به نام خدا

امنیت و حریم خصوصی در یادگیری ماشین (۴۰۸۱۶)

نیم‌سال اول سال تحصیلی ۱۴۰۵-۱۴۰۴

استاد درس: دکتر امیرمهدی صادق‌زاده

طراح: علی نجار

دانشکده مهندسی کامپیوتر

تمرین شماره ۴ | مهلت تحویل: ساعت ۲۳:۵۹:۵۹ ۳۰ دی ۱۴۰۴

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. از تحویل تمارین نظری به شکل تایپ شده (فایل Word-Latex یا هر نوع تایپ شده دیگری) بپرهیزید و حتما پاسخ دست نویس تحویل دهید.

۴. 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.

۵. همه‌ی فایل‌های مربوط به پاسخ خود را در یک فایل فشرده و با نام SPML_HW۴_StdNum_FirstName_LastName ذخیره کرده و ارسال نمائید.

سوال ۱ (۱۰ 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 $\tilde{f}$ مقدار $O(\epsilon)$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 تخمین میانگین محرمانه یک توزیع گاوسی (۶۰ نمره)

در این مسئله، محدودیت‌های تخمین میانگین یک توزیع گاوسی را تحت قید حریم خصوصی تفاضلی خالص $(\varepsilon, 0)$ بررسی می‌کنیم.

فرض کنید $X_1, \ldots, X_n$ نمونه‌های مستقل با توزیع یکسان (i.i.d) باشند که از یک توزیع گاوسی $\mathcal{N}(\mu, 1)$ نمونه‌برداری شده‌اند، به طوری که میانگین نامعلوم در یک بازه کران‌دار $\mu \in [-R, R]$ قرار دارد. هدف ما طراحی یک الگوریتم کارآمد $\mathcal{A}$ است که تخمینی نزدیک به μ ارائه دهد و در عین حال حریم خصوصی تفاضلی را برآورده کند.

۱. رویکرد برش‌دهی ساده (**The Naive Truncation Approach**)

ابتدا یک رویکرد ساده را در نظر بگیرید که در آن داده‌ها را برش می‌دهیم (Truncate) تا اطمینان حاصل کنیم حساسیت (Sensitivity) محدود است و سپس نویز لاپلاس اضافه می‌کنیم.

فرض کنید $C > 0$ یک عدد ثابت به اندازه کافی بزرگ باشد. تخمین‌گر میانگین تجربی برش‌خورده زیر را در نظر بگیرید:

$$\tilde{\mu}(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i \cdot \mathbf{1}\left[|X_i| \leq R + C\sqrt{\log n}\right]$$

فرض کنید الگوریتم $\mathcal{A}$ مکانیزمی باشد که نویز لاپلاس را متناسب با حساسیت $\tilde{\mu}$ به آن اضافه می‌کند.

ثابت کنید که یک الگوریتم $(\varepsilon, 0)$ـتفاضلی محرمانه $\mathcal{A}$ وجود دارد به طوری که با احتمال حداقل 0.97:

$$|\mathcal{A}(X_1, \ldots, X_n) - \mu| \lesssim \sqrt{\frac{1}{n}} + \frac{R + \sqrt{\log n}}{n\varepsilon}$$

۲. روش هیستوگرام (بهبود وابستگی به R)

کران به دست آمده در بخش (آ) هزینه خطی نسبت به R (دامنه میانگین) دارد. اگر R بزرگ باشد، این هزینه زیاد است. می‌توانیم با پیدا کردن یک تقریب اولیه از μ برای محدود کردن فضای جستجو، این وضعیت را بهبود بخشیم.

فرض کنید R یک عدد صحیح است. بازه $[-R, R]$ را به $2R - 2$ بازه با طول واحد I_j افراز می‌کنیم. می‌توانیم از یک الگوریتم هیستوگرام محرمانه استفاده کنیم تا بازه‌ای که بیشترین نقاط را در بر دارد شناسایی کنیم.

فرض کنید f تابع هیستوگرام برای این افراز باشد و I_j سطلی (Bucket) باشد که بیشترین نقاط در آن قرار دارند.

ثابت کنید با احتمال $\geq 1 - \exp(-\Omega(n))$، اگر j اندیس سطلی با بیشترین تعداد شمارش باشد، آنگاه $|\mu - j| \leq 5$. علاوه بر این، نشان دهید که $f_j(X_1, \ldots, X_n) \geq 0.1n$ و تعداد شمارش‌ها برای سطل‌های دورتر ناچیز است $(< 0.02n)$.

۳. ترکیب تطبیقی و کران‌های بهبود یافته

اکنون می‌توانیم ایده‌های بخش (آ) و (ب) را ترکیب کنیم. می‌توانیم از روش هیستوگرام برای یافتن یک بازه کوچک با اندازه ثابت (مثلاً شعاع ۵) که شامل μ است استفاده کنیم، و سپس الگوریتم برش‌دهی بخش (آ) را محدود به این بازه کوچک‌تر اجرا کنیم.

ثابت کنید یک الگوریتم کارآمد $(\varepsilon, 0)$ـتفاضلی محرمانه $\mathcal{A}$ وجود دارد به طوری که اگر $n \gtrsim \frac{\log R}{\varepsilon}$، آنگاه با احتمال حداقل $0.99 - \exp(-\Omega(n))$ داریم:

$$|\mathcal{A}(X_1, \ldots, X_n) - \mu| \leq 5$$

توضیح دهید چرا این نتیجه نشان‌دهنده یک بهبود نمایی در وابستگی به R در مقایسه با بخش (الف) است.

