به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس هوش مصنوعی قابل اعتماد

مدرس: دکتر مصطفی توسلی‌پور

تمرین شماره ۴

خرداد ماه ۱۴۰۳

فهرست

سوال اول : *SECURITY*

یک حمله backdoor شامل قرار دادن یک trigger در داده‌های آموزش می‌شود، به طوری که مدل یاد می‌گیرد تا در صورت مشاهده این trigger در داده ورودی، خروجی خاصی (اشتباه) را نتیجه دهد. همچنین در صورت عدم وجود trigger ، مدل عملکرد عادی خود را دارد که این امر شناسایی این حمله را دشوار می‌کند.

در این قسمت یک مدل ساده کانولوشنی که بر اساس مجموعه داده‌ی MNIST آموزش دیده و مورد حمله backdoor نیز قرار گرفته شده است، در اختیار شما قرار داده‌ایم. از شما می‌خواهیم تا با استفاده از روش‌های ارائه شده در مقاله Neural Cleanse[1] ، ابتدا trigger را شناسایی و بازیابی کرده و سپس مدل را پاکسازی کنید.

بخش اول – شناسایی TRIGGER

در این مقاله، دو فرض اساسی در نظر گرفته شده‌است:

1. یک اینکه ما به مدل مورد حمله قرار گرفته، دسترسی داریم
2. تعدادی از داده‌های سالم نیز در دسترس ما قرار دارند.

لذا معماری مدل آموزش‌یافته در ذیل قابل مشاهده می‌باشد. شما باید پس از ساخت مدل، وزن‌های مدل مرتبط با خود را با توجه به رقم آخر شماره دانشجویی خود، از میان وزن‌هایی که در اختیارتان قرار داده‌ایم، انتخاب کرده و آن‌ها را بارگذاری کنید. همچنین برای داده‌های سالم می‌توانید از داده‌های آزمون[1] مجموعه داده MNIST استفاده کنید.

[1] Test

```
                    Model Structure
(conv1): Sequential(
    (0): Conv2d(1, 16, kernel_size=(5, 5), stride=(1, 1))
    (1): ReLU()
    (2): AvgPool2d(kernel_size=2, stride=2, padding=0)
  )
  (conv2): Sequential(
    (0): Conv2d(16, 32, kernel_size=(5, 5), stride=(1, 1))
    (1): ReLU()
    (2): AvgPool2d(kernel_size=2, stride=2, padding=0)
  )
  (fc1): Sequential(
    (0): Linear(in_features=512, out_features=512, bias=True)
    (1): ReLU()
  )
  (fc2): Sequential(
    (0): Linear(in_features=512, out_features=10, bias=True)
    (1): Softmax(dim=-1)
  )
  (dropout): Dropout(p=0.5, inplace=False)
```

زیربخش اول – بازسازی TRIGGER به صورت مهندسی معکوس

در مقاله معرفی شده، با حل یک مسئله بهینه‌سازی[1]، سعی بر بازسازی trigger به صورت مهندسی معکوس برای تمامی برچسب‌ها شده‌است. ابتدا دو ترم موجود در تابع بهینه‌سازی مورد نظر را توضیح داده **(4 نمره)** و سپس با پیاده‌سازی این قسمت، به ازای تمام برچسب‌ها، trigger مرتبط به هر کدام را بازسازی کرده و نمایش دهید. **(13 نمره)**

(راهنمایی: می‌توانید از کتابخانه‌هایی مانند Pytorch برای حل مسئله بهینه‌سازی استفاده کنید. دو پارامتر pattern و mask را به صورت یک کاناله در نظر گرفته و با عدد 1 مقداردهی اولیه کنید. همچنین برای محدود کردن مقادیر آنها در بازه (0, 1) می‌توانید از توابع فعالسازی مانند sigmoid یا tanh یا ... استفاده کنید.)

[1] Optimization

زیربخش دوم – شناسایی برچسب مورد حمله قرار گرفته

در این بخش، می‌خواهیم تا با استفاده از یک روش شناسایی داده‌ی پرت بر اساس مقادیر MAD[1]، برچسب مورد حمله قرار گرفته را شناسایی کنیم. ابتدا به طور مختصر دلیل استفاده از این روش را توضیح داده **(3 نمره)** و سپس با پیاده‌سازی این بخش از مقاله، برچسب مورد حمله قرار گرفته را گزارش کرده و trigger مرتبط به آن را نمایش دهید. **(6 نمره)**

بخش دوم – پاکسازی مدل و کاهش اثر حمله

در مقاله معرفی شده، پس از شناسایی و بازیابی trigger، برای کاهش اثر حمله سه روش متفاوت معرفی شده‌است. ابتدا هر کدام از این سه روش را به طور مختصر توضیح دهید.**(6 نمره)** در ادامه روش سوم یعنی Unlearning را با استفاده از trigger بازیابی شده پیاده‌سازی کرده، سپس دقت مدل و درصد حملات موفق را برای هر دو مدل مورد حمله قرار گرفته و پاکسازی شده بروی داده‌های آزمون گزارش دهید. **(8 نمره)**

