به نام خدا

دانشکده مهندسی کامپیوتر

امنیت و حریم خصوصی در یادگیری ماشین

(۴۰۸۱۶)(نیم‌سال دوم سال تحصیلی ۱۴۰۱-۱۴۰۲)

استاد درس: دکتر امیرمهدی صادق‌زاده

دستیاران آموزشی: مهدی غزنوی، زینب گلگونی، الهه فرشادفر، محمدرضا کاظمی، حمید دشتبانی

تمرین ششم

مهلت تحویل: ساعت ۶:۰۰ پنجشنبه ۸ تیر ۱۴۰۲

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_HW۶_StdNum_FirstName_LastName ذخیره کرده و ارسال نمائید.

سوال ۱ تولید نمونه‌های مسموم (۱۷ نمره)

الگوریتم Poisoning Example Generation را با تکمیل دفترچه‌ی poisoning example generation.ipynb پیاده کنید. با در نظر گرفتن دو عکس frog.jpeg و ship.jpeg، هر بار یکی را به عنوان نمونه‌ی هدف و دیگری را به عنوان نمونه‌ی پایه در نظر گرفته، تصویر تولید شده توسط الگوریتم را نمایش دهید و برچسب پیش‌بینی شده توسط مدل آموزش داده شده برای آن‌ها پیش و پس از اعمال الگوریتم گزارش کنید.

سوال ۲ استخراج رگرسیون لاجستیک (۱۸ نمره)

در این سوال قصد داریم مبحث استخراج مدل را برای رگرسیون لاجستیک چند کلاسه بررسی کنیم. یک مسئله دسته‌بندی را از فضای ورودی $\mathcal{X}$ به $\mathcal{Y} = \{1, ..., c\}$ در نظر بگیرید. ابتدا یک نمونه از فضای ورودی توسط یک شبکه عصبی به فضای $\mathcal{H} = \mathbb{R}^n$ نگاشت می‌شود. سپس یک دسته‌بند رگرسیون لاجستیک ابتدا توسط رابطه $z = Wh + b$ بردار لاجیت z را تولید می‌کند که در آن $h \in \mathcal{H}$ و $z \in \mathbb{R}^c$ و $W \in \mathbb{R}^{c \times n}$ و $b \in \mathbb{R}^c$ است. سپس بردار لاجیت z با استفاده از تابع Softmax به بردار احتمال p تبدیل می‌شود که تابع Softmax به صورت $p_j = \frac{exp(z_j)}{\sum_{k=1}^{c} exp(z_k)}$ تعریف می‌شود که در آن z_j عنصر jـاُم بردار لاجیت z است. فرض کنید به تعداد دلخواه N می‌توانیم به مدل درخواست بدهیم. حال به سوالات زیر پاسخ دهید:

(الف) (۶ نمره) اگر مدل به ازای ورودی $x^{(i)}$ بردار لاجیت $z^{(i)}$ را خروجی دهد، روشی برای بازیابی وزن‌های رگرسیون لاجستیک ارائه دهید.

(ب) (۶ نمره) نشان دهید حداقل چند ورودی باید به مدل دهیم تا وزن‌های بازیابی شده قابل اعتبار باشند. آیا استفاده از ورودی‌های بیشتر تاثیری در وزن‌های بازیابی شده دارد؟ چرا؟

(ج) (۶ نمره) حال فرض کنید که مدل به ازای ورودی $x^{(i)}$ برادر احتمال $p^{(i)}$ را خروجی دهد، روشی برای بازیابی وزن‌های رگرسیون لاجستیک ارائه دهید. چه رابطه‌ای بین وزن‌های بازیابی شده با وزن‌های واقعی وجود دارد؟ (راهنمایی: از بردار احتمال لگاریتم بگیرید.)

سوال ۳ JbDA (۳۲ نمره)

در این سوال به بررسی و پیاده‌سازی روش Jacobian-based Dataset Augmentation (JbDA) می‌پردازیم.