سوال ۳ تکنیک‌های پیشرفته برای اثبات **DP**-(ε, δ) (۳۰ نمره)

تعریف استاندارد حریم خصوصی تفاضلی (ε, δ) نیازمند بررسی یک شرط برای تمام زیرمجموعه‌های اندازه‌پذیر فضای خروجی است. با این حال، در عمل اغلب ساده‌تر است که «متغیر تصادفی اتلاف حریم خصوصی» (Privacy Loss Random Variable) یا واگرایی‌های آماری خاص بین توزیع‌های خروجی را تحلیل کنیم.

در این مسئله، شما دو ابزار قدرتمند برای اثبات حریم خصوصی تفاضلی را استخراج خواهید کرد: مشخصه‌سازی واگرایی هاکی‌ـاستیک (Hockey-Stick) و روش کران دنباله (Tail Bound).

فرض کنید $M(X)$ و $M(X')$ خروجی‌های یک مکانیزم تصادفی روی مجموعه داده‌های مجاور X و X' باشند. فرض کنید $f_{M(X)}$ و $f_{M(X')}$ به ترتیب نشان‌دهنده توابع چگالی احتمال آن‌ها باشند.

متغیر تصادفی اتلاف حریم خصوصی (**PLRV**) ،به صورت زیر تعریف می‌گردد:

$$L_{X' \| X}(Y') \;=\; \log\left(\frac{f_{X'}(Y')}{f_X(Y')}\right), \qquad Y' \sim M(X').$$

الف) مشخصه‌سازی واگرایی هاکی‌ـاستیک

یک تفسیر هندسی از حریم خصوصی تفاضلی شامل «واگرایی هاکی‌ـاستیک» است.

تعریف: واگرایی هاکی–استیک

برای توزیع‌های احتمال P و Q و پارامتر $\alpha > 0$، واگرایی هاکی–استیک به صورت زیر تعریف می‌شود:

$$H_\alpha(P\|Q) = \mathbb{E}_{y\sim Q}\left[\frac{dP}{dQ}(y) - \alpha\right]_+$$

که در آن $[x]_+ = x \cdot \mathbf{1}(x \geq 0)$ و $\frac{dP}{dQ}$ نسبت چگالی‌های P نسبت به Q است.

قضیه زیر را که این واگرایی را به حریم خصوصی تفاضلی مرتبط می‌کند، اثبات کنید.

قضیه ۱

مکانیزم M (ε, δ)-DP است اگر و تنها اگر برای هر دو مجموعه‌داده‌ی مجاور X, X' داشته باشیم:

$$H_{e^\varepsilon}\big(M(X')\|M(X)\big) = \int \big[f_{X'}(y) - e^\varepsilon f_X(y)\big]_+ \, dy \;\leq\; \delta.$$

ب) کران دنباله (**The Tail Bound Lemma**)

در حالی که مشخصه‌سازی واگرایی یک شرط معادل است، محاسبه مستقیم آن گاهی دشوار است. یک شرط کمی ضعیف‌تر (Looser) اما اغلب کاربرپسندتر، شامل کران‌دار کردن احتمال دنباله PLRV است.
قضیه زیر را اثبات کنید.

قضیه ۲ (کران دنباله برای DP)

اگر برای هر دو مجموعه‌داده‌ی مجاور X, X' داشته باشیم

$$\Pr_{Y'\sim M(X')}\Big[L_{X'\|X}(Y') \geq \varepsilon\Big] \leq \delta,$$

آنگاه M در (ε, δ)-DP صدق می‌کند.

ج) دقت (**Tightness**) کران‌ها

بر اساس اثبات خود در بخش (ب)، به اختصار بحث کنید که آیا شرط موجود در لم ۶ دقیق (Tight) است؟ به عبارت دیگر، آیا (ε, δ)-DP لزوماً ایجاب می‌کند که $\Pr[L \geq \varepsilon] \leq \delta$؟

سوال ۴ مکانیزم گاوسی (**Gaussian Mechanism**) (۳۰ نمره)

فرض کنید $f : \mathcal{X}^n \to \mathbb{R}^d$ یک تابع d–بعدی باشد، و بگذارید

$$\Delta_2 f = \max_{H(X,X')\leq 1} \|f(X) - f(X')\|_2$$

حساسیت ℓ_2 آن باشد. مکانیزم گاوسی با پارامتر σ، نویز مستقل $\mathcal{N}(0, \sigma^2) \sim$ را به هر یک از d مؤلفه خروجی f اضافه می‌کند.

قضیه

فرض کنید $\varepsilon \in (0, 1)$ و $\delta \in (0, 1]$ برای $c^2 > 2\log(\frac{1.32}{\delta})$ باشند، آنگاه مکانیزم گاوسی با پارامتر $\sigma \geq \frac{c\Delta_2 f}{\varepsilon}$ دارای ویژگی محرمانگی تفاضلی (ε, δ)–DP است.

راهنمایی: با استفاده از تقارن توزیع گاوسی بگویید

$$f(X') - f(X) = (\Delta, 0, \ldots, 0)$$

و سپس از مسئله قبل استفاده کنید.