FIT5230 Malicious AI

All about AI vs Security

Overview

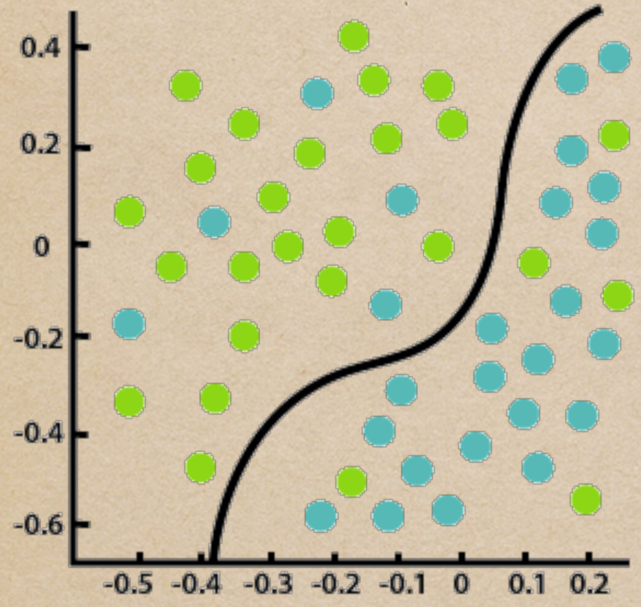
- The case for AI+Security
- AI vs Security
  - AI for Security: AI → Sec
  - Security attacks AI: AI ⇤ Sec
  - Security meets AI: AI ↔ Sec
  - AI attacks Security: AI ⇥ Sec

AI @Monash…

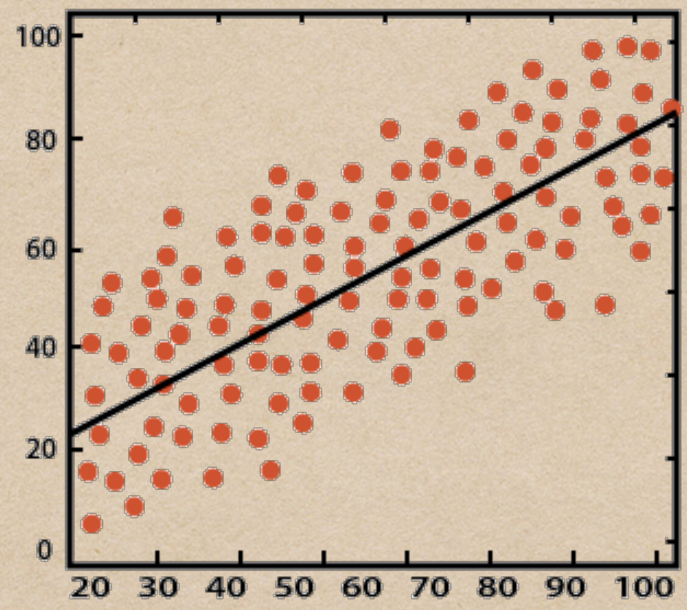
- Your journey into the AI world @PG
- FIT5047 Fundamentals of AI
- FIT5201 Machine Learning
- FIT5215 Deep Learning
- FIT5216 Modelling Discrete Optimization Problems
- FIT5217 Natural Language Processing (NLP)
- FIT5221 Intelligent Image & Video Analysis
- FIT5222 Planning & Automated Reasoning
- FIT5226 Multi Agent Systems & Collective Behaviour
- FIT5230 Malicious AI

AI

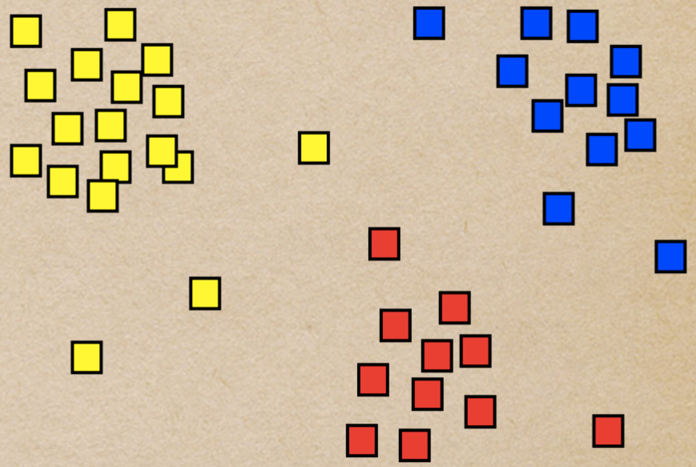
- supervised learning, unsupervised learning, …
- samples have/not labels



Classification



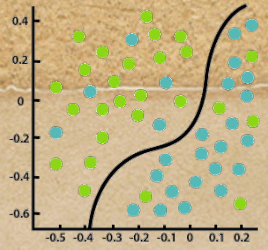
Regression



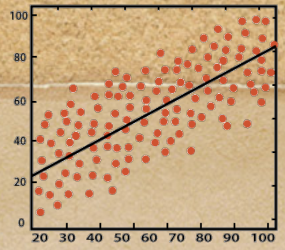
Clustering

- samples assumed benign/correct
- Q: does each sample affect learning outcome?

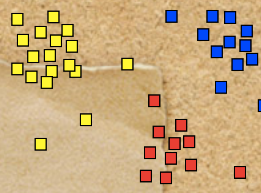
AI



Classification



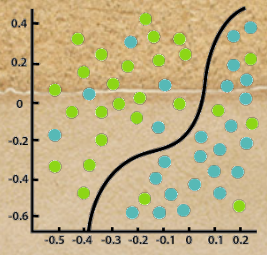
Regression



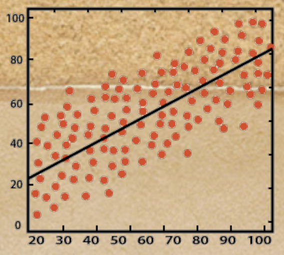
Clustering

