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 پنجم

مهلت تحویل: ساعت ۲۳:۵۹ دوشنبه ۱۵ خرداد ۱۴۰۲

نکات و قواعد

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

۲. محل بارگذاری تمرین تا یک هفته پس از مهلت ارسال باز خواهد بود. در طول ترم، در مجموع می‌توانید از ۲۶ روز تاخیر مجاز به صورت ساعتی استفاده کنید و پس از آن به ازای هر روز ۲۰ درصد جریمه بر روی نمره‌ی کسب شده اعمال خواهد شد.

۳. 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.

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_HW۵_StdNum_FirstName_LastName ذخیره کرده و ارسال نمائید.

سوال ۱ مسمومسازی (۱۳ نمره)

(الف) (۳ نمره) با در نظر داشتن،

- $\mathcal{L}_{attack}$: تابع هدف طراحی شده‌ی مهاجم (که باید بیشینه شود)
- $\mathcal{L}_{train}$: تابع هزینه‌ای که فرد مورد تهاجم با کمینه کردن آن مدلش را آموزش می‌دهد
- D_{train}: مجموعه دادگان اولیه‌ی فرد مورد تهاجم
- D_{test}: مجموعه دادگانی که تابع هزینه‌ی مهاجم روی آن اندازه‌گیری می‌گردد.

یافتن مجموعه‌ی بهینه از دادگان D_{poison} را که قرار است به منظور اجرای حمله به D_train اضافه شوند را به صورت یک بهینه‌سازی دو سطحی بنویسید.

(ب) (۱۰ نمره) برای دفاع در برابر حملات مسمومسازی به صورت ضمانت شده، رویکرد خاصی استفاده شده است که در ادامه معرفی می‌شود. مجموعه دادگان آموزش (S) به عنوان ورودی به سیستم داده شده و با کمک تابع h ابتدا به k دسته تقسیم می‌شوند. k مدل دسته‌بند برای مسئله‌ی مورد نظر تعیین می‌شود. هر یک از این k دسته مجموع دادگان آموزش یکی از مدل‌ها را تشکیل می‌دهند. دسته‌بندی دادگان، توابع متناظر و تابع نهایی به صورت زیر هستند:

$$P_i^S := \{s \in S | h(s) \overset{i}{=} 0\}$$
$$f_i^S(x) := \{f(P_i^S, x)\}$$
$$n_c^S := |\{i \in [k] | f_i^S(x) = c\}|$$

Algorithm **Simple Blackbox Attack (SimBA)**

procedure SIMBA($\mathbf{x}, y, Q, \epsilon$)
  $\delta = \mathbf{0}$
  $\mathbf{p} = p_h(y \mid \mathbf{x})$
  while $\mathbf{p}_y = \max_{y'} \mathbf{p}_{y'}$ **do**
    Pick randomly without replacement: $\mathbf{q} \in Q$
    for $\alpha \in \{\epsilon, -\epsilon\}$ **do**
      $\mathbf{p}' = p_h(y \mid \mathbf{x} + \delta + \alpha\mathbf{q})$
      if $\mathbf{p}'_y < \mathbf{p}_y$ **then**
        $\delta = \delta + \alpha\mathbf{q}$
        $\mathbf{p} = \mathbf{p}'$
        break
  return δ

$$g(S, x) := \arg\max_c n_c^S(x)$$

طراح این رویکرد دفاعی مبتنی بر رویکردهای دفاعی که در درس دیده‌اید معتقد است که با این روش نسبت به حملات مسموم‌سازی مقاومت ضمانت شده وجود دارد. اما در بخش‌هایی از این طراحی اشتباهاتی وجود دارد که این نتیجه‌گیری را زیر سوال می‌برد. اشتباهی که در این طراحی وجود دارد و بایستی ویرایش شود را به صورت مشخص و با ذکر دلیل بیان کنید.

