به نام ایزد یکتا

تمرین اول درس روش پژوهش و ارائه

استاد: دکتر رضا صفابخش

تهیه کننده: بردیا اردکانیان

برسی و مهار مثال‌های خصمانه در یادگیری ماشین

• بیان مسئله

در سال‌های گذشته، هوش مصنوعی[1] پیشرفت‌های بسیاری به خود دیده است. در حال حاضر، یادگیری ماشین[2] برای حل بسیاری از مسائل چالش برانگیز مانند ماشین‌های خودران، ترجمه خودکار زبان، تشخیص کلاه‌برداری به کار گرفته می‌شود. گستردگی یادگیری ماشین به قدری ادامه پیدا کرده است که شماری از بسترهای شناخته شده همچون آمازون و گوگل، یادگیری ماش را به عنوان یک سرویس به توسعه دهندگان ارائه می‌کنند.

یادگیری ماشین متخاصم[3] یک تکنیک یادگیری ماشین است که تلاش می‌کند با بهره‌گیری از اطلاعات قابل دستیابی مدل و استفاده از آن برای ایجاد حملات مخرب، از مدل‌ها سوء استفاده کند. بیشتر تکنیک‌های یادگیری ماشین برای کار بر روی مجموعه‌ای از مشکلات خاص طراحی شده‌اند که در آن داده‌های آموزشی و آزمایشی از توزیع آماری یکسان (IID[4]) تولید می‌شوند. وقتی این مدل‌ها در دنیای واقعی اعمال می‌شوند، مهاجمان ممکن است داده‌هایی را ارائه دهند که این فرض آماری را نقض می‌کند. این داده ها ممکن است برای سوء استفاده از آسیب‌پذیری‌های خاص و به خطر انداختن نتایج تنظیم شوند.

چندین مدل یادگیری ماشین، از جمله شبکه‌های عصبی، به‌طور مداوم نمونه‌های متخاصم[5] را به اشتباه طبقه‌بندی می‌کنند - ورودی‌های ساخته‌شده مخصوصاً به گونه‌ای طراحی شده‌اند که برای انسان‌ها «معمولی» به نظر برسند، اما باعث طبقه‌بندی اشتباه در مدل یادگیری ماشین می‌شوند. اغلب، یک شکل از «نویز» طراحی شده ویژه برای برانگیختن طبقه بندی اشتباه استفاده می شود - نتایج را در خروجی مدل با یک پاسخ نادرست با اطمینان بالا نشان می دهد. هدف این پژوهش بررسی انواع این تکنیک‌ها برای تولید نمونه‌های متخاصم و مهار کردن آنها می‌باشد.

• منابع

[1] Alexey Kurakin, Ian J. Goodfellow, Samy Bengio, “Adversarial Examples in the Physical World,” in Artificial Intelligence Safety and Security, *Roman V. Yampolskiy*, Ed.Boca Raton: Chapman and Hall/CRC, 2018, pp. 99-112. doi.org/10.1201/9781351251389

[2] Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy. “Explaining and Harnessing Adversarial Examples,” unpublished

[3] Honggang Yu, Shihfeng Zeng, Teng Zhang, Ing-Chao Lin, Yier Jin, “EXPLORING ADVERSARIAL EXAMPLES FOR EFFICIENT ACTIVE LEARNING IN MACHINE LEARNING CLASSIFIERS,” unpublished

[4] Bc. Matˇej Koci´an, “Adversarial Examples in Machine Learning,” Master Thesis, Dept. Theoretical CS & Mathematical Logic., Charles Univ., Prague, Czechia, 2018

- ## هدف آرمانی پژوهش

درک و رسیدن به بهترین مدل مهار کردن انواع مختلف مثال‌های خصمانه با حداکثر دقت و سرعت، و گرفتن خروجی صحیح با اطمینان بالا.

- ## هدف کلی پژوهش

تشخیص، مقایسه و درک انواع مختلف مثال‌های خصمانه در یادگیری ماشین و شکل دادن راه‌حلی کلی برای مهار کردن آن.

