استنتاج عضویت (Membership Inference Attack) یک آسیب‌پذیری حریم خصوصی است که در آن یک مهاجم تلاش می‌کند بفهمد که آیا یک نمونه داده خاص در مجموعه داده‌های آموزشی یک مدل قرار داشته است یا خیر. این حمله از الگوها یا سوگیری‌هایی که مدل ممکن است به طور ناخواسته در طول آموزش به خاطر سپرده باشد، بهره می‌گیرد.

ارزیابی عضویت در حمله MIA با تابع امتیازدهی $M\big(x, Access(\Theta)\big)$ بصورت زیر است:

x: نمونه ورودی است که وضعیت عضویت آن باید بررسی شود.

$Access(\Theta)$: دسترسی مهاجم به مدل را نشان می‌دهد، که ممکن است شامل احتمالات، لاجیت‌ها یا گرادیان‌های مدل با پارامترهای Θ باشد.

خروجی $M\big(x, Access(\Theta)\big)$: یک امتیاز عددی است که احتمال اینکه x بخشی از داده‌های آموزشی باشد را کمی می‌کند.

این فرآیند شامل مقایسه این امتیاز با یک آستانه از پیش تعیین‌شده τ است. اگر $M\big(x, Access(\Theta)\big) \geq \tau$ باشد نمونه به عنوان بخشی از مجموعه آموزشی شناسایی می‌شود. در غیر این صورت، به عنوان یک نمونه خارجی در نظر گرفته می‌شود. این تابع امتیازدهی با شناسایی overfit یا رفتارهای غیرعادی در پاسخ مدل به x که ممکن است ناشی از به خاطر سپردن داده‌ها باشد، عمل می‌کند.

ب)

۱. Proximal Policy Optimization (PPO)

یک روش یادگیری تقویتی که در آموزش مدل‌های زبانی بزرگ (LLM) استفاده می‌شود. از ترجیحات انسانی از طریق آموزش یک مدل پاداش بر اساس داده‌های ترجیحات جفتی بهره می‌برد. این فرآیند شامل سه مرحله است:

Supervised Fine-Tuning (SFT): مدل اولیه را آموزش می‌دهد.

Reward Modeling: یک تابع پاداش با استفاده از جفت‌های ترجیح داده‌ها آموزش می‌دهد.

Policy Optimization: مدل را با استفاده از تابع پاداش با روش PPO تنظیم می‌کند، به گونه‌ای که به‌روزرسانی‌ها به طور قابل توجهی از سیاست اولیه فاصله نگیرند.

داده‌های ترجیح از طریق تابع پاداش یادگرفته‌شده به طور غیرمستقیم بر مدل نهایی تأثیر می‌گذارند.

۲. Direct Preference Optimization (DPO)

یک روش بهینه‌سازی مستقیم که آموزش را ساده کرده و مرحله مدل‌سازی reward را حذف می‌کند. مدل را مستقیماً بر اساس داده‌های ترجیحی تنظیم می‌کند، به طوری که به پاسخ‌های ترجیح داده شده (با احتمال بالاتر) تراز شود و پاسخ‌های با ارجحیت کمتر را (با احتمال پایین‌تر) جریمه کند. شامل یک فرآیند بهینه‌سازی تک‌مرحله‌ای است که مستقیماً از جفت‌های ترجیحی در آموزش استفاده می‌کند.

چرا مدل‌های DPO نسبت به MIA حساس‌تر هستند؟

۱. قرار گرفتن مستقیم در معرض داده‌های ترجیحی:

مدل‌های DPO مستقیماً با داده‌های ترجیحی آموزش داده می‌شوند که به مدل اجازه می‌دهد تا به راحتی این ورودی‌ها را overfit کند و حفظ نماید. این overfit الگوهایی ایجاد می‌کند که حملات MIA می‌توانند از آن‌ها بهره‌برداری کنند.

۲. مدل‌سازی پاداش ضمنی:

برخلاف PPO که از یک مدل پاداش جداگانه استفاده می‌کند، DPO ترجیحات را مستقیماً در مدل رمزگذاری می‌کند. این رویکرد ضمنی فاقد اثر تنظیمی مرحله مدل‌سازی پاداش است که خطرات حفظ حریم خصوصی را افزایش می‌دهد.

۳. تبادل بین حریم خصوصی و سادگی:

روش DPO به دلیل هم‌راستایی ساده و مستقیم با داده‌های ترجیحی، کارایی محاسباتی را افزایش می‌دهد، اما با هزینه افزایش خطرات حریم خصوصی همراه است.

۴. شکاف تعمیم(Generalization Gap):

PPO بین پیروی از ترجیحات انسانی و حفظ قابلیت‌های تعمیم مدل تعادل ایجاد می‌کند، که منجر به کمتر شدن overfit نسبت به داده‌های آموزشی در مقایسه با DPO می‌شود.

