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 سوم

مهلت تحویل: ساعت ۲۳:۵۹ دوشنبه ۴ اردیبهشت 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 ۲۱ روز تاخیر مجاز به صورت ساعتی استفاده کنید و پس از آن به ازای هر روز ۲۰ درصد جریمه بر روی نمره‌ی کسب شده اعمالط 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_HW۳_StdNum_FirstName_LastName** ذخیره کرده و ارسال نمائید.

سوال ۱ آموزش خصمانه دسته بند خطی (۹ نمره)

میخواهیم یک مسئله دسته‌بندی دوکلاسه را با استفاده از توابع خطی ساده حل کنیم. فضای فرضیه انتخابی ما (H) مجموعه تابع‌های خطی (h) به صورت زیر هستند که با پارامتر W و b مشخص می‌شوند.

$$W \in \mathbb{R}^n,\ b \in \mathbb{R}$$
$$x \in \mathbb{R}^n,\ y \in \{\pm 1\}$$
$$h(x) = W^T x + b$$

یادگیری استاندارد به صورت این مسئله بهینه سازی خواهد بود:

$$\min_{W,b} \mathbb{E}_{(x,y)} \left[loss(h(x), y) \right] \quad (۱)$$

تابع هزینه مورد استفاده ما در این مسئله به صورت زیر است:

$$loss(h(x), y) = \log(1 + exp(-y.h(x)))$$

(الف) یکی از اولین فرضیه های مطرح شده برای چرایی وجود نمونه‌های خصمانه، پیچیدگی و ذات غیرخطی شبکه‌های عصبی بوده است. در این سوال ما از تابع خطی ساده استفاده کرده‌ایم. آیا در شرایطی ممکن است مقدار خصمانه(δ) برای نمونه x وجود داشته باشد که $||\delta||_\infty \leq \epsilon$ و با وجود کوچک بودن مقدار ϵ ، پیش‌بینی تابع h خطی ما را تغییر دهد؟ (در صورت پاسخ منفی، چرایی آن و در صورت پاسخ مثبت، توضیحی برای حالتی که آسیب پذیری وجود دارد بیان کنید.)

(ب) میخواهیم به جای آموزش استاندارد، آموزش خصمانه در نرم بی نهایت با مقدار مجاز ϵ و با حمله FGSM داشته باشیم. فرم ساده سازی شده و نهایی مسئله بهینه سازی متناظر با این آموزش خصمانه در مسئله حاضر را بنویسید.

(ج) اگر یادگیری استاندارد را با منظم ساز نرم ۱ ترکیب کنیم[1] ، تابع بهینه سازی نیز تغییر می‌کند. این حالت را با حالت آموزش خصمانه که در بخش (الف) به دست آوردید از منظر فرم تابع بهینه سازی و همچنین نحوه پیشرفت در فرآیند آموزش مقایسه کنید.

سوال ۲ حمله هدفمند (۶ نمره)

یک تیم مهاجم قصد پیداکردن حمله به یک مدل دسته‌بندی موجود (h_w) با کمک عبارت بهینه‌سازی زیر را دارند:

$$\max_{x'} l(h_w(x'), y) \quad (۲)$$

ولی به پیشنهاد یکی از اعضا برای داشتن حملات قوی‌تر می‌خواهند به سراغ حملات هدفمند (targetd) بروند.
برای این کار بایستی مشخص کنند که اولا کلاس هدف برای حمله را بر چه مبنایی انتخاب کنند و دوما حمله فعلی خود را چگونه به حالت هدفمند تغییر بدهند. شما به عنوان یکی از اعضای فرضی تیم پاسخ خود را به صورت زیر پیشنهاد کنید:

(الف) دو رویکرد برای انتخاب کلاس هدف در حمله پیشنهاد کنید. (در صورتی که از مقاله یا منبع دیگری برای پاسخ خود استفاده کرده اید، آن را ذکر کنید.)

(ب) عبارت بهینه سازی این حمله را به گونه‌ای تغییردهید تا یک حمله هدفمند داشته باشید.

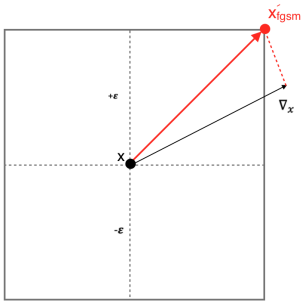
سوال ۳ حملات گوناگون (۱۵ نمره)

نمونه خصمانه برای یک نمونه (x) در حملات مختلف با رابطه‌های مختلف به دست می‌آید. به عنوان نمونه برای حمله FGSM داریم:

$$\delta_{fgsm} = \epsilon \; sign(\nabla_x l(f(x), y))$$

$$x'_{fgsm} = x + \delta_{fgsm}$$

شکل زیر به صورت نمادین نمونه خصمانه استخراج شده توسط FGSM برای یک نمونه دو بعدی را نشان می دهد.



(توجه داشته باشید که در این سوال نرم بی نهایت را برای محدوده مجاز حمله در نظر داریم.)
رابطه مربوط به نمونه خصمانه ساخته شده با حمله PGD و حمله FGSM-RS که در مقاله Fast is better than free را بنویسید. همچنین مشابه شکل به صورت شماتیک استخراج نمونه خصمانه در هر یک از این دو حمله را نمایش بدهید و با در نظرگرفتن آن تفاوتش با FGSM را بیان کنید.

