به نام خدا

دانشکده مهندسی کامپیوتر

امنیت و حریم خصوصی در یادگیری ماشین (۴۰۸۱۶)

نیم‌سال اول سال تحصیلی ۱۴۰۴-۱۴۰۳

استاد درس: دکتر امیرمهدی صادق‌زاده

طراحان: متین علی‌نژاد، رئوف زارع، علیرضا سخایی‌راد

تمرین پنجم

مهلت تحویل: ساعت ۲۳:۵۹ جمعه ۱۴ دی ۱۴۰۳

نکات و قواعد

۱. سوالات خود را زیر پیام مربوطه در Quera مطرح نمایید.

۲. لطفا مطابق تاکید پیشین، حتما آداب‌نامه‌ی انجام تمرین‌های درسی را رعایت نمایید. در صورت تخطی از آیین‌نامه، در بهترین حالت مجبور به حذف درس خواهید شد.

۳. در صورتی که پاسخ‌های سوالات نظری را به صورت دست‌نویس آماده کرده‌اید، لطفا تصاویر واضحی از پاسخ‌های خود ارسال کنید. در صورت ناخوانا بودن پاسخ ارسالی، نمره‌ای به پاسخ ارسال شده تعلق نمی‌گیرد.

۴. همه‌ی فایل‌های مربوط به پاسخ خود را در یک فایل فشرده و با نام SPML_HW۵_StdNum_FirstName_LastName ذخیره کرده و ارسال نمائید.

سوال ۱ (۲۵ نمره)

فرض کنید که می‌خواهیم به یک پرسش شمارشی پاسخ دهیم:

$$f(X) = \sum_{i=1}^{n} X_i,$$

که در آن $X_i \in \{0,1\}$ است. در مکانیزم لاپلاس، نویز لاپلاس با پارامتر مقیاس $1/\epsilon$ به سادگی به مجموع اضافه می‌شود. حال فرض کنید به جای آن، Z یک متغیر تصادفی پیوسته و یکنواخت باشد که به طور یکنواخت از بازه $[-3/\epsilon, 3/\epsilon]$ گرفته شده است. آماره زیر را در نظر بگیرید:

$$\tilde{f}(X) = \sum_{i=1}^{n} X_i + Z.$$

آیا $\tilde{f}$ $O(\epsilon)$ـمحرمانه تفاضلی است؟ اگر بله، آن را اثبات کنید. اگر خیر، توضیح دهید که چه ویژگی‌ای در توزیع لاپلاس باعث می‌شود که محرمانگی تفاضلی به الگوریتم داده شود و چرا توزیع یکنواخت توانایی ایجاد محرمانگی را ندارد.

سوال ۲ (۲۵ نمره)

فرض کنید یک پایگاه داده $x \in \mathcal{X}^n$ و $m \in \{0,1,\dots,n\}$ داده شده باشد. یک mـزیرنمونه تصادفی از x یک پایگاه داده جدید $x' \in \mathcal{X}^m$ است که با انتخاب تصادفی m ردیف از x و حذف مابقی $n-m$ ردیف ساخته می‌شود. نشان دهید که برای هر $n \in \mathbb{N}$، $|\mathcal{X}| \geq 2$، $m \in \{1,\dots,n\}$، $\varepsilon > 0$ و $\delta < \frac{m}{n}$، مکانیزم $M(x)$ که mـزیرنمونه تصادفی از $x \in \mathcal{X}^n$ را تولید می‌کند، دارای حریم خصوصی تفاضلی (ε, δ) نیست.
(راهنمایی: یک پیشمامد مناسب تعریف کنید و از برهان خلف استفاده کنید.)

سوال ۳ (۳۰ نمره)

مکانیزم نمایی یکی از ابزارهای اساسی در حریم خصوصی تفاضلی است که برای مشکلاتی طراحی شده که در آن‌ها نیاز به انتخاب یک گزینه از میان مجموعه‌ای از گزینه‌ها وجود دارد، به طوری که این انتخاب مبتنی بر داده‌های حریم خصوصی باشد، اما خود داده‌های حریم خصوصی افشا نشوند. برای درک بهتر، یک مثال کاربردی ارائه می‌دهیم:

یک مثال از حراج کالای دیجیتال: فرض کنید یک فروشنده تعداد نامحدودی از یک کالا (مانند نسخه دیجیتالی یک کتاب، فیلم یا بازی ویدئویی) در اختیار دارد. n نفر علاقه‌مند به خرید این کالا هستند، اما هر فرد فقط به یک نسخه از کالا نیاز دارد. هر فرد i حداکثر تا مبلغ ارزش‌گذاری‌شده‌ی خود، یعنی v_i، حاضر به پرداخت است. حال، فروشنده باید قیمت کالا p را به گونه‌ای تعیین کند که درآمد خود را حداکثر کند. روش متداول این است که فروشنده قیمت p را طوری انتخاب کند که درآمد حاصل از فروش، یعنی $p\times$ تعداد افرادی که حاضر به پرداخت هستند، به حداکثر برسد. با این حال، این رویکرد غیربهینه از نظر حفظ حریم خصوصی است؛ چرا که اگر فردی ارزش‌گذاری بسیار بالایی داشته باشد، این موضوع می‌تواند از انتخاب قیمت p توسط فروشنده مشخص شود.

