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: علیرضا سخایی‌راد، رامتین مسلمی، علی اکبری

تمرین سوم

مهلت تحویل: ساعت ۲۳:۵۹ پنج‌شنبه ۱ آذر 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_HW۳_StdNum_FirstName_LastName ذخیره کرده و ارسال نمائید.

سوال ۱ (۲۰ نمره)

فرض کنید برای مسئله دسته بندی با برچسب های $y \in \{-1, 1\}$ از یک مدل دسته‌بند خطی استفاده می‌کنیم. خروجی این مدل به صورت

$$f(x) = sign(w^T x + b)$$

است که در آن w شامل وزن‌های مدل است.

الف)

شعاع تضمین مقاومت این مدل در برابر اغتشاش جمع‌شونده با ورودی را بیابید.

ب)

فرض کنید مشابه با روش هموارسازی تصادفی ورودی‌های x را ابتدا با یک نویز گاوسی به صورت $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ جمع می‌کنیم و سپس به عنوان ورودی به دسته‌بند خطی می‌دهیم. حال تابع زیر را تعریف می‌کنیم:

$$g(x) = argmax_{c \in 0,1} P[f(x + \epsilon) = c]$$

که در آن تابع g محتمل‌ترین کلاس برای حالات مختلف خروجی دسته‌بند هنگام جمع با نویز را نشان می‌دهد. نشان دهید همواره داریم $g(x) = f(x)$.
به عبارت دیگر نشان دهید هرگاه یک نویز گاوسی با میانگین صفر با ورودی یک دسته بند خطی جمع شود، فارغ از پارامتر واریانس آن، همواره محتمل ترین کلاس برای خروجی داده ی جدید، همان خروجی دسته بند به ازای ورودی اصلی است.

سوال ۲ (۲۰ نمره)

همان طور که در درس مطرح شد، با شناسایی نمونه های خصمانه، تلاش برای دفاع دربرابر حملات و مقاوم سازی مدل ها به یک مسئله جدی تبدیل شده است. روش های مختلفی نیز در این سال ها پیشنهاد شده اند. نکته مهم در این میان، ارزیابی جامع و قابل اطمینان موفقیت روش های دفاع پیشنهادی است.

الف)

یک تیم پژوهشی روش جدیدی برای بهبود مقاومت مدل ها نسبت به نمونه های خصمانه به نام *prpd* پیشنهاد کرده اند. در این روش تابع *ReLU* با تابع دیگری جایگزین شده است و ادعا می شود در مقابل حملات مختلف مقاومت پایدار بالاتری دارد. این تیم برای اثبات اثربخشی روش خود عملکرد آن را در مقابل حملات *FGSM*، *PGD-20*، *CW* و همچنین *BB* که یک حمله جعبه سیاه نسبتا ساده مبتنی بر مدل دیگر است، گزارش کرده اند که نتایج آن در جدول زیر نمایش داده شده است.

BB	*CW*	*PGD-20*	*FGSM*	*Clean*	
46.93	*47.12*	*47.25*	*47.62*	*81.40*	*prpd*

نظر شما در باره ارزیابی درستی ادعای موفقیت این روش پیشنهادی چیست؟ آیا در همین نتایج و ارزیابی صورت گرفته، نکته یا نکات سوال برانگیزی وجود دارد که موفقیت این روش را برای شما زیر سوال ببرد؟ تحلیل خود را بیان کنید و همچنین در صورتی که به نظرتان نیاز به بررسی بیشتر وجود دارد، دو راه برای ادامه ارزیابی این روش دفاع پیشنهاد کنید.

ب)

یکی از نتایجی که در کارهای اخیر برای ارزیابی مقاومت روش های پیشنهادی گزارش می شود، دقت خصمانه در مقابل مجموعه ای از حملات موسوم به *AutoAttack* است. مقاله *Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks* مجموعه‌ای از حملات متنوع را در کنار هم معرفی و پیاده سازی کرده است که با توجه به قدرت و اثربخشی آن در کارهای بعدی به عنوان معیار شناخته شده ای برای ارزیابی مقاومت مدل ها استفاده شده است. هر یک از حمله های این مجموعه را با بیان نوع و خصوصیات اصلی آن به صورت کوتاه معرفی کنید.

سوال ۳ (۲۰ نمره)

۱. فرض کنید که می‌خواهید رفتار یک مدل را بررسی کنید که متوجه شوید دفاعی از نوع *Gradients Obfuscating* در آن انجام شده است یا نه، پنج نوع رفتاری که بررسی می‌کنید برای فهم این موضوع و دلایل بررسی این رفتارها را بیان کنید. در هر کدام شرح دهید اعمال این نوع از دفاع‌ها منجر به چه تغییر رفتاری می‌شود. (۶ نمره)

۲. توضیح دهید هر یک از روش‌های *Reparametrization*، *EOT* و *DBPA* بر کدام یک از روش‌های *Gradients Obfuscate* غلبه می‌کنند و به تفصیل نحوه عملکرد هر یک را شرح دهید. (۶ نمره)

۳. با توجه به مقاله *Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples* که در اسلایدهای درس برای این موضوع آورده شده است به این پرسش پاسخ دهید: نحوه عملکرد *l-Level Encoding Thermometer* را شرح دهید و توضیح دهید چگونه می‌توان *Adversarial Training* بر روی این نوع انکودینگ داشت. توضیح دهید در این نوع انکودینگ و در این مقاله چگونه از *BPDA* برای حمله استفاده شده است. (۸ نمره)

سوال ۴ تمرین عملی (۱۰۰ نمره)

موفق باشید.