# CSS 590 Adversarial Machine Learning

## Course Description:

Machine learning has emerged as a technique that is used to add intelligence to numerous applications. This course will examine the case when machine learning algorithms are attacked by an adversary. The course will cover the basics of machine learning, the machine learning threat model, and the existing methods for securing applications built on machine learning algorithms. Students will identify weaknesses in machine learning algorithms and develop solutions to secure them. The course will include machine learning in traditional environments such as intrusion detection and emerging environments such as pervasive computing applications. The goal of this course is to provide students with necessary background and experience with innovative research in adversarial machine learning, so that they have the confidence and training to pursue further research in the field.

## Work Load and Grading:

| **Course Work** | **Percentage** | **Grades** | **Approximately Corresponding Numeric Grade** |
|---|---|---|---|
| Research Proposal | 30% | 90s | 3.5 - 4.0 |
| Reinforcement Exercises | 20% | 80s | 2.7 - 3.4 |
| Final Project/Paper | 50% | 70s | 1.7 - 2.7 |

## Textbooks/References:

This course has no required text books. Our reading material will generally be from recently published research papers.

Many of these can be found via Google Scholar, Microsoft Research, or Citeseer. Additionally, papers can often be found at:

IEEE XPloreLinks to an external site. -- Search for publications through IEEE
ACM Digital LibraryLinks to an external site. -- Search for publications through ACM.
Usenix ConferencesLinks to an external site. -- Listing of Usenix conferences

http://www.cs.cornell.edu/info/misc/latex-tutorial/latex-home.htmlLinks to an external site. -- LaTeX tutorial

http://www.maths.tcd.ie/~dwilkins/LaTeXPrimer/Links to an external site. -- LaTeX help

## Software:

http://miktex.org/downloadLinks to an external site. -- Tex typesetting software http://texstudio.sourceforge.net/Links to an external site. -- LaTex editor

http://www.texniccenter.org/Links to an external site. -- Another LaTeX editor

http://www.bobsoft-mac.de/texnicle/texnicle.htmlLinks to an external site. -- LaTeX editor (Mac)

https://www.sharelatex.com/ (Links to an external site.)Links to an external site. -- online latex editor

## Course Goals:

The overall goals of CSS 590, "Adversarial Machine Learning" include:

- Understand the vulnerabilities present in machine learning based systems
- Analyze and extend existing state of the art work in the security of data analysis systems

## Assignments:

Assignments will consist primarily of research-based projects:

1. Reinforcement Exercises -- These reinforce the topics learned in class. They will be due (or will occur) throughout the quarter. They will include checkpoint exercises, quizzes, programming assignments, and discussions about adversarial machine learning. Expect between 5-8 of these over the quarter, depending on the intensity of the exercises. Approximately 20 hours over the course of the quarter is expected.
2. Research Proposal -- You will write a research proposal. As other assignments are kept to a minimum, you are expected to spend most of your class-related work on this in the first 3 weeks. 30-50 hours of work is expected.
3. Final project -- This will be a culmination of the skills and techniques we have learned in the class. You will test your hypothesis about an open problem in the field of security in emerging environments and writeup and present your results. During the remaining 7 weeks, you are expected to spend most of your class-related work on this project. Approximately 100 hours of work outside of your research proposal is expected.

## Topics covered and tentative 490/590 schedule:

Note that this is an approximate ordering of topics. Chapters will take about the allotted time and not all sections in all chapters are covered. If there are major changes they will be announced to the class.

| Week | Topics | | Reading | Additional Information |
|---|---|---|---|---|
| 0 | Introduction and ML | | | |

| | | | | |
|---|---|---|---|---|
| | Adversarial ML Overview | | | |
| 1 | Overview of Poisoning Attacks | | | |
| | Overview of Poisoning Defenses | | | |
| 2 | Overview of Adversarial Examples | | | |
| | Overview of Adv Ex. Defenses | | | |
| 3 | SVM-Specific | | | |
| | | | | |
| 4 | Neural Network-Specific | | | |
| | | | | |
| 5 | Online Learning | | | |
| | | | | |
| 6 | Clustering and Anomaly Detection | | | |
| | | | | |
| 7 | Application-specific attacks | | | |
| | | | | |
| 8 | | | | |
| | | | | |
| 9 | Human Factors | | | |
| | | | | |
| 10 | Final Presentations | | | |
| | | | | |