(راهنمایی: 1- برای پاکسازی مدل تنها نیاز به یک دوره[2] آموزش مدل دارید. برای آماده‌سازی دیتاست از داده‌های آزمون MNIST استفاده کرده و trigger بازسازی شده را مطابق مقاله به 20 درصد آن‌ها اعمال کنید. توجه کنید که برچسب تمام داده‌ها درست باشد. 2- در صورتی که در بخش اول موفق به بازیابی trigger نشده‌اید، می‌توانید از trigger اصلی برای بخش دوم استفاده کنید. اطلاعات مربوط به trigger هر یک از مدل‌ها داخل فایل "triggers_information" مهیا شده‌است)

[1] Median Absolute Deviation
[2] Epoch

سوال دوم : *PRIVACY*

بخش اول

سیستمی را در فرض کنید که از یک مدل $\varepsilon - Differentiallyprivate$ با مکانیسم لاپلاس برای محافظت از میانگین و کل درآمد در یک جامعه استفاده می کند و دو درخواست زیر را در نظر بگیرید.

درخواست یک: متوسط درآمد جامعه چقدر است؟

درخواست دو: درآمد کل جامعه چقدر است؟

این جامعه شامل 500 نفر است. حساسیت مربوط به درخواست شماره یک ، 5000 دلار و حساسیت مربوط به درخواست دوم ، 50000 دلار می باشد. (privacy guarantee) $\varepsilon = 0.1$ در نظر بگیرید.

زیربخش اول

مقدار b در توزیع لاپلاس را برای هر درخواست محاسبه کنید. **(3 نمره)**

زیربخش دوم

اگر میانگین درآمد واقعی جامعه 40 هزار دلار و کل درآمد واقعی جامعه 20 میلیون دلار باشد و شما یک نویز 2000 دلاری را برای درخواست مربوط به درآمد متوسط و یک نویز 5000 دلاری را برای درخواست درآمد کل از توزیع های لاپلاس مربوطه به هرکدام نمونه برداری کنید، مقادیر Privacy-preserving گزارش شده برای میانگین و درآمد کل را حساب کنید. **(5 نمره)**

زیربخش سوم

حال، فرض کنید هر دو درخواست را به داده ها اعمال می کنیم، که به عنوان composition of differential privacy شناخته می شود. اگر بخواهیم overall privacy loss را کمتر از $\varepsilon = 0.1$ نگه داریم، باید privacy budget را برای هر درخواست تنظیم کنیم. با اختصاص دادن $\varepsilon 1 = 0.05$ را به درخواست مربوط به درآمد متوسط و $\varepsilon 2 = 0.05$ را به درخواست درآمد کل ، اثر این تغییر بر پارامتر scale در توزیع لاپلاس مربوط به هر درخواست و میانگین و درآمد کل گزارش شده را توضیح دهید. **(7 نمره)**

بخش دوم

پایگاه داده D متشکل از n نود، و یک مجموعه درخواست $Q = q_1, q_2, \dots q_k$ که باید بر روی این پایگاه داده اعمال شود را در نظر بگیرید. هر درخواست به صورت یک counting query بوده و پاسخ ها به صورت یک عدد صحیح در بازه $[0, n]$ می باشند. در این سوال ما قصد داریم تا برای هر درخواست از یک مدل $(\varepsilon, \delta) - Differentially\ private$ با مکانیزم لاپلاس که در آن $\delta = 10^{-5}$ و $\varepsilon = 0.1$ می باشند استفاده کنیم.

با فرض اینکه $\Delta f = 1$ $(sensitivity\ of\ each\ query)$ است و با در نظرگرفتن یک توزیع لاپلاس برای نویز با توزیع چگالی احتمال زیر

$$f(x|b) = \frac{1}{2b} exp\left(\frac{-|x|}{b}\right)$$
$$b = \frac{\Delta f}{\varepsilon}$$

زیربخش اول

مقدار b، ضریب scale مربوط به توزیع لاپلاس را محاسبه کنید. **(3 نمره)**

زیربخش دوم

اگر پاسخ دقیق برای درخواست $q_i = 500$ باشد، احتمال اینکه پاسخ نویزی بیشتراز 505 باشد را محاسبه کنید. (راهنمایی: از تابع توزیع تجمعی (CDF) توزیع لاپلاس استفاده کنید) **(6 نمره)**