۲) تحلیل ریاضی

الف) شاخص Area Under the Receiver Operating Characteristic (AUROC) یک معیار عملکرد است که توانایی یک مدل برای تمایز بین دو کلاس برای مثال، "داده‌های آموزشی" در مقابل "داده‌های غیرآموزشی" در MIA را ارزیابی می‌کند. این معیار، توازن بین True Positive Rate (TPR)و False Positive Rate (FPR) را در سطوح مختلف آستانه تصمیم‌گیری اندازه‌گیری می‌کند.

۱. True Positive Rate (TPR):

$$TPR = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

نشان‌دهنده بخشی از نمونه‌های آموزشی است که به درستی شناسایی شده‌اند.

۲. False Positive Rate (FPR):

$$FPR = \frac{False\ Positives}{False\ Positives + True\ Negatives}$$

نشان‌دهنده بخشی از نمونه‌های غیرآموزشی است که به اشتباه به عنوان داده‌های آموزشی دسته‌بندی شده‌اند.

تفسیر:

- مقدار AUROC برابر 1.0 نشان‌دهنده تمایز کامل بین داده‌های آموزشی و غیرآموزشی است.
- مقدار AUROC برابر 0.5 به معنی حدس تصادفی است.
- مقادیر بالاتر AUROC نشان می‌دهند که مدل حساس‌تر به حملات MIA است، زیرا مهاجم می‌تواند راحت‌تر بین داده‌های آموزشی و غیرآموزشی تمایز قائل شود.

در زمینه MIA، یک AUROC بالا نشان می‌دهد که مدل داده‌های آموزشی را تا حدی حفظ کرده است که آن را در برابر نقض حریم خصوصی آسیب‌پذیر می‌کند.

ب) تحلیل تابع زیان DPO و تأثیر آن بر حساسیت بهMIA

تابع زیان DPO (معادله ۳)

$$L_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right].$$

۱. مفهوم اصلی:

هم‌راستایی مستقیم با ترجیحات: این تابع زیان اختلاف بین احتمالاتی که به پاسخ ترجیح داده‌شده (y_w) و پاسخ کمتر ترجیحی (y_l) تخصیص داده می‌شود را کاهش می‌دهد و آن‌ها را در مقابل مدل مرجع (π_{ref}) مقیاس‌بندی می‌کند.

$\pi_\theta(y \mid x)$: احتمال تخصیص داده شده به یک پاسخ y توسط مدل فعلی با توجه به پرامپت x.

$\pi_{ref}(y \mid x)$: احتمال تخصیص داده شده به پاسخ توسط مدل مرجع (معمولاً مدلSFT)

۲. فاکتور مقیاس‌بندی:

β: وزن تفاوت ترجیحات را کنترل می‌کند.

مقادیر بزرگ‌تر β تفاوت بین ترجیحات را با شدت بیشتری برجسته می‌کند و مدل را به سمت پاسخ ترجیح داده شده سوق می‌دهد.

۳. تفاوت احتمالات لگاریتمی:

این عبارت احتمال پاسخ ترجیح داده شده را افزایش داده و احتمال پاسخ کمتر ترجیحی را کاهش می‌دهد.

۴. تبدیل سیگموئید:

استفاده از تابع سیگموئید σ تضمین می‌کند که احتمالات بین ۰ و ۱ مقیاس‌بندی شوند و مدل در طول آموزش همگرا شود.

تأثیر معادله DPO بر حساسیت به MIA :

۱. تأثیر مستقیم داده‌های ترجیحی:

مدل مستقیماً بر اساس داده‌های ترجیحی (x, y_w, y_l) بهینه‌سازی می‌شود، که باعث می‌شود الگوهای خاص مرتبط با این جفت‌های آموزشی را حفظ کند.

برخلاف PPO، که از یک مدل پاداش واسطه استفاده می‌کند، DPO خروجی‌های مدل را به داده‌های آموزشی نزدیک‌تر می‌کند و حساسیت بیشتری ایجاد می‌کند.

۲. بزرگ‌نمایی تفاوت‌ها:

تمرکز بر حداکثرسازی شکاف بین $\pi_\theta(y_w \mid x)$ و $\pi_\theta(y_l \mid x)$ می‌تواند مدل را به شدت نسبت به داده‌های آموزشی حساس کند. این حساسیت منجر به رفتارهای متمایزی در مواجهه با داده‌های آموزشی می‌شود که حملات MIA می‌توانند از آن‌ها بهره‌برداری کنند.

