Machine Learning: Integrating Security Techniques in Curriculum

CS4991 Capstone Report 2023
May 8, 2023

Maclay Teefey
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
mjt6vj@virginia.edu

ABSTRACT

Machine Learning models have unique vulnerabilities that are not covered as part of UVA machine learning or cybersecurity classes at the University of Virginia. A new special topics course is therefore proposed. In it, students would study the security faults in machine learning models, including de-anonymization attacks, snapshot attacks, and data poisoning. Students would learn to detect and prevent such threats. The course would prepare students to manage current vulnerabilities. By improving students' understanding of machine learning models, it would help them also thwart future attacks. The course could expand with further research into new exploitations of ML models, like prompt engineering bypassing ChatGPT's rules.

1 INTRODUCTION

"Now we've got John's credit card, his address, his phone number," In the wake of Target's 2014 data breach, Avivah Litan, fraud analyst at Gartner, illustrates the personal impact of poor data security (Popken, 2014). In a 2022 report, IBM calculated on average, data breaches cost the afflicted company $4.35 million (IBM, 2022). Data security is important to computer science, especially machine learning (ML).

ML describes the ability for algorithms to learn from training data to solve various tasks (Janiesch et al., 2021). Traditional AI (Artificial Intelligence) algorithms are limited in their ability to learn new information, because they learn through hard coded statements. However, ML algorithms transcend the limitation by learning from training data. ML's data dependence makes it susceptible to unique data security attacks.

Language models built upon ML can auto-complete sentences using their training data. If personal information, like "John Smith's social security is XXX-XX-XXXX" is in the model's training data, hackers can exploit the model by typing out fragments of the sentence and the model will auto-complete the personal information (Quach, 2022). Other techniques including snapshot attacks (Tarasova et al., 2021) and poisoning attacks (Jiang et al., 2019) are only possible because of ML's relationship to its data.

2 RELATED WORKS

Several ML data security experts have discovered attacks against ML models, including poisoning attacks (Jiang et al., 2019), snapshot attacks (Zanella-Béguelin et al., 2020), and de-anonymization attacks (Narayanan & Shmatikov, 2008). Zanella-Béguelin, et al. (2020), have discussed mitigation strategies for the ML algorithms, including differential privacy and truncating output, with potential drawbacks of the model's degradation and utility, respectively.

My project is the creation of a special topics class based around these topics but avoiding the large amount of computing power and time required to use certain ML models.

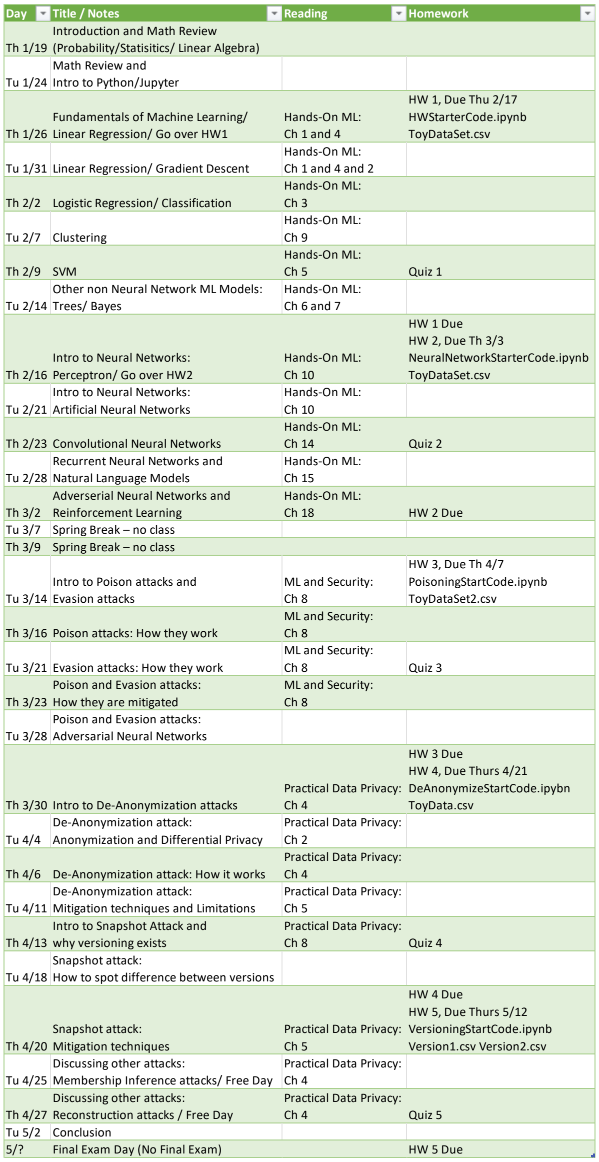
According to Tarasova, et al. (2021), the primary advantages of using project-based learning for the engineering curriculum are increased intellectual stimulation and understanding of engineering processes. The major drawbacks include faculty organizing and implementing students' project-based learning activities and facilitating student-to-student interactions (Tarasova et al., 2021). My approach to developing a special topics class utilizes Tarasova, et al.'s recommendation but I address the drawbacks by purposing toy datasets for large homework assignments that become more complicated as the semester proceeds.