(الف) (۱۲ نمره) سه عبارت زیر را برای قسمت افزودن (Augmentation) داده‌ها در نظر بگیرید که عبارت ۱، عبارت اصلی افزودن موجود در این روش است و دو عبارت دیگر، نسخه‌های تغییر یافته آن هستند. با توجه به هدف روش JbDA که تخمین کران‌های تصمیم‌گیری مدل قربانی یا همان اُراکل (Oracle) است، عملکرد هر کدام از سه عبارت را بررسی و مقایسه نمایید.

$$S_{\rho+1} = \{x + \lambda_{\rho+1}.sgn(J_F[\tilde{O}(x)]) : x \in S_\rho\} \cup S_\rho \quad (۱)$$

$$S_{\rho+1} = \{x - \lambda_{\rho+1}.sgn(J_F[\tilde{F}(x)]) : x \in S_\rho\} \cup S_\rho \quad (۲)$$

$$S_{\rho+1} = \{x - \lambda_{\rho+1}.sgn(J_F[\tilde{O}(x)]) : x \in S_\rho\} \cup S_\rho \quad (۳)$$

که F مدل جایگزین، O مدل اُراکل، J ماتریس ژاکوبین خروجی مدل F نسبت به ورودی، S_ρ نمونه‌های موجود در مرحله ρ افزودن و $\sim$ در بالای نماد یک مدل، نشان‌دهنده برچسب پیش‌بینی شده توسط آن مدل است ($\lambda > 0$).

(ب) (۲۰ نمره) دفترچه‌ی JbDA.ipynb را کامل کنید. در این قسمت قصد داریم با استفاده از الگوریتم JbDA برای شبکه‌ای که در اختیارتان قرارگرفته است (با وزن‌های موجود در checkpoint) ، یک مدل جایگزین (Substitue) بسازیم. در این قسمت تنها می‌توانید از ۱۰۰ نمونه تصادفی از مجموعه دادگان آموزش استفاده کنید. برای شبکه‌ی جایگزین از یک معماری کوچک ۲ لایه‌ی پیچشی + ۲ لایه خطی استفاده نمایید. بعد از آموزش، دقت شبکه‌ی جایگزین را بر روی مجموعه دادگان ارزیابی گزارش کنید.

سوال ۴ یک الگوریتم خصوصی[۱] دیگر! (۱۷ نمره)

فرض کنید ما میخواهیم به جستاری شمارشی [۲] بفرم $f(X) = \sum_{i=1}^{n} X_i$ که در آن $X_i \in \{0,1\}$ است پاسخ دهیم. در کلاس در باره ی مکانیزم لاپلاس یعنی افزودن نویز با پارامتر مناسب با $1/\epsilon$ آموختیم اما اگر به نویز لاپلاس دسترسی نداشتیم چه؟ فرض کنید که Z یک متغیر تصادفی پیوسته ی یکنواخت دربازه ی $[-3/\epsilon, 3/\epsilon]$ باشد. آماره ی $\tilde{f}(X) = \sum_{k=1}^{n} X_i + Z$ را در نظر بگیرید. آیا $\tilde{f}$ بصورت $O(\epsilon)$-differential private است؟ اگر جواب شما بله است آنرا بصورت بهترین ثابتی که privacy را تضمین میکند اثبات نمایید. اگر خیر،چرا؟

سوال ۵ نشر خصوصی هیستوگرام‌ها(چندجمله‌ای‌ها) (۱۷ نمره)

فرض کنید در صدد تخمین یک توزیع چندجمله‌ای یا به عبارتی یک هیستوگرام هستیم. به صورتی که ما مشاهدات $X \in \{1, 2, ..., k\}$ که در آن k میتواند بزرگ باشد و ما میخواهیم $p_j := \mathbb{P}(X = j)$ را برای $j = 1, ..., k$ تخمین بزنیم. برای نمونه ی x_j^n بردار شمارنده ی تجربی $\hat{p}_n$ با مقادیر $\hat{p}_{n,j} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i = j\}$ ابتدا ثابت کنید

$$Lip_{1,d_{ham}}(\hat{p}_n) \leq \frac{2}{n}$$

که در آن عبارت $Lip_{1,d_{ham}}(f)$ برای هر تابع $f : \mathcal{X}^n \to \mathbb{R}^d$ بصورت زیر تعریف می‌شود

$$Lip_{1,d_{ham}}(f) = \sup\{||f(x_1^n) - f(y_1^n)||_1 \mid d_{ham}(x_1^n, y_1^n) \leq 1\} \leq L$$