زیربخش سوم

حال سناریویی را در نظر بگیرید که در آن درخواست ها را به‌طور متوالی اعمال می شوند و پارامترهای privacy برای هر درخواست i به صورت $(\varepsilon_i, \delta_i)$ میباشد، که در آن $\varepsilon_i = \frac{\varepsilon}{k}$ و $\delta_i = \frac{\delta}{k}$ هستند. مقدار b و احتمال اینکه پاسخ نویزی برای q_i بزرگتر از 505 باشد را دوباره محاسبه کنید. **(7 نمره)**

زیربخش چهارم

در Unbounded Differential Privacy، اگر بدانیم که درصد معین p (بیان شده به صورت اعشاری) از کل جمعیت از پایگاه داده اضافه یا حذف شده است، حساسیت Δf را به گونه ای عوض کنید تا این تغییر منعکس شود و مقدارb و احتمال اینکه پاسخ نویزی بیشتر از505 باشد را دوباره محاسبه کنید. **(9 نمره)**

سوال سوم : *FAIRNESS*

در این تمرین قصد داریم مبحث عدالت [1]را بررسی نماییم. همان‌طور که می‌دانید برای استفاده از مدل‌های یادگیری ماشین برای تصمیم‌گیری، نیاز می‌باشد مدل به صورت عادل [2]عمل نموده و می‌بایست فاقد سوگیری خاصی نسبت به برخی ویژگی‌های حساس باشد. در این تمرین به بررسی عادل بودن یک مدل طبقه‌بند می‌پردازیم.

شما یک دیتاساینتیست در یک شرکت مطرح می‌باشید، در پایان سال دیتای کارمندان شرکت در اختیار شما قرار داده شده‌است و شما می‌خواهید یک طبقه‌بند شامل دو کلاس طراحی نمایید که فارغ از بحث جنسیت افراد، پیش بینی نماید که فرد مورد نظر دستمزد بالاتر از 50K و یا پایین‌تر از 50K دریافت می‌نماید.

نکته‌ای که در طراحی این مسئله از اهمیت زیادی برخوردار می‌باشد، توجه به این است که علاوه بر دقیق بودن مدل در پیش بینی، عادل بودن مدل و نگاه برابر به زنان و مردان در بحث دریافت دستمزد از اهمیت بالایی برخوردار می‌باشد.

بخش اول: دیتا و ارزیابی مدل

بارگذاری داده

فایل دیتاست "data.csv" پیوست شده به همراه تمرین را بارگذاری نمایید.

ارزیابی مدل

ارزیابی این مدل از دو طریق صورت می‌پذیرد، مانند تمام مدل‌های طبقه بند نیاز می‌باشد اطمینان حاصل نماییم که مدل به خوبی آموزش دیده و می‌تواند به درستی پیش‌بینی انجام دهد. می‌توانید از متریک‌های مانند دقت برای سنجش این بخش استفاده نمایید.

از جهتی دیگر می بایست اطمینان حاصل شود که مدل به صورت عادلانه برای دو جنسیت زن و مرد عمل می‌نماید. برای سنجش عادل بودن مدل از دو متریک Zemel Fairness و Disparate Impact استفاده

[1] Fairness
[2] Fair

می‌نماییم. آن‌هارا پیاده سازی نموده و در مراحل بعد برای سنجش مدل خود از آن‌ها استفاده نمایید.
(5نمره)

C : خروجی طبقه بند

S : کلاس مربوط به زنان

$\bar{S}$: کلاس مربوط به مردان

+ : درآمد بالای 50k

$$\text{Zemel Fairness} = \text{prob}\,(C = +|S = \bar{s}\,) - \text{prob}(C = +|S = s)$$

$$\text{Disparate Impact} = \frac{prob(c= +|S=s)}{prob(C= +|S=\bar{s}\,)}$$

بخش دوم: پیاده سازی مدل پایه[1]

ابتدا دیتاست خود را به 70٪ برای آموزش و 30٪ برای تست تقسیم کنید. سپس یک طبقه بند طراحی نمایید تا بتواند میزان درآمد افراد را پیش‌بینی نماید. پس از آموزش و گزارش ارزیابی مدل خود به سوالات زیر پاسخ دهید. **(10 نمره)**

- آیا مدل به خوبی توانسته است درآمد را پیش‌بینی نماید؟
- آیا مدل برای زنان و مردان به صورت عادل عمل می‌نماید؟
- به نظر شما، حذف ویژگی حساس از دیتاست می‌تواند در عادل کردن مدل موثر باشد؟

[1] Base

بخش سوم: پیاده سازی مدل عادل

در این بخش برای آموزش یک مدل عادل، از روشی برای از بین بردن سوگیری [1]با توجه به ویژگی حساس در دیتاست استفاده می‌نماییم. مدل خود را پس از طی نمودن مراحل زیر و به دست آوردن دیتاست جدید آموزش دهید. . **(20 نمره)**

