**EMCS2020: Advanced Topics in Computer Security**

Assignment: Machine Learning and Security

Brian Russel Davis, brian\_davis@brown.edu

**Question 1**

According to Yann LeCun (Professor of Computer Science at NYU and VP and Chief AI Scientist at Facebook), the technology of Generative Adversarial Networks (GANs) is one of the most promising directions in machine learning. After doing your own research to learn more about this technology, describe the high-level idea of GANs and mention one application scenario where the results of applying GANs surprised you the most.

I am an outlier on this topic as I don’t believe that true AI exists. Judging from his previous lectures, Professor LeCun also doesn’t believe that true AI exists[[1]](#footnote-0), which is why I am perplexed by his comments on GANs being promising. This idea of two machines facing off to sharpen each other skills sounds interesting indeed, but I think it ignores some of the staples of ML and Data Science. First, the tools don’t have inherent value on their own, meaning outside of human direction/guidance. You cannot teach a machine to play like a master chess player by feeding it basic chess rules and facing it off against another machine that is a beginner chess player. I can see slight time savings by giving a GAN remedial work, but all the valuable ML we have now is always sourced and verified by human intelligence. Even with all of the choices and all the data, “machines” still cannot predict, discern, qualify or create much past the level of a 5-year-old child. This should not be surprising, it’s very much analogous to the way the study of thermodynamics and non-equilibrium thermodynamics are related but the rules of the first seldom hold in the latter. Simulations of intelligence have a long way to go, and pitting them against one another in GAN as a way of discovering something new, is at best a *McNamara fallacy ( quantitative fallacy )*: “making a decision based solely on quantitative observations and ignoring all others.”[[2]](#footnote-1) Having all the data is not the only factor in making a ‘good’ decision.

**Question 2**

A colleague comes up with the following idea to break standard AES block cipher encryption:

“Since machine learning algorithms are so good at learning functions, let’s train an ML system with a set of AES ciphertexts labeled with the corresponding plaintexts (assume for simplicity that the encryption key is fixed). This would generate an ML model that is able to invert AES, that is, decrypt ciphertexts into the corresponding plaintexts. The model will successfully guess future unobserved plaintexts by processing ciphertexts through the learned function, thus breaking all modern secure communication and storage solutions!”

Explain why this idea does not work.

Hint: Recall that AES and other block ciphers satisfy the property that their output is indistinguishable from a random output of the same length.

This doesn’t work because there is no pattern for the machine to learn as assumed by the statement: “let’s train an ML system with a set of AES ciphertexts labeled with the corresponding plaintexts.” The colleague is falsely assuming that ML can “find” key by looking at pairs of ciphertexts and plain texts, however there is not magic algoritm to undo a one way trap door function. Of course there is always brute force, trying all the combinations until you find the right one. But there is no way to “learn” from such a small dataset. Machine learning is about analyzing large data sets, training and finding the rules the exist to tie things together with math. In this case however the math used to relate the cyphertext and the plain text is designed to be irreversible and the variables can only be discovered with brute force.

**Question 3**

Consider a chatbot trained to interact with the customers of a bank to provide automated customer service. (a) Outline what kind of training data should be used so that the chatbot will learn how to interact with the clients fluently. (b) Suppose we want to improve the performance of the chatbot by using subsequent actual chat sessions. Discuss whether it is a bad or a good idea to retrain the chatbot from unedited chat transcripts.

Hint: Keep in mind that in a chat session, customers may type in private information (such as names, addresses, account numbers, account balances, income, etc.) or inappropriate language.

Usually bots are trained by looking at user chat records and discovering the highest mode among customer questions. Since find the first order mode for an array of questions, we can use the question as a first order intent and use natural language processing to record all the possible ways a user can ask the same question. We can look and see if there is a high correlation between the satisfication of the user and the agent providing any one answer or solution. If there is a high correlation between the users This is would be the algoritmic way of determining the most common question and the most satisfying answer. For this problem would recommend a Supervised Learning Model ( Discrete Variable Prediction ) with a Decision Tree Classifier that is power by rich set of curated user intents. Retraining the bot based on unedited chat transcripts is fine ( and is probably best ) as long at the models are preprocessed to dealing with missing data, handle data imputation, handle categorical data, encoding class labels for classification problems, indluces features to transform the data and deal with dimensionality reduction, such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA).

**Question 4**

Write a summary of the talk by Michael Littman on "Fairness, Accountability, and Transparency in Machine Learning: An Overview." Your summary should be intended for a layperson.

Littman’s presentation is an overview of how the field of data science differentiates itself from other parts of computer science. The first half of his presentation seems to focus on the scope of problems that data science is trying to solve and what these problems have in common. It’s pretty apparant that the problems that data science is able ot solve best are the ones with quantifiable relations. A quantifiable relation would be a relation that can made in to a mathematical rule. However there are many problems with this approach. In logic we even have an informal fallacy called the *McNamara fallacy ( quantitative fallacy ).* As I mentioned above there are more factors to solution than the details that can be quantified. Morever there are many that cannot be quantified at all. Using machine learning, artificial intelligence or computer vision to solve problems that don’t have quantifiable outcomes may result in mimiced bias, unethical ( or more precisely aethical behavior) or worse a combiniation of both. Data science has a great deal of promise, but it should considered a tool wielded by humans, not a tool that wll one day become human.

1. (2017, October 17). Retrieved November 18, 2019, from https://www.youtube.com/watch?v=aCCotxqxFsk. [↑](#footnote-ref-0)
2. Baskin, J.S. 2014. "According To U.S. Big Data, We Won The Vietnam War," Forbes Magazine. https://www.forbes.com/sites/jonathansalembaskin/2014/07/25/according-to-big-data-we-won-the-vietnam-war/#395975054c5a [↑](#footnote-ref-1)