Adversarial and Secure Learning

Summer Semester 2018, Chair I20, TUM

Bojan Kolosnjaji, TUM
Huang Xiao, Fraunhofer AISEC
Machine learning is everywhere

- Computer vision
- Speech recognition
- Biometrics
- Text processing
- Recommendation systems
- Spam detection
- Malware Detection
- ...

MACHINE LEARNING EVERYWHERE
Learning in adversarial environment

- Problem considered in the research community at least since early 2000s
- With the hype over machine learning (deep learning) the problem gains importance
- Adversarial perturbations studied in vision, text, malware...
Danger - ML systems are vulnerable

- Easy to perturb data and cause misclassification

Danger - ML systems are vulnerable


Not only computer vision...


Rising interest

● In the research community
  ○ Many new papers in top level ML and security conferences, especially since 2015.
  ○ Still unsolved problems

● In the tech media and general public
  ○ AI unreliable?
  ○ What if AI is hacked, we are doomed...
We need to consider attacks (security)

- Potentially unreliable:
  - Training data - poisoning
  - Test data - evasion

- Evaluate security under adversarial environment

- Think about designing robust systems
Arms race

1. Analyze learner (Exploring)
2. Perform adversarial attacks
3. Analyze attacks’ impact
4. Countermeasures (e.g., retraining, adding/removing features)

1. Adversary modelling
2. Simulate potential attacks
3. Evaluate attacks’ impact
4. Enhance system security by design

reactive
proactive
Seminar goals

- Investigate inherent vulnerabilities of ML methods
- Special interest for: SVM, Neural Networks, Random Forest
- Consider attack types and countermeasures
- Study problems in various application scenarios
- Be aware of security when applying ML in the future
- Prepare for further research in this area
Seminar topics (1)

- **Evasion** of machine learning classification algorithms
- **Feature selection** in adversarial environment
- Attacks on **Support Vector Machines (SVM)**
- Connections of **Robustness** and **Regularization** in SVM
- Analysis of adversarial examples for **Neural Networks**
- Adversarial attacks on reinforcement learning, sequence labeling, structured prediction
Seminar topics (2)

● **Generative Adversarial Networks**, Adversarial Autoencoders
● Techniques for increasing robustness of **Neural Networks**
● Adversarial attacks on spam detection
● **Poisoning** malware clustering
● Evading malware detection systems
● Attacks on graph-based **anomaly detection** in DNS data

...
Seminar plan

- 12 students, 12 topics, 12+1 seminar meetings
- Each student gets a topic with two state-of-the-art research papers
- Every student presents his papers on one seminar meeting
- Students write a short report to summarize their topic
- Grading based on the presentation and report
Prerequisites

- Student of Informatics or similar (advantage to Master students)
- Machine Learning - basic knowledge
- Interest in deeper knowledge of ML methods
How to apply?

● Send an e-mail to kolosnjaji@sec.in.tum.de with the following information:
  ○ Previous knowledge that qualifies you for the seminar (Machine Learning courses, internships, independent projects,...)
  ○ Optional: what topics are of your special interest

● Apply through the matching system

● We divide the papers after the student registration
More information

- Follow the course website:
  https://www.sec.in.tum.de/i20/teaching/ss2018/adversarial-and-secure-machine-learning

- Ask course organizers:
  Bojan Kolosnjaji, TUM: kolosnjaji@sec.in.tum.de
  Huang Xiao, Fraunhofer AISEC: huang.xiao@aisec.fraunhofer.de