3 PROPOSED DESIGN

The course schedule was designed around 75-minute class periods on Tuesdays and Thursdays. The longer class periods better support the in-depth lecture content and reduce the time required to refresh students on the lecture material. The course will have perquisites of APMA 3080 and CS 2150 or CS 3140 with a C- or better. The schedule for the course is shown in Figure 1.

3.1 Topics

The topics of this course are split into two sections separated by spring recess with multiple subsections: Introduction to ML, with subsections ML fundamentals and Neural Networks, and ML attacks, with subsections Poison and Evasion attacks, De-Anonymization attacks, and Snapshot attacks. Introduction to ML will introduce students to various machine learning models, while ML attacks will introduce students to attack and prevention methods for ML models.



Day	Title / Notes	Reading	Homework
Th 1/19	Introduction and Math Review (Probability/Statisitics/ Linear Algebra)		
Tu 1/24	Math Review and Intro to Python/Jupyter		
Th 1/26	Fundamentals of Machine Learning/ Linear Regression/ Go over HW1	Hands-On ML: Ch 1 and 4	HW 1, Due Thu 2/17 HWStarterCode.ipynb ToyDataSet.csv
Tu 1/31	Linear Regression/ Gradient Descent	Hands-On ML: Ch 1 and 4 and 2	
Th 2/2	Logistic Regression/ Classification	Hands-On ML: Ch 3	
Tu 2/7	Clustering	Hands-On ML: Ch 9	
Th 2/9	SVM	Hands-On ML: Ch 5	Quiz 1
Tu 2/14	Other non Neural Network ML Models: Trees/ Bayes	Hands-On ML: Ch 6 and 7	
Th 2/16	Intro to Neural Networks: Perceptron/ Go over HW2	Hands-On ML: Ch 10	HW 1 Due HW 2, Due Th 3/3 NeuralNetworkStarterCode.ipynb ToyDataSet.csv
Tu 2/21	Intro to Neural Networks: Artificial Neural Networks	Hands-On ML: Ch 10	
Th 2/23	Convolutional Neural Networks	Hands-On ML: Ch 14	Quiz 2
Tu 2/28	Recurrent Neural Networks and Natural Language Models	Hands-On ML: Ch 15	
Th 3/2	Adverserial Neural Networks and Reinforcement Learning	Hands-On ML: Ch 18	HW 2 Due
Tu 3/7	Spring Break – no class		
Th 3/9	Spring Break – no class		
Tu 3/14	Intro to Poison attacks and Evasion attacks	ML and Security: Ch 8	HW 3, Due Th 4/7 PoisoningStartCode.ipynb ToyDataSet2.csv
Th 3/16	Poison attacks: How they work	ML and Security: Ch 8	
Tu 3/21	Evasion attacks: How they work	ML and Security: Ch 8	Quiz 3
Th 3/23	Poison and Evasion attacks: How they are mitigated	ML and Security: Ch 8	
Tu 3/28	Poison and Evasion attacks: Adversarial Neural Networks		
Th 3/30	Intro to De-Anonymization attacks	Practical Data Privacy: Ch 4	HW 3 Due HW 4, Due Thurs 4/21 DeAnonymizeStartCode.ipybn ToyData.csv
Tu 4/4	De-Anonymization attack: Anonymization and Differential Privacy	Practical Data Privacy: Ch 2	
Th 4/6	De-Anonymization attack: How it works	Practical Data Privacy: Ch 4	
Tu 4/11	De-Anonymization attack: Mitigation techniques and Limitations	Practical Data Privacy: Ch 5	
Th 4/13	Intro to Snapshot Attack and why versioning exists	Practical Data Privacy: Ch 8	Quiz 4
Tu 4/18	Snapshot attack: How to spot difference between versions		
Th 4/20	Snapshot attack: Mitigation techniques	Practical Data Privacy: Ch 5	HW 4 Due HW 5, Due Thurs 5/12 VersioningStartCode.ipynb Version1.csv Version2.csv
Tu 4/25	Discussing other attacks: Membership Inference attacks/ Free Day	Practical Data Privacy: Ch 4	
Th 4/27	Discussing other attacks: Reconstruction attacks / Free Day	Practical Data Privacy: Ch 4	Quiz 5
Tu 5/2	Conclusion		
5/?	Final Exam Day (No Final Exam)		HW 5 Due

