مبانی یادگیری ماشین

نیم‌سال دوم ۰۲-۰۱

دانشکده‌ی مهندسی کامپیوتر

مدرس: دکتر حامد ملک

پروژه

موعد تحویل: ۱۴۰۲/۰۳/۳۰

پروژه‌ی ۱. تشخیص ناهنجاری

تشخیص ناهنجاری‌ها در داده‌ها در حوزه‌های متعددی مورد توجه است. به عنوان مثال بانک ها علاقه‌مند به کشف تقلب در برنامه‌ها و تراکنش‌ها؛ تحلیلگران امنیت سایبری علاقه‌مند به تشخیص رخدادهای غیرمنتظره مربوط به الگوهای ترافیک در گزارشات[1] سیستم؛ مدیران سیستم‌ها علاقه‌مند به رفتارهای غیرمنتظره در سیستم‌ها و برنامه های کاربردی مستقرشده هستند.

بررسی حملات محروم‌سازی از سرویس[2]

تشخیص ناهنجاری نقش مهمی در زمینه امنیت شبکه دارد. در روش‌های قدیمی تشخیص ناهنجاری نفوذ به سیستم[3] و حملات محروم‌سازی از سرویس برنامه‌نویس‌ها بصورت دستی با جستجوی کلمات کلیدی و مطابقت عبارت منظم متکی بودند. اگرچه برنامه‌نویس‌ها می‌توانند از سیستم‌های تشخیص نفوذ برای کاهش حجم کاری خود استفاده کنند، با این حال داده‌های گزارش سیستم بسیار زیاد، حملات متنوع و مهارت‌های هک در حال بهبود است، که باعث می‌شود روش‌های تشخیص قدیمی به اندازه کافی کارآمد نباشند.

در این پروژه، از شما خواسته می‌شود رفتارهای ناهنجار را در داده‌های یک سیستم دوربین امنیتی شناسایی کنید. از آنجایی که هیچ برچسبی به شما داده نمی شود که نشان دهنده رفتار ناهنجار باشد، باید یک راه حل تشخیص مبتنی بر ناهنجاری با استفاده از روش‌های بدون نظارت[4] با تمرکز بر تجزیه و تحلیل ردپای کاربر ارائه دهید.

باید اشاره کنیم که نگاه کردن به این مشکل به عنوان یک کار یادگیری ماشینی نظارت شده، یک اشکال اساسی دارد. فقط می تواند الگوهای شناخته شده ای را که قبلاً در یک مجموعه داده توضیح داده شده اند را شناسایی کند. بنابراین رویکرد صحیح، ساخت مدلی از رفتار عادی سیستم و سپس جستجوی انحرافات در داده های گزارشات خواهد بود. روش‌های بدون نظارت در این بخش باید به دور از داده‌های ناهنجار در مجموعه آموزشی خود باشند تا خطر سوگیری در انتخاب الگوهای شناخته‌شده را به حداقل برسانیم.

پیاده‌سازی

مراحل اصلی پروژه شامل پیش‌پردازش داده‌ها[5]، استخراج ویژگی‌ها[6] و اعمال روش‌های یادگیری ماشینی بدون نظارت و ارزیابی عملکرد مدل‌ها است. هدف اصلی پروژه ارزیابی و مقایسه کارایی الگوریتم‌های یادگیری ماشینی در تشخیص ناهنجاری است. همچنین، در صورت امکان، پیشنهاد راهکارهای بهبود عملکرد و تعمیم قابلیت این روش‌ها به سیستم‌های واقعی را در نظر خواهیم داشت.

[1] Logs
[2] DoS attacks
[3] Intrusion
[4] Unsupervised
[5] Data pre-processing
[6] Feature extraction

پروژه‌ی ۲. پیش‌بینی قیمت لپ‌تاپ

در این پروژه بایستی قیمت لپ‌تاپ‌های موجود در سایت ترب را پیش‌بینی کنید. برای این منظور با استفاده از ابزارهای خزنده وب، داده‌های مدنظر را جمع‌آوری کرده و با دو مدل یادگیری ماشین اعم از رگرسیون خطی، xgboost، شبکه‌های عصبی و ... مدل خود را آموزش دهید.

در بخش جمع آوری دیتا باید داده‌های مربوط به لپ‌تاپ را به همراه قیمت آن جمع‌آوری کنید. پیشنهاد می‌شود که از ابزارهای خزنده وب مانند selenium استفاده کنید. تعداد داده‌ها بایستی حداقل ۱۰۰۰ عدد باشد.

بخش پیش‌پردازش داده یکی از مهم ترین گام های حل مسئله خواهد بود، زیرا داده‌ای که جمع‌آوری می کنید به احتمال زیاد داده قابل استفاده‌ای نخواهد بود. در این قسمت شما باید با تکنیک‌های مهندسی ویژگی که پیش‌تر آموختید، مجموعه داده‌های جمع‌آوری شده را تمیز کنید. برای قیمت هر محصول از ارزان‌ترین قیمت موجود استفاده کنید.

در نهایت در بخش ساخت مدل و ارزیابی نتایج، شما باید مدل خود را بسازید. برای ساخت مدل مجاز به انتخاب هر مدل در حوزه یادگیری ماشین و نه در حوزه یادگیری عمیق هستید. پس از ساخت مدل شما باید نتایج مدل را ارزیابی کنید و مصورسازی کنید. این ارزیابی شامل معیار خطا (مثلاً خطای میانگین مربعات) و نمودار کاهش خطای آموزش و اعتبارسنجی می‌باشد. همچنین برای ارزیابی مدل از اعتبارسنجی متقابل k-fold با مقادیر kهای مختلف استفاده کنید.