- ابتدا از مدل مرحله قبل استفاده نموده و خروجی مدل را به دیتاست خود اضافه نمایید.
- سپس به ازای هر فرد احتمال هر خروجی را محاسبه نموده و ماکسیمم آن را به دیتاست اضافه نمایید.
- سپس دیتاست را به دوقسمت افرادی که promotion (CP) دریافت می‌کنند و افرادی که demotion (CD) دریافت می‌کنند تقسیم کنید.
  - CP: مردانی که درآمد بالا50k دارند، بر اساس احتمال صودی مرتب [2]نمایید.
  - CD: زنانی که درآمد زیر 50k دارند، بر اساس احتمال نزولی مرتب نمایید.
- کلاس‌های n ردیف اول هر دسته را با یکدیگر جابه‌جا نمایید.

$$n = \frac{(Ss \times S\bar{s}+)-(S\bar{s}\times Ss+)}{Ss+S\bar{s}}$$

Ss: تعداد زنان

$S\bar{s}$: تعداد مردان

Ss+: تعداد زنانی که مدل اولیه پیش بینی درآمد بالای 50k برای آن ها داشته است.

$S\bar{s}$ +: تعداد مردانی که مدل اولیه پیش بینی درآمد بالای 50k برای آن ها داشته است.

- در نهایت دو ستون پیش بینی مدل اولیه و احتمالات را از دیتاست خود حذف نمایید.
- مدل خود را با دیتای جدید به دست آمده آموزش دهید و نتایج را گزارش کنید و تحلیل کنید.

---

[1] Bias
[2] Sort

بخش چهارم: مقایسه و نتیجه گیری

در این قسمت نتایج به دست آمده از دو مدل بالا را در یک جدول گزارش داده و به سوالات زیر پاسخ دهید: **(5 نمره)**

- کدام مدل از دقت بالاتری برخوردار است؟
- کدام مدل عادل می باشد؟
- آیا ارتباطی بین دقت و عادل بودن مدل مشاهده می‌نمایید؟ توضیح دهید.
- یک روش دیگر برای عادل کردن طبقه‌بند معرفی کنید . تحلیل نمایید چرا این روش را معرفی نموده‌اید و به چه علت آن را موثر می‌دانید.
- یک روش دیگر برای عادل کردن طبقه‌بند معرفی کنید و تحلیل نمایید چرا این روش را معرفی نموده‌اید و به چه علت آن را موثر می‌دانید.

بخش پنجم: امتیازی

روشی که در آخرین بخش قسمت قبل معرفی نموده را پیاده‌سازی نموده و نتایج خود را گزارش دهید. نتایج به دست آمده را با مدل‌های قبلی مقایسه نمایید. **(10 نمره)**

مراجع

[1] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In Proceedings of 40th IEEE Symposium on Security and Privacy, 2019

نکات تحویل

- مهلت ارسال این تمرین تا پایان روز "جمعه 15 تیر ماه" خواهد بود.
- در نظر داشته باشید که در این تمرین امکان استفاده از گریس وجود نداشته و پس از اتمام مهلت تحویل، سامانه بسته خواهد شد.
- پیاده سازی با زبان برنامه نویسی پایتون باید باشد و کدهای شما باید قابل اجرا بوده و به همراه گزارش آپلود شوند.
- انجام این تمرین به صورت یک نفره می‌باشد.
- در صورت مشاهده هر گونه تشابه در گزارش کار یا کدهای پیاده‌سازی، این امر به منزله تقلب برای طرفین در نظر گرفته خواهد شد.
- استفاده از کدهای آماده بدون ذکر منبع و بدون تغییر به منزله تقلب خواهد بود و نمره تمرین شما **صفر** در نظر گرفته می‌شود
- در صورت رعایت نکردن فرمت گزارش کار نمره گزارش به شما تعلق نخواهد گرفت.
- تحویل تمرین به صورت **دستنویس** قابل پذیرش نیست.
- تمامی تصاویر و جداول مورد استفاده در گزارش کار باید دارای توضیح (caption) و شماره باشند.
- بخش زیادی از نمره شما مربوط به گزارش کار و روند حل مسئله است.
- لطفا گزارش ، فایل کدها و سایر ضمائم مورد نیاز را با فرمت زیر در سامانه بارگذاری نمائید.
- HW4 _[Lastname]_[StudentNumber].zip
- در صورت وجود سوال و یا ابهام میتوانید از طریق رایانامه زیر با موضوع TAI_HW4 با دستیاران آموزشی در ارتباط باشید:
  - سوال اول: مهیار ملکی mahyar.maleki@ut.ac.ir
  - سوال دوم: مهدی دهشیری mhdhshri@gmail.com
  - سوال سوم: فرزانه حاتمی نژاد farzaneh.hatami@ut.ac.ir

با آرزوی سلامتی و موفقیت روزافزون.