۳. وابستگی به مدل مرجع:

اگر مدل مرجع π_{ref} به خوبی تعمیم یابد، انحرافات از آن در مدل آموزش دیده با DPO، overfit را بیشتر کرده و آن را بیشتر در معرض حمله قرار می‌دهد.

۴. نقش β:

یک مقدار بزرگ‌تر برای β این اثرات را تشدید می‌کند و حساسیت مدل را به حفظ و فاش کردن الگوهای خاص داده‌های آموزشی افزایش می‌دهد.

۳) مقاله بیان می‌کند که مدل‌های بزرگ‌تر به دلیل ظرفیت بالاتر برای حفظ داده‌های آموزشی، معمولاً حساسیت بیشتری به حملات Membership Inference Attack (MIA) دارند. این حساسیت افزایش یافته در مقادیر بالاتر AUROC هنگام ارزیابی اثربخشی حملات MIA منعکس می‌شود.

- ✓ AUROCبالاتر برای مدل‌های بزرگ‌تر:

آزمایش‌های مقاله نشان می‌دهند که مدل‌های بزرگ‌تر مانند GPT2-xl و Mistral-7B اغلب AUROC بالاتری در مقایسه با مدل‌های کوچک‌تر به دست می‌آورند. به عنوان مثال:

در دیتاست Stack-Exchange، مقادیر AUROC برای مدل‌های بزرگ به طور قابل توجهی بیشتر از مدل‌های کوچک‌تر است. این نشان می‌دهد که مدل‌های بزرگ جزئیات بیشتری از داده‌های آموزشی خود را حفظ می‌کنند و در نتیجه هدف آسان‌تری برای حملات MIA هستند.

- ✓ سادگی وظیفه و تعمیم‌پذیری:

در وظایف ساده‌تر (مانند داده‌ست IMDB)، مدل‌های بزرگ مانند Mistral-7B و GPT2-xl تعمیم بهتری نشان می‌دهند و اثربخشی حملات MIA کاهش می‌یابد. با این حال، در داده‌ست‌های پیچیده‌تر، حفظ الگوهای دقیق باعث افزایش حساسیت به MIA می‌شود.

- ✓ تشدید اثر داده‌های ترجیحی:

برای مدل‌های DPO بزرگ‌تر، مواجهه مستقیم با داده‌های ترجیحی مشکل حفظ الگوها را تشدید می‌کند و مقادیر AUROC حتی بیشتر از مدل‌های PPO می‌شود.

روش‌های پیشنهادی برای کاهش حساسیت مدل‌های بزرگ به MIA

برای کاهش آسیب‌پذیری مدل‌های بزرگ در برابر حملات MIA، می‌توان چندین روش را به کار برد:

۱. تکنیک‌های حریم خصوصی تفاضلی (Differential Privacy - DP)

از روش‌هایی مانند Differentially Private Stochastic Gradient Descent (DP-SGD) برای اطمینان از این که مشارکت هر داده خاص در فرآیند آموزش قابل تمایز نباشد میتواند استفاده شود. محدود کردن میزان تأثیر داده‌های خاص بر آموزش مدل، باعث کاهش overfit و حفظ الگوها می‌شود.

۲. تکنیک‌های منظم‌سازی (Regularization)

Dropout: غیرفعال کردن تصادفی نورون‌ها در طول آموزش برای جلوگیری از overfit.

Weight Regularization: مجازات مقادیر بزرگ وزن‌ها در مدل با استفاده از نرم‌های L_1 یا L_2.

این تکنینک توانایی تعمیم مدل‌های بزرگ را بهبود می‌بخشد و وابستگی آن‌ها به داده‌های خاص آموزشی را کاهش می‌دهد.

۳. استخراج دانش (Knowledge Distillation)

استفاده از یک مدل کوچک‌تر که بر اساس خروجی‌های مدل بزرگ آموزش دیده است و تعمیم را حفظ می‌کند در حالی که الگوهای خاص داده‌ها را کنار می‌گذارد. با این کار ظرفیت حفظ الگوها را با فشرده‌سازی مدل کاهش می‌دهد.

۴. افزایش داده (Data Augmentation)

Data Augmentation معرفی تنوع از طریق تولید نمونه‌های آموزشی مصنوعی یا افزایش داده‌های موجود. این کار نمایندگی داده‌های آموزشی را متنوع‌تر کرده و از overfit بیش از حد جلوگیری می‌کند.

۵. تنظیم دقیق با حفظ حریم خصوصی در نظر گرفته شده (Privacy-Aware Fine-Tuning)

استفاده از روش‌هایی مانند PPO به جای DPO، زیرا PPO مدل را به صورت غیرمستقیم با ترجیحات هماهنگ می‌کند و قرار گرفتن آن در معرض داده‌های حساس آموزشی را کاهش می‌دهد.