UNIVERSITY OF GOTHENBURG
CHALMERS UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF
GOTHENBURG

CHALMERS

DAT340 / DIT867 Applied Machine Learning

Writing Assignment 2: Machine learning meets the real world

Neha Devi Shakya
University of Gothenburg
CY Tech (Exchange student)
gusshane@student.gu.se

Khushi Chitra Uday
University of Gothenburg
CY Tech (Exchange student)
guschikh@student.gu.se

Sarvesh Meenowa
University of Gothenburg
CY Tech (Exchange student)
gusmeesa@student.gu.se

Date Last Edited: May 23, 2022

Artificial Intelligence (AI) and Machine Learning (ML) are the two most popular terms in each business in recent years. For many businesses, innovation relies upon these two terms or technological advancements. AI is a broader term that refers to machines efficiently performing numerous tasks that humans perceive as "clever." On the other hand, ML is an AI application that grants machines access to data and requires them to analyse it themselves. Surprisingly, the expansion of the internet and the amount of broad virtual statistics prepared the way for methods of ML enhancement. Both technologies have assisted several industries in informing themselves to innovate and sustain themselves. Regardless of whatever industry each technology has brought to, it is exceptional to discuss the changes in many sectors.

Machine Learning is a branch of artificial intelligence (AI) that enables computers to investigate and improve themselves robotically without being explicitly programmed. Machine Learning strives to create programming languages that can access statistics and apply them for study. The learning process begins with observations or information, such as examples, direct experience, or education. On the other hand, people are looking for trends in records and making better decisions based solely on the criteria provided. The primary goal is to enable computer systems to analyse routinely without human intervention or assistance and change movements. Machine learning and algorithms are currently influencing many facets of our life. Some of these effects are well known, while others are less so.

Most people are already aware that when websites with recommendation engines, such as YouTube, Netflix, or Amazon, are used, every choice is made, from the videos liked and disliked to how long a film was watched and the things bought, tracked and logged. Powered by machine learning, these sites use this data to "recommend" or "suggest" other related products, videos, or films that might be enjoyed. Recommendation engines are an utterly innocuous application of machine learning in our daily lives. An implicit assent to such uses of machine learning is (usually) provided, and the outcomes are sometimes precious.

The further along the spectrum of machine learning moved from "benign and harmless" to "disturbing and insidious", the more it is found that machine learning and algorithms influence people in ways they are not always comfortable with. "Targeted advertising" is somewhere in the middle, flooding social media ads that seem to know more about the users' buying habits and online activities than they are willing to accept.

However, what if learning machines had a significant impact on our pensions instead of prospective vacation spots? The vast majority of trades in financial markets are now virtually totally handled by algorithms and learning machines. This information has made a few people money. In May of 2010, the Dow Jones Industrial Average dropped 1,000 points in seconds before regaining equilibrium roughly 20 minutes later. Navinder Singh Sarao, a trader with his troublesome learning robots, exploited a flaw in Dow's micro trading method.

Another example, CRUSH, which stands for Crime Reduction Using Statistical History, is an algorithm that uses predictive analysis to reduce crime in a particular geographic area [4][9]. In 2005, the CRUSH IBM predictive analytics application was developed as a trial project in Memphis, Tennessee. The Memphis police force set a new record for arrests in a single day in its first three days of existence, with 1,200 arrests. Over the next six years, the city's crime rate decreased by more than 24%, and police departments worldwide began to covet it.

The idea was simple enough to analyse historical data on where and when crimes were com-

mitted in the past to predict where they are most likely to occur. Increase police patrols in these targeted areas and watch arrests increase. Civil rights organisations have raised concerns about this method, claiming that it can become a self-fulfilling prophecy perpetuating misery in historically poor communities with populations disproportionately targeted by police.

Human biases are well documented and range from implicit tests that reveal predispositions they may not be aware of to studies that show how these biases can affect outcomes. Although humans are prone to error and bias, this does not mean that algorithms are inherently superior. AI bias is a variation in the output of machine learning algorithms caused by biased assumptions during the algorithm development process or predispositions in the training data. Nonetheless, these systems can be impacted by how they are designed, built, and ultimately used, a phenomenon known as algorithmic bias. It is hard to recognise how frameworks are defenceless to algorithmic predisposition since this technology frequently works in a corporate black box.

Cognitive biases are unintentional errors in thinking that influence people's judgments and decisions and result from the brain's attempt to simplify processing information about the world. Developers could inadvertently introduce cognitive biases into machine learning algorithms or through a training dataset that contains these biases. On the other hand, incomplete data may not be representative and contain biases. When one thinks of data, one may think of formal studies where demographics and representation are carefully considered, constraints weighed, and peer-assessed results. That is not necessarily the case with the AI systems that might decide for us.

An example of how training data can lead to racism in an algorithm happened a couple of years prior, in 2016, when Microsoft was about to launch its new Twitter chatbot, Tay [10]. It was portrayed as an "experiment in conversational understanding." It was intended to draw individuals into a discussion through tweets or direct messages, imitating the style and jargon of a teenage girl. Unlike other chatbots, such as Joseph Weizenbaum's Eliza, Tay was designed to learn more about the English language and, over time, have the option to engage in conversations about virtually any topic. When Tay was first released on Twitter, she entertained her growing number of followers with light banter and terrible jokes. However, after just a few hours, Tay began tweeting incredibly offensive things, with a disturbing percentage of her tweets being abusive and offensive. What the company had planned as a fun experiment in "conversational understanding" turned into a golem spiralling out of control through the power of language, and a bot intended to impersonate a youngster's discourse became vicious. Under certain circumstances, it may be impossible to find biassed training data.

