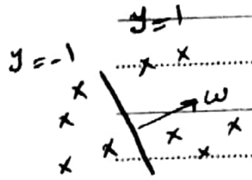
حمیدرضا امیرزاده
۴۰۱۲۰۶۹۹۹

تمارین سری سوم درس SPML

① الف) $h(x) = w^T x + b$ $\quad loss(h(x), y) = \lg(1+e^{-y\,h(x)})$

$x' = x + \delta \quad h(x') - h(x) = w^T\delta$

$\arg\max\ w^T\delta$ s.t. $\|\delta\|_\infty \le \varepsilon$ $\xrightarrow{\text{جواب}}$ $\delta^* = \varepsilon . sign(w)$



بنابراین میزان تغییر $h(x)$ حداکثر برابر $\varepsilon\|w\|_1 = h(x') - h(x) = w^T\delta^*$ می‌باشد.

اگر $y = 1$ باشد آنگاه تغییر x به میزان $-\varepsilon\, sign(w)$ یعنی $x' = x - \varepsilon\, sign(w)$ به میزان $\varepsilon\|w\|_1$ در خروجی $h(x)$ تغییر ایجاد می‌کند که اگر این میزان باعث شود مرز تصمیم‌گیری را رد کنیم، x' یک نمونه خصمانه حساب می‌شود.

اگر $y = -1$ باشد آنگاه تغییر x به میزان $\varepsilon\, sign(w)$ یعنی $x' = x + \varepsilon\, sign(w)$ به میزان $\varepsilon\|w\|_1$ در خروجی $h(x)$ تغییر ایجاد می‌کند که اگر این میزان باعث شود مرز تصمیم‌گیری را رد کنیم x' یک نمونه خصمانه حساب می‌شود.

ب) در حمله FGSM مقدار نویز بهینه برابر $\eta^* = \varepsilon\, sign(\nabla_x f(x))$ است. از آنجایی که در این مسئله داریم $f(x) = w^T x + b$ بنابراین داریم: $\eta = \varepsilon . sign(w)$. از آنجایی که اگر $y = 1$ باشد باید در جهت $-sign(w)$ حرکت کنیم تا نتیجه دسته‌بندی به عوض شدن نزدیک شود، بنابراین تابع هزینه در حالت آموزش خصمانه ما به صورت زیر در می‌آید:

$$\min_{w,b} \frac{1}{2} E_{(x,y)}\left[loss(h(x), y)\right] + \frac{1}{2} E_{(x-\varepsilon\, sign(w),\, y)}\left[loss(h(x), y)\right]$$

$$= \frac{1}{2}\lg\left(1+e^{-y(w^T x + b)}\right) + \frac{1}{2}\lg\left(1+e^{-y(w^T(x-\varepsilon\, sign(w)) + b)}\right)$$

$$= \frac{1}{2}\lg\left(1+e^{-y(w^T x + b)}\right) + \frac{1}{2}\lg\left(1+e^{-y(w^T x + b - \varepsilon\|w\|_1)}\right)$$

البته این ضرایب $\frac{1}{2}$ می‌تواند متفاوت باشد و بر اساس میزان ضریب داده‌های خصمانه یک و داده‌های عادی صفر باشد.

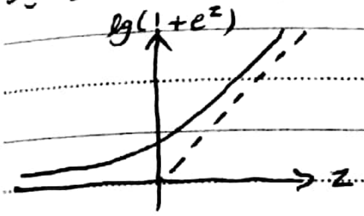
ج) اگر از همان یادگیری استاندارد با استفاده از منظم‌سازی نرم ۱ استفاده کنیم آنگاه تابع هزینه به صورت مقابل است:

$$\min_{w,b} E_{(x,y)}[loss(h(x),y)] + \lambda\|w\|_1$$

$$= lg(1+e^{-y(w^Tx+b)}) + \lambda\|w\|_1$$

با مقایسه تابع هزینه آموزش خصمانه و آموزش استاندارد با منظم‌سازی نرم ۱ می‌توان مشاهده کرد که در حالت آموزش استاندارد، هزینه اضافی $\lambda\|w\|_1$ جداگانه به تابع هزینه اضافه می‌کند و اهمیتی نمی‌دهد که اطمینان مدل در پیش‌بینی داده چقدر است.

اما در حالت آموزش خصمانه، هزینه اضافی $\epsilon\|w\|_1$ درون تابع softplus قرار دارد و ثابت می‌کند که اگر مدل با اطمینان خوبی داده را پیش‌بینی می‌کند، هزینه اضافی ناچیزی به تابع هزینه اعمال می‌شود. هرچه اطمینان مدل در پیش‌بینی کمتر و کمتر و به سمت پیش‌بینی غلط می‌رود، تأثیر این هزینه اضافی بیشتر و بیشتر می‌شود. این رفتار با مشاهده تابع softplus واضح است.



حال می‌توان گفت استفاده از آموزش استاندارد به علت تحمیل یک هزینه ثابت به مدل، عملکرد و [illegible] آن را محدود می‌کند، باعث افت عملکرد مدل می‌شود. اما استفاده از آموزش خصمانه به علت تحمیل هزینه اضافی به جا و مناسب به مدل، می‌تواند عملکرد بهتری نسبت به حالت آموزش استاندارد با منظم‌سازی نرم ۱ داشته باشد.