سوال ۲ حمله‌ی جعبه‌سیاه (۱۵ نمره)

الگوریتم داده شده، الگوریتم مربوط به یک حمله‌ی ساده‌ی جعبه‌سیاه است، که در آن x یک تصویر ورودی، y برچسب صحیح مربوط به تصویر x ، Q یک مجموعه از بردارهای پایه‌ی رندوم $Orthonormal$ (با اندازه‌ی واحد و دوبه‌دو عمود بر یکدیگر)، ϵ یک اندازه‌ی کنترل‌کننده‌ی میزان تغییرات و $P_h(b|a)$ میزان احتمال دسته‌بندی تصویر ورودی a در کلاس b (برچسب نرم یا $Soft\ label$) می‌باشند.
با توجه به این الگوریتم به سوالات زیر پاسخ دهید:

(الف) (۴ نمره) همانطور که می‌دانیم حملات نمونه‌خصمانه هدفمند هستند و یا بی‌هدف؛ این الگوریتم در کدام دسته قرار می‌گیرد؟ نشان دهید چگونه می‌توان آن را به حالت دیگر تغییر داد.

(ب) (۳ نمره) علت اینکه مجموعه‌ی Q مجموعه‌ای از بردارهای عمود بر یکدیگر است و اینکه مطابق با خط هفتم الگوریتم، بردار q را هر بار بدون جایگذاری انتخاب می‌کنیم، چیست؟

(ج) (۵ نمره) آیا الگوریتم به‌لحاظ تعداد جستارهای (کوئری‌های) ارسالی به مدل هدف (تعداد محاسبات $P_h(b|a)$) بهینه است یا می‌توان بدون تغییر در کارکرد، آن را بهبود بخشید؟ پاسخ خود را با دلیل بیان کرده و در صورت امکان ایجاد تغییر، نحوه‌ی اعمال آن را نشان دهید.

(د) (۳ نمره) برای T تکرار ($Iteration$)، حد بالای میزان آشفتگی به این صورت خواهد بود: $\|\delta\|_2^2 \leq T\epsilon^2$
این حد بالا دقیق است، اگر همه‌ی جستارها منجر به یک گام ϵ یا $-\epsilon$ شوند. رابطه‌ی بالا یک مصالحه ($Tradeoff$) را نشان می‌دهد. آن را تحلیل نمایید.

سوال ۳ علامت گرادیان (۱۲ نمره)

حمله‌ی جعبه‌سیاه را به‌عنوان یافتن جهت با کوچکترین اندازه تا مرز تصمیم‌گیری (کوتاه‌ترین فاصله تا مرز تصمیم‌گیری) درنظر می‌گیریم. به‌طور خاص، برای یک نمونه‌ی x داده شده، برچسب واقعی y و تابع جعبه‌سیاه برچسب سخت ($Hard\ label$) f با ورودی R^d و خروجی یکی از برچسب‌های ۱ تا k برای k کلاس، تابع هدف g را (برای حمله‌ی بدون هدف) می‌توان به‌صورت زیر نوشت:

$$min_\theta\, g(\theta)\, where\, g(\theta) \;=\; argmin_{\lambda>0}\, (f(x \,+\, \lambda\frac{\theta}{\|\theta\|}) \;\neq\; y) \qquad (۱)$$

که در آن θ جهت انتخابی و $g(\theta)$ برابر با فاصله‌ی نقطه‌ی x در جهت θ تا مرز تصمیم‌گیری می‌باشد. نشان داده شده است که این تابع عموما هموار می‌باشد، و تابع هدف g را می‌توان با یک رویه‌ی جست‌وجوی دودویی محلی ارزیابی نمود. همچنین می‌توان مقدار بهینه‌ی آن را به‌عنوان یک تابع

هدف، از طریق یک رویه‌ی بهینه‌سازی، به‌دست آورد. همانطور که می‌دانیم، برای بهینه‌سازی نیاز به گرادیان تابع هدف داریم؛ در اینجا تنها می‌توانیم تخمینی از گرادیان تابع هدف g را با کمک روش تفاوت متناهی ($Finite\ differences$) به‌دست بیاوریم:

$$\nabla g(\theta;\ u)\ \approx\ \frac{g(\theta\ +\ \epsilon u)\ -\ g(\theta)}{\epsilon}u \tag{۲}$$