Another application is the use of ML to improve password-guessing algorithms. Although this application is being explored with moderate success, it is essential to point out that the same application has already proven to be more efficient than traditional approaches. Therefore, it is not far-fetched to assume that AI-based algorithms and tools are constantly being designed and developed by individuals or groups who could misuse them.

Traditional password-guessing tools, such as HashCat and John the Ripper, typically compare many different variants against the password's hash value to determine the password that matches the hash value [2][6]. The trials are generated from a dictionary of commonly used passwords; then, variations are made based on the composition of the password. Using neural networks and generative adversarial networks (GANs) makes it possible to analyse

a large dataset of passwords and generate variations that fit the statistical distribution, e.g. for password leaks, leading to more targeted and effective password guessing [1]. The first attempt in this direction can already be seen in a post on an underground forum from February 2020, which mentions a GitHub repository from the same year where software can analyse 1.4 billion credentials and create rules for password variations based on its findings. The PassGAN system, released in 2019, uses a GAN to learn the statistical distribution of passwords from password leaks and generates high-quality guesses [3]. This system was able to match 51% to 73% more passwords compared to HashCat alone.

DeepFake audio-visual content could be presented as 'legitimate' evidence to frustrate criminal investigations and court proceedings, thereby calling into question audio-visual evidence as a whole category of evidence [8]. This evidence would create new legal hurdles for investigators and lawyers to undermine the proceedings' credibility, including the institutions and individuals involved in the same proceedings. Even the administration of the judicial system could be called into question. Moreover, existing vulnerabilities in widely used technologies such as CCTV or police cameras could be used to replace actual footage with DeepFakes. Ultimately, the ability to hack surveillance camera systems opens the door to the concealment of criminal and malicious practices.

After seeing the arguments above about malicious applications and abuses of AI, one should be aware of how they may improve their preparedness to deal with the current and probable future evolution of such usage. AI can be utilised for good by leveraging the technology's potential as a crime-fighting tool to future-proof the cybersecurity sector and law enforcement, building on ongoing work to ensure that AI is trustworthy: lawful, ethical, and technically sound. Implementing a human-centric approach to AI, such as the EU's Trustworthy AI framework, should be encouraged. Developing globally applicable, technology-agnostic policy responses to avoid the potentially malicious use of AI without limiting AI's innovative and promising applications should be encouraged.

While it can bring enormous benefits to society and help solve some of the biggest challenges currently faced, AI could also enable a range of digital, physical, and political threats. Therefore, AI systems' risks and potentially criminal abuse must be well-understood to protect society and critical industries and infrastructures from malicious actors. Risk management approaches could help classify the threats stemming from AI's current and future uses and abuses and prioritise the response measures accordingly.

Criminals are likely to use AI to facilitate and improve their attacks by maximising opportunities for profit in a shorter period, exploiting more victims, and creating new, innovative criminal business models — all the while reducing their chances of being caught[7]. Consequently, as AI becomes a more widespread service, it will also lower entry barriers by reducing the skills and technical expertise required to facilitate attacks, which further exacerbates the potential for AI to be abused by criminals and become a driver of future crimes[5].

Despite its possible misuse, AI has potential for constructive applications, including support for investigating all types of crimes and completing important initiatives worldwide. Understanding the capabilities, scenarios, and attack vectors are critical for improving preparedness, building resilience, and guaranteeing the positive usage of AI to realise its full potential.

References

[1] Jason Brownlee. *18 Impressive Applications of Generative Adversarial Networks (GANs)*. en-US. June 2019. URL: https://machinelearningmastery.com/impressive-applications-of-generative-adversarial-networks/.

[2] *hashcat - advanced password recovery*. URL: https://hashcat.net/hashcat/.

[3] Briland Hitaj et al. "PassGAN: A Deep Learning Approach for Password Guessing". en. In: (Sept. 2017). DOI: 10.48550/arXiv.1709.00440. URL: https://arxiv.org/abs/1709.00440v3.

[4] *IBM100 - Predictive Crime Fighting*. en-US. CTB14. Aug. 2017. URL: http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/crimefighting/.

[5] *Internet Organised Crime Threat Assessment (IOCTA) 2020*. en. URL: https://www.europol.europa.eu/publications-events/main-reports/internet-organised-crime-threat-assessment-iocta-2020.

[6] *John the Ripper password cracker*. URL: https://www.openwall.com/john/.

[7] *Malicious Uses and Abuses of Artificial Intelligence*. en. URL: https://www.europol.europa.eu/publications-events/publications/malicious-uses-and-abuses-of-artificial-intelligence.

[8] *Snapshot Paper - Deepfakes and Audiovisual Disinformation*. en. URL: https://www.gov.uk/government/publications/cdei-publishes-its-first-series-of-three-snapshot-papers-ethical-issues-in-ai/snapshot-paper-deepfakes-and-audiovisual-disinformation.

[9] Tony Thompson. "Crime software may help police predict violent offences". en-GB. In: *The Observer* (July 2010). ISSN: 0029-7712. URL: https://www.theguardian.com/uk/2010/jul/25/police-software-crime-prediction.

[10] James Vincent. *Twitter taught Microsoft's friendly AI chatbot to be a racist asshole in less than a day*. en. Mar. 2016. URL: https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.