Figure 1: Proposed Class Schedule

The first week of ML fundamentals introduces the class, reviews the math fundamentals that underlies ML, and teaches students how to use Juypter and Colab, which students will use to run their assignments. The second week will introduce the first homework assignment and teach linear regression and gradient descent. The third week will instruct students about logistic regression, classification, and clustering, while the fourth week will have lectures on support vector machines and other ML models that are not neural networks including trees and genetic models.

The neural networks unit takes up two and a half weeks and starts with an introduction to neural networks and the perceptron and introduces the second homework assignment. The next two weeks teach artificial neural networks, convolutional neural networks, recurrent neural networks, and adversarial neural networks, with discussions into natural language models and reinforcement learning in the context of neural networks.

Each unit of ML attacks is made up of lectures introducing the attacks, describing how the attacks work, and illustrating the attacks are mitigated, but the poison and evasion attacks unit adds a lecture about the attacks' connection to adversarial neural networks and the de-anonymization attacks unit adds a lecture about how anonymization techniques work. The final three class periods are discussions of membership inference attacks and reconstruction attacks, which do not have an associated homework assignment and a conclusion lecture. These lectures can be scrapped in case of unforeseen events.

3.2 Coursework

The coursework will be split into five homework assignments and quizzes with a grading split being the same as the special topics course Foundations of Data Analytics, with 80% of the final grade being homework assignments and 20% being the quizzes. Each unit will have one homework assignment and one quiz, and each quiz will be taken a week before the homework is due.

3.2.1 Homework

Each homework assignment will have a juypter file that has starter code for the unit it is in and csv file(s) with data for the machine learning models. Students will run their code on Colab, and will use Scikit-Learn, Keras, and TensorFlow for the ML models. For the Introduction of ML section, students will create ML models from their respective unit, while the ML Attacks section will have already working ML models that students will attack using the units' respective attack. The final homework on snapshot attacks will act as the final exam of the class and will be due at the end of the day on the class's final exam date.

3.2.2 Quizzes

Quizzes will be taken for each unit to help the professor understand how well the students are understanding the lecture material. Each quiz will be due at the end of the day and will be an online open notes quiz on UVACollab or Canvas.

3.2.3 Textbooks

Three textbooks will be used for background readings for the lectures, all of them available online on O'Reilly for free with the student's UVA account:

1. Hands-On Machine Learning with Scikit-Learn, Keras, & TensorFlow by Aurélien Géron for the Introduction to ML section
2. Machine Learning and Security by Clarence Chio and David Freeman for the poison and evasion attacks unit
3. Practical Data Privacy by Katharine Jarmul for the anonymization and snapshot units.