توجه شود که d_{ham} نیز بصورت زیر تعریف شده

$$d_{ham}\left(\{x_1, ..., x_n\}, \{x_1^{'}, ..., x_n^{'}\}\right) = \sum_{i=1}^{n} \mathbf{1}\{x_i \neq x_i^{'}\}$$

حال با کمک نامساوی بالا و مفاهیم تدریس شده در کلاس برای مکانیزم لاپلاس بصورت

$$Z = \hat{p}_n + W, \qquad W_j \overset{iid}{\sim} Laplace(\frac{2}{n\epsilon})$$

ثابت کنید که

$$\mathbb{E}[||Z - p||_2^2] = \frac{8k}{n^2\epsilon^2} + \frac{1}{n}\sum_{j=1}^{n} p_j(1 - p_j) \leq \frac{8k}{n^2\epsilon^2} + \frac{1}{n}$$

[۱] Private
[۲] Count query

سوال ۶ **Randomized Response** (۱۷ نمره)

پس از برگزاری آزمون در یک کلاس n نفره، میخواهیم مکانیزم حافظ حریم خصوصی طراحی کنیم که میزان تقلب دانشجویان در امتحان را مشخص کند. در این مکانیزم، در ابتدا از دانشجویان سوال میشود که آیا در آزمون تقلب کرده اند یا خیر. فرض کنید پاسخ واقعی دانشجوی iام به این سوال با تابع مشخصه‌ی X_i تعیین میگردد

$$X_i = \begin{cases} 1, & \text{دانشجوی } i\text{ام در این آزمون تقلب کرده است} \\ 0, & \text{دانشجوی } i\text{ام در این آزمون تقلب نکرده است} \end{cases}$$

در مکانیزم حریم خصوصی به جای آنکه مقدار X_i به‌عنوان پاسخ هر دانشجو بازگردانده‌شود، ما از دانشجویان می‌خواهیم کهه با احتمال $\frac{1}{2}+\alpha$ مقدار X_i و با احتمال $\frac{1}{2}-\alpha$ مقدار $1-X_i$ را برگردانند. خروجی این مکانیزم را با Y_i مشخص میکنیم.

$$Y_i = \begin{cases} X_i, & \text{با احتمال } \frac{1}{2}+\alpha \\ 1-X_i, & \text{با احتمال } \frac{1}{2}-\alpha \end{cases}$$

به دانشجویان تکه کاغذی داده شده است که موظف هستند مقدار Y_i را رد آن یادداشت نمایند و درنهایت اطلاعات مورد نظر از کاغذهای یادداشت شده توسط دانشجویان به‌دست می‌آید. در رابطه با این مکانیزم به سوالات زیر پاسخ دهید.

(الف) تخمین‌گری ناریب[۳] برای میزان تقلب با استفاده از Y_iهای بدست آمده از دانشجویان برحسب پارامتر α ارائه دهید.

(ب) میزان دقت تخمینگر و حریم خصوصی دانشجویان چه رابطه ای با پارامتر α دارد؟ حالات $\alpha=0$ و $\alpha=\frac{1}{2}$ را بررسی نمایید.

(ج) برای میزان خطای تخمین با استفاده از نامساوی چبی‌شف[۴] یک کران بالا به‌دست آورید. با استفاده از این کران تحلیل کنید که برای رسیدن به خطای γ به چه تعداد دانشجو نیاز است.

سوال ۷ الگوریتم گرادیان کاهشی تصادفی خصوصی تفاضلی (۳۳ نمره)

با تکمیل دفترچه‌ی differentially private SGD.ipynb الگوریتم را پیاده‌سازی کرده و ارزیابی‌های خواسته شده را انجام دهید.

موفق باشید

[۳] Ubiased
[۴] Chebyshev