حساسیت تابع درآمد: مشکل دیگر این است که تابع درآمد ممکن است به تغییرات کوچک در قیمت بسیار حساس باشد. برای مثال، فرض کنید سه فرد ارزش‌گذاری‌های 1 دلار، 1 دلار و 3.01 دلار برای یک بازی دارند: اگر قیمت $p = 1$ دلار باشد، درآمد 3 دلار خواهد بود. اگر قیمت به 1.01 دلار افزایش یابد، درآمد به 1.01 دلار کاهش می‌یابد. اگر قیمت $p = 3.01$ دلار باشد، درآمد مجدداً به 3.01 دلار می‌رسد. اما اگر قیمت به 3.02 دلار افزایش یابد، درآمد به صفر می‌رسد. این مثال نشان می‌دهد که تغییرات کوچک در قیمت می‌توانند تأثیرات بزرگی بر درآمد داشته باشند و تلاش برای خصوصی‌سازی قیمت با افزودن نویز به آن مؤثر نیست.

ایده مکانیزم نمایی: برای حل این مشکل، مکانیزم نمایی به جای در نظر گرفتن قیمت به عنوان یک «مقدار»، آن را به عنوان یک «شیء» در نظر می‌گیرد. قیمت‌هایی مانند 1 دلار یا 3.01 دلار به‌عنوان اشیاء با کیفیت بالا (زیرا درآمد زیادی تولید می‌کنند) شناخته می‌شوند، در حالی که قیمت‌هایی مانند 3.02 دلار اشیاء با کیفیت پایین هستند.

تعریف رسمی مکانیزم نمایی: مکانیزم نمایی ورودی‌های زیر را دریافت می‌کند:

- یک مجموعه داده $X \in X^n$،
- یک مجموعه اشیاء H،
- یک تابع امتیاز $s : X^n \times H \to \mathbb{R}$.

تابع امتیاز s میزان «کیفیت» هر شیء $h \in H$ را با توجه به مجموعه داده X مشخص می‌کند. در مثال بالا، مجموعه داده شامل ارزش‌گذاری‌های افراد است، مجموعه اشیاء تمام قیمت‌های ممکن را تشکیل می‌دهد، و تابع امتیاز درآمد حاصل از انتخاب هر قیمت را محاسبه می‌کند.

حریم خصوصی و حساسیت: در این تنظیم، فرض می‌کنیم مجموعه اشیاء و تابع امتیاز عمومی هستند و نیازی به حفظ حریم خصوصی آن‌ها نیست. تنها اطلاعات خصوصی مجموعه داده X است. برای سنجش میزان حساسیت تابع امتیاز به مجموعه داده، از رابطه زیر استفاده می‌کنیم:

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

که در آن X و X' دو مجموعه داده‌ی مجاور هستند.

نحوه عملکرد مکانیزم نمایی: مکانیزم نمایی شیء $h \in H$ را به گونه‌ای انتخاب می‌کند که احتمال انتخاب هر شیء h متناسب با مقدار زیر باشد:

$$\exp\left(\frac{\epsilon s(X,h)}{2\Delta}\right),$$

که در آن:

- ϵ یک پارامتر حریم خصوصی است که کنترل می‌کند چقدر انتخاب نهایی به تابع امتیاز وابسته باشد،
- Δ حساسیت تابع امتیاز است.

حال ثابت کنید مکانیزم نمایی ϵ حریم تفاضلی میباشد.

سوال ۴ (۲۰ نمره)

به منظور بررسی حملات استنتاج عضویت بر روی داده‌های ترجیحی برای تنظیم مدل‌های زبانی بزرگ (LLM)، مقاله *Exposing Privacy Gaps: Membership Inference Attack on Preference Data for LLM Alignment* را مطالعه کرده و به سوالات زیر پاسخ دهید:

لینک دانلود مقاله

۱. مفاهیم اولیه:

(الف) حمله استنتاج عضویت (MIA) را تعریف کنید. در این تعریف، نحوه عملکرد یک تابع امتیاز $M(x, \text{Access}(\Theta))$ برای شناسایی عضویت نمونه‌های آموزشی را توضیح دهید.

(ب) تفاوت‌های کلیدی بین روش‌های *PPO* و *DPO* در تنظیم مدل‌های زبانی بزرگ را شرح دهید. چرا مدل‌های تنظیم‌شده با *DPO* نسبت به *PPO* به حملات MIA حساس‌تر هستند؟

۲. تحلیل ریاضی:

(الف) مقاله از معیار *AUROC* برای ارزیابی اثربخشی حملات MIA استفاده می‌کند. این معیار را تعریف کرده و توضیح دهید چگونه نشان‌دهنده حساسیت مدل به حملات است.

(ب) معادله (۳) در مقاله، بهینه‌سازی مستقیم ترجیحات در روش *DPO* را مدل‌سازی می‌کند:

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

این معادله را تحلیل کنید و توضیح دهید چگونه به افزایش حساسیت مدل به حملات MIA منجر می‌شود.

۳. مقاله پیشنهاد می‌کند که مدل‌های بزرگ‌تر به دلیل ظرفیت بیشتر برای حفظ داده‌های آموزشی، به حملات MIA حساس‌تر هستند. با استفاده از نتایج آزمایشات در مقاله، تأثیر اندازه مدل بر *AUROC* را توضیح دهید و پیشنهاد دهید که چگونه می‌توان حساسیت مدل‌های بزرگ به حملات را کاهش داد.

سوال ۵ تمرین عملی (۲۰ + ۱۰۰ نمره)

نوت‌بوک *SPML_HW5_DP.ipynb* را تکمیل کنید.

موفق باشید.