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: رامتین مسلمی، علی اکبری، متین علی‌نژاد

تمرین چهارم

مهلت تحویل: ساعت ۲۳:۵۹ یکشنبه ۲۵ 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_HW۴_StdNum_FirstName_LastName ذخیره کرده و ارسال نمائید.

سوال ۱ بندیت! (۲۰ نمره)

در این سوال می خواهیم به صورت گام به گام به بررسی این مقاله بپردازیم. این مقاله با استفاده از دانش پیشین در خصوص گرادیان تابع هزینه، راهکاری برای کاهش تعداد *query* برای حملات جعبه سیاه ارائه می دهد.

۱. این مقاله بیان می کند تخمین گر ارائه شده برای گرادیان در روش *NES* میان تخمین گر های نااریب با تعداد *query* های کم، کمترین واریانس را دارد. به صورت مختصر توضیح دهید این نتیجه چگونه حاصل می شود (نیازی به ذکر روابط ریاضی نیست و توضیح کفایت می کند).

۲. طبق مقاله، می توان با استفاده از دانش پیشین در خصوص گرادیان تابع هزینه نسبت به ورودی، تعداد *query* ها را حتی نسبت به *NES* کاهش داد. انواع دانش های پیشینی را که می توانیم در خصوص گرادیان نسب به یک داده داشته باشیم توضیح دهید.

۳. مقاله مسئله ی تخمین گرادیان در هر گام از یک روش تکرار شونده همراه دانش پیشین را به یک مسئله ی *bandit* تبدیل می کند. توضیح دهید که مسئله ی *bandit* به طور کلی چیست. سپس بیان کنید در روش ارائه شده دانش پیشین، خروجی و تابع هزینه چیست.

۴. الگوریتم به دست آوردن خروجی بر اساس ورودی در حملات l_2 را به صورت دقیق توضیح دهید (الگوریتم ۲ مقاله و روابط بعد از آن).

سوال ۲ انتقال پذیری (۱۰ نمره)

پیش از این در درس با روش های حمله ی جعبه سیاه مبتنی بر انتقال حملات خصمانه آشنا شدیم. در این سوال قصد داریم به بررسی دقیق تر انتقال پذیری حملات خصمانه بین مدل های مختلف بپردازیم. بر اساس این مقاله به سوالات زیر پاسخ دهید:

۱. سه عامل موثر در انتقال پذیری حملات خصمانه را بیان کنید.

۲. فرض کنید یک داده ی خصمانه ی هدفمند با یک روش مبتنی بر گرادیان برای یک مدل به دست آمده است. به نظر شما آیا مدل هدف نیز این داده را در کلاس مشابه مدل اولیه دسته بندی خواهد کرد؟ توضیح دهید.

سوال ۳ مدل های مولد (۲۰ نمره)

در این سوال با حمله به مدل های مولد آشنا خواهید شد. طبق این مقاله به سوالات زیر پاسخ دهید:

۱. در این مقاله مدل های تهدید چه مواردی هستند؟ هر کدام را به اختصار توضیح دهید.

۲. نویسندگان برای دفاع در برابر این حملات چه مواردی را پیشنهاد می دهند؟ به اختصار توضیح دهید.

۳. با توجه به بخش سوم مقاله، دانش پیشین حمله گر چه مواردی است؟

سوال ۴ تمرین عملی (۱۵۰ نمره)

نوت‌بوک‌های *SPML_PHW4_Extraction.ipynb* و *SPML_PHW4_Poisoning.ipynb* را تکمیل کنید.

موفق باشید.