سوال ۴ عملی (۷۰ نمره)

در این بخش به سراغ مجموعه دادگان CIFAR۱۰ می‌روید و به صورت عملی به آموزش استاندارد، حمله به مدل یادگرفته شده و آموزش خصمانه خواهید پرداخت. برای این سوال از معماری مدل Resnet استفاده می کنید. در ادامه مراحل مختلف و نتایجی که بایستی گزارش شوند، بیان شده‌اند. برای حل این سوال و انجام پیاده‌سازی ها به فایل ضمیمه مراجعه کنید.

(الف) ابتدا مدل خود را به صورت استاندارد با مجموعه دادگان آموزش CIFAR۱۰ آموزش بدهید.

(۱) دقت نهایی مدل یادگرفته شده روی مجموعه دادگان آموزش و ارزیابی را گزارش کنید.

(۲) حال حمله FGSM را پیاده سازی کنید و سپس براساس آن، دقت خصمانه مدل یادگرفته شده را در مقابل حمله FGSM به ازای سه مقدار مختلف $\epsilon = 4/255, 8/255, 12/255$ گزارش کنید.

[1] 1\-regularization

(۳) به ازای ۵ نمونه از دادگان ارزیابیCIFAR۱۰، تصویر اصلی نمونه، نویز اضافه شده براساس حمله FGSM و تصویر تغییریافته نهایی با حمله FGSM را نمایش بدهید و خروجی مدل در حالت تمیز و حالت تغییریافته را همراه با برچسب واقعی داده گزارش کنید.

(۴) از بین دادگان ارزیابی، نمونه اول را در نظر بگیرید. بررسی کنید تا مطمئن باشید مدل آن را به درستی دسته‌بندی می‌کند (اگر دسته‌بندی داده اشتباه بود از داده دوم استفاده کنید و به همین ترتیب ادامه دهید). گرادیان تابع هزینه مدل نسبت به این ورودی را محاسبه کنید.
بازه $\epsilon = [-0.5, 0.5]$ را با فواصل 0.01 به صورت ۱۰۱ مقدار در نظر بگیرید و به ازای هر کدام نمونه جدیدی با اضافه کردن $\epsilon\ sign(\nabla_x(l(W, x, y)))$ به نمونه اصلی (x) بسازید و به مدل بدهید.
مقدار logit (بردار ۱۰ تایی ورودی به Softmax) متناظر با هرکدام را به دست آورید و در نموداری این ۱۰۱ مقدار را برحسب مقدار اپسیلون (محور افقی نمودار) نمایش دهید (مشابه نمودار سمت چپ در شکل ۴ مقاله Explaining and harnessing adversarial examples).
حال مشابه این کار را با یک جهت تصادفی انجام دهید. برای این کار با اضافه کردن $\epsilon\ sign(U(-1, +1))$ به نمونه اصلی که $U(-1, 1)$ توزیع یونیفرم در بازه $(1, -1)$ است، نمونه‌های جدید متناظر با ϵ ها را بسازید و نمودار مشابه را در این حالت رسم کنید.
نتیجه‌گیری خود براساس دو نمودار ایجاد شده را شرح دهید.

(ب) در این قسمت مدلی با معماری مشابه قسمت اول را دوباره ساخته و این بار به جای آموزش استاندارد با کمک حمله FGSM ای که پیاده کرده‌اید به صورت خصمانه با $\epsilon = 8/255$ آموزش بدهید.

(۱) نتایج نهایی مدل یادگرفته شده روی مجموعه دادگان آموزش و ارزیابی، شامل دقت استاندارد (بدون حمله) و دقت خصمانه با حمله fGSM به ازای $\epsilon = 8/255$ گزارش کنید.

(۲) حال حمله PGD را پیاده سازی کنید و عملکرد مدل را در مقابل حمله PGD با $\epsilon = 8/255$ به ازای تعداد گام های ۲ و ۴ روی مجموعه دادگان ارزیابی گزارش کنید.

(۳) نتایج نهایی مدل یادگرفته شده روی مجموعه دادگان آموزش و ارزیابی، شامل دقت استاندارد (بدون حمله) و دقت خصمانه با حمله PGD به ازای $\epsilon = 8/255$ گزارش کنید. ارزیابی و جمع بندی خود از نتایج و مقایسه ها را در گزارش بیان کنید.

(۴) در این قسمت عملکرد مدل اول که به صورت استاندارد آموزش داده اید و همینطور مدلی دوم که به صورت خصمانه آموزش داده اید را روی مجموعه دادگان ارزیابی در شرایطی که به نمونه ها نویز گاوسی با میانگین صفر و واریانس $6/255$ اضافه کنید، ارزیابی و گزارش کنید. نتایج این قسمت را با نتایج دو مدل نسبت به نمونه های خصمانه با یکدیگر مقایسه کنید و جمع‌بندی خود را بیان کنید.

(۵) به ازای ۵ نمونه از دادگان ارزیابیCIFAR۱۰، تصویر اصلی نمونه، نویز اضافه شده و تصویر تغییریافته نهایی براساس حمله PGD به ازای تعداد گام های ۴ را نمایش بدهید و خروجی مدل در حالت تمیز و حالت تغییریافته را همراه با برچسب واقعی داده گزارش کنید.
(!نکته: ترجیحا نمونه های یکسانی را در بخش نمایش تصویر در دو بخش (الف) و (ب) برای بررسی انتخاب کنید.)

برای پاسخ قسمت عملی، علاوه بر فایل نوت بوک خود، لازم است گزارش خود (شامل روند کار و مشاهدات نهایی، نتایج خواسته شده در بخش های مختلف و توضیحات لازم دیگر) را در فایل جدا به فرمت pdf ارسال کنید.

موفق باشید