که در آن u یک بردار رندوم گاوسی و $\epsilon\ >\ 0$ یک پارامتر هموارسازی بسیار کوچک است. اما مشکل اینجاست که هر محاسبه‌ی این تخمین گرادیان، نیاز به تعداد زیادی جستارهای برچسب سخت برای جست‌وجوی دودویی دارد.
اگر از رابطه‌ی زیر به‌جای تخمین مقدار دقیق گرادیان استفاده کنیم:

$$Sign(g(\theta\ +\ \epsilon u)\ -\ g(\theta))\ =\ \begin{cases}+1, & f(x\ +\ g(\theta)\frac{(\theta\ +\ \epsilon u)}{\|\theta\ +\ \epsilon u\|})\ \neq\ y\\ -1, & \text{Otherwise.}\end{cases} \tag{۳}$$

(الف) (۹ نمره) آیا منطق رابطه‌ی بالا جهت محاسبه‌ی علامت تفاوت متناهی درست است؟ با رسم شکل درستی یا نادرستی آن را نشان دهید.

(ب) (۳ نمره) آیا استفاده از علامت گرادیان مانند رابطه‌ی ارائه‌شده، بجای مقدار گرادیان تخمینی، می‌تواند جایگزین مناسب و کافی از نظر حجم اطلاعاتی که ارائه می‌دهد، برای گرادیان تخمینی باشد؟

سوال ۴ پیشنیازها و یافتن گرادیان به کمک بهینه‌سازی (۱۲ نمره)

در زمان تخمین گرادیان در حملات جعبه‌سیاه، معمولا فرض بر این است که گرادیان هدف، یک بردار کاملا ناشناخته می‌باشد. اما همانطور که در اینجا نشان داده شده است، حجم زیادی از دانش پیش زمینه‌ای درباره‌ی گرادیان تابع هدف وجود دارد که می‌توان در زمان تخمین گرادیان از آن بهره برد.

(الف) (۴ نمره) انواع این دانش پیش‌زمینه‌ای را نام برده و هر کدام را به‌طور خلاصه توضیح دهید (توضیح در حد یک یا دو خط برای هر مورد کافیست).

(ب) (۸ نمره) در این مقاله نیز مانند بسیاری از حملات جعبه‌سیاه دیگر، نیاز به گرادیان تابع خطا در نقاط مشخص داریم، اما نحوه‌ی به‌دست آوردن این گرادیان به‌صورت بهینه، بسیار متفاوت است؛ به‌این منظور یک تابع هدف جدید برحسب گرادیان تخمینی g که مقدار صحیح آن جهت بهینه‌سازی تابع خطا مدنظر ما است، تعریف می‌شود، که این تابع هدف جدید در واقع قرینه‌ی ضرب داخلی گرادیان تخمینی g و گرادیان واقعی تابع خطا $\nabla L(x,\ y)$ می‌باشد:

$$l_t(g)\ =\ -\left\langle \nabla L(x,\ y),\ \frac{g}{\|g\|} \right\rangle \tag{۴}$$

که در آن t شماره‌ی گام تکرار می‌باشد. در صورت به‌حداقل رسانی مقدار این تابع هدف، این دو بردار بر یکدیگر منطبق شده و تخمین گرادیان g مقدار مناسب را پیدا خواهد نمود و می‌توانیم در نهایت از این گرادیان تخمینی در روشی مانند حمله‌ی PGD در حالت جعبه‌سیاه بهره ببریم. اما پیدا کردن تخمین مناسب از g نیز خود یک مسئله‌ی بهینه‌سازی می‌باشد که نیاز به محاسبه‌ی گرادیان دارد، که آن را Δ_t می‌نامیم. در واقع Δ_t گرادیان تابعی برحسب گرادیان g است. به کمک الگوریتم ۲ در صفحه‌ی ۸ مقاله، نحوه‌ی محاسبه‌ی Δ_t را توضیح دهید.
(نکته: v در این الگوریتم همان اطلاعات پیشنیازی یا $Prior$ است که در بخش الف به آن پرداخته شد.)

سوال ۵ NES (۲۰ نمره)

دفترچه‌ی NES and RND.ipynb را کامل کنید. در این دفترچه، بخش‌هایی از یک الگوریتم جعبه سیاه به نام NES را کامل خواهید کرد و سپس اثر دفاع نویز تصادفی [۱] بر این حمله را بررسی خواهید نمود.

موفق باشید

[۱] Random Noise Defence (RND)