۲۲

سه‌شنبه
Tuesday
11 Apr 2023
۲۰ رمضان ۱۴۴۴
فروردین

② الف) طبق پیشنهاد مقاله Towards evaluating the Robustness of NN سه رویکرد زیر را می‌توان برای انتخاب کلاس هدف در پیش گرفت:

۱- حالت میانگین (Average): کلاس دارای کلاس‌هایی که کلاس صحیح داره مدنظر نیستند به صورت تصادفی با احتمال مساوی انتخاب کنیم.

۲- حالت بهترین (Best): حمله را برای تمام دسته‌های غیرصحیح انجام داده و پس دسته‌ای را انتخاب کنیم که ساده‌ترین یا کم‌ترین میزان perturbation را لازم داشت.

۳- حالت بدترین (Worst): حمله را برای تمام دسته‌های غیرصحیح انجام داده و پس دسته‌ای را انتخاب کنیم که سخت‌ترین یا بیشترین میزان Perturbation را لازم داشت.

ب) برای اینکه یک حمله صدمه‌زننده‌تر باشیم باید سعی در کم کردن هزینه دسته‌بندی داده با در نظر گرفتن دسته درست برای یا دسته هدف کنیم. عبارت بهینه‌سازی به صورت زیر می‌شود:

$$\min_{x'} loss(h_\theta(x'), t)$$

(t: دسته هدف)

---

③ رابطه بازگشتی به‌روزرسانی ساخته شده با حمله PGD به صورت زیر است:

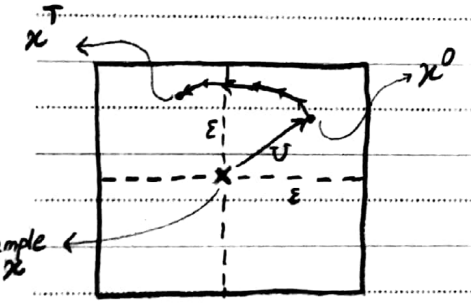
for i = 0 to T:

$$\|\delta\|_\infty \leq \varepsilon$$

$$x' = clip_{[0,1]}\{x' + U(-\varepsilon, \varepsilon)\}$$

$$\delta' = \alpha \cdot sign(\nabla_x loss(\theta, x', y))$$

$$x'^{i+1} = clip_{[\max(0, x-\varepsilon), \min(1, x+\varepsilon)]}\{x' + \delta'\}$$



PGD برخلاف FGSM روشی iterative است. در روش FGSM با یک تکرار قصد می‌توان به یکی از گوشه‌های مجموعه $\|\delta\|_\infty = \varepsilon$ برای نمونه‌خواهد رسید (با در نظر گرفتن x به عنوان مبدا) اما در روش PGD با تغییر روش گام‌به‌گام به دنبال یافتن نمونه‌خواهد ان هستیم که تابع هزینه به ازای آن ماکسیمم شده، همانطور که در شکل مشاهده می‌شود فضای جستجو مجموعه $\|\delta\|_\infty \leq \varepsilon$ است.

نکته قدر: اگر منظر سوال حمله PGD معرفی شده در مقاله باشد، تفاوت آن با این روش گفته شده این است که شروع تصادفی ندارد.

۲۳
چهارشنبه
Wednesday
12 Apr 2023
۲۳ رمضان ۱۴۴۴
فروردین

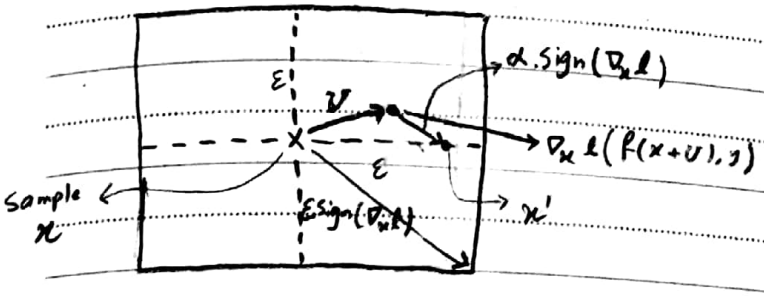
رابطه مربوط به نمونه خصمانه ساخته شده با حمله FGSM_RS به صورت زیر است:

$$\delta = U(-\varepsilon, \varepsilon)$$

$$\delta = \delta + \alpha \cdot sign\left(\nabla_x loss\left(f(x+\delta), y\right)\right)$$

$$\delta = \max\left(\min(\delta, \varepsilon), -\varepsilon\right)$$

$$x' = x + \delta$$



تفاوت روش FGSM_RS با روش FGSM در
نوع مقادیر آن است. این نوع مقادیر
برگشت پذیر است (sign(∇x l)) با ضریب α
عملاً کمک می‌کند که بتوانیم به تمام نقاط
ε همسایگی نمونه سالم نمونه خصمانه بسازیم
داشته باشیم. بنابراین عملکرد این روش
در حد عملکرد روش پیشنهادی PGD می‌باشد.