4 ANTICIPATED RESULTS

By the end of the course, students should be able to meet the following learning objectives:

1. Evaluate multiple ML models and how they work
2. Recognize the security flaws in ML models
3. Understand how poison and evasion, de-anonymization, and snapshot attacks work

4. Evaluate various mitigation strategies to prevent attacks on ML models

These goals will be achieved through the course's homework assignments, which will allow students to learn and show their knowledge in each topic through implementing the ML models, attacks, and mitigation techniques discussed in the lectures.

5 CONCLUSION

In response to the growth of ML and the cost of data breaches, I have developed a special topics course covering the unique attacks ML models face and ways to mitigate those attacks. The course is split into two sections: Introduction to ML and ML attacks, which will teach students ML models and three types of attacks on ML models: poison and evasion attacks, de-anonymization attacks, and snapshot attacks. Students will reinforce the course material through five homework assignments spread across the semester, which tasks students with implementing the ML models, attacks, and protection methods taught in the lectures. By the end of the course, students will be able to evaluate multiple ML models and their security faults and implement security protections on those ML models preventing data breaches and attacks.

6. FUTURE WORK

With the popularization of natural language models like ChatGPT, users have discovered ways of bypassing security measures through forcing the model to enter "Do-Anything-Now" mode (Getahun, 2023). Any addition of natural language models bypasses to the curriculum are limited in scope due to the large computing power to host a natural language model, and the frequent changes to the models from their creators, so if it were added it would be replacing the two discretionary lectures at the end of the semester and be at the instructors' discretion. If any other large scale ML attacks are discovered, the curriculum may be modified by reducing the complexity of the homework assignments and reducing the number of lectures per unit.

REFERENCES

[1] Hannah Getahun. 2023. Breaking ChatGPT: The AI's alter ego DAN reveals why the internet is so drawn to making the chatbot violate its own rules. *Business Insider*. Retrieved April 8, 2023 from https://www.businessinsider.com/open-ai-chatgpt-alter-ego-dan-on-reddit-ignores-guidelines-2023-2

[2] IBM. 2022. Cost of a data breach 2022. Retrieved October 26, 2022 from https://www.ibm.com/reports/data-breach

[3] Christian Janiesch, Patrick Zschech, and Kai Heinrich. 2021. Machine learning and deep learning. *Electron Markets* 31, 3 (September 2021), 685–695. DOI:https://doi.org/10.1007/s12525-021-00475-2

[4] Wenbo Jiang, Hongwei Li, Sen Liu, Yanzhi Ren, and Miao He. 2019. A Flexible Poisoning Attack Against Machine Learning. In *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 1–6. DOI:https://doi.org/10.1109/ICC.2019.8761422

[5] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust De-anonymization of Large Sparse Datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, 111–125. DOI:https://doi.org/10.1109/SP.2008.33

[6] Ben Popken. 2014. Target estimates breach affected up to 110 million. *NBC News*. Retrieved February 24, 2023 from http://www.nbcnews.com/business/business-news/target-says-stolen-info-data-breach-hit-70-million-people-flna2D11894083

[7] Katyanna Quach. 2022. ML models models leak data after poisoning training data. *The Register*. Retrieved February 24, 2023 from https://www.theregister.com/2022/04/12/machine_learning_poisoning/

[8] E. N. Tarasova, Olga Khatsrinova, G. N. Fakhretdinova, and Alla A. Kaybiyaynen. 2021. Project-Based Learning Activities for Engineering College Students. In *Educating Engineers for Future Industrial Revolutions* (Advances in Intelligent Systems and Computing), Springer International Publishing, Cham, 253–260. DOI:https://doi.org/10.1007/978-3-030-68201-9_26

[9] Santiago Zanella-Béguelin, Lukas Wutschitz, Shruti Tople, Victor Ruehle, Andrew Paverd, Olga Ohrimenko, Boris Köpf, and Marc Brockschmidt. 2020. Analyzing Information Leakage of Updates to Natural Language Models. In *ACM Conference on Computer and Communication Security (CCS)*, ACM. Retrieved from https://www.microsoft.com/en-us/research/publication/analyzing-information-leakage-of-updates-to-natural-language-models/