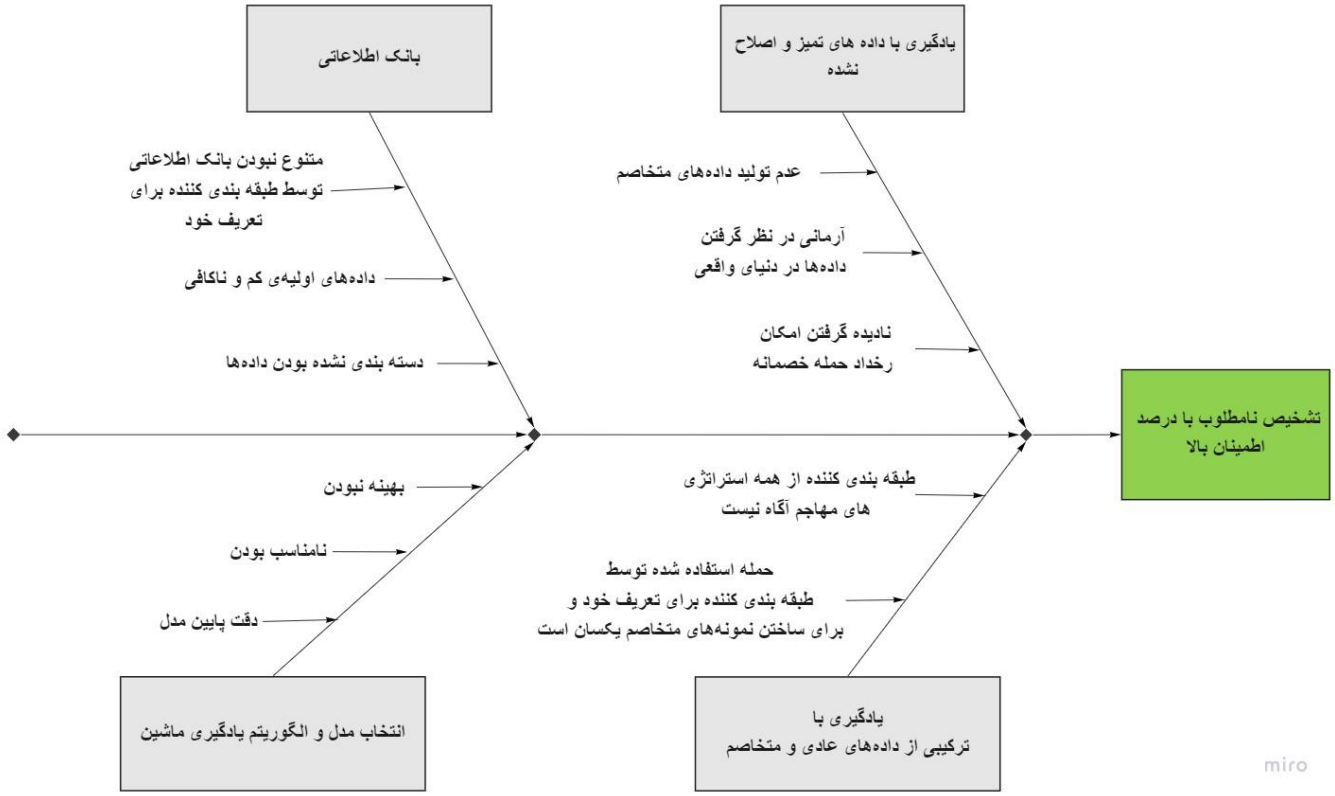
- ## اهداف ویژه‌ی پژوهش

1. شناخت رایج ترین راهبرد های یادگیری ماشین خصمانه.
2. شناخت انواع مختلف حملات خصمانه.
3. شناخت تکنیک‌های فعلی برای تولید نمونه‌های متخاصم.
4. شناخت سازوکارهای‌های دفاعی فعلی در برابر راهبردهای یادگیری ماشین متخاصم.
5. تعیین رویکرد چند مرحله‌ای برای محافظت از یادگیری ماشین.

- ## اهداف کاربردی پژوهش

1. تعیین انواع مختلف حملات خصمانه.
2. تعیین نتایج حملات خصمانه در سیستم از راه دور در دنیای واقعی.
3. تعیین تکنیک‌هایی برای تولید نمونه‌های متخاصم.
4. تعیین مناسب‌ترین سازوکارهای در برابر راهبردهای یادگیری ماشین متخاصم.
5. تعیین مناسب‌ترین رویکرد برای محافظت از یادگیری ماشین.

- **دیاگرام استخوان ماهی:**



- **نقشه ذهن:**

[1] Artificial Intelligence (AI)
[2] Machine learning
[3] Adversarial machine learning
[4] Independent and identically distributed random variables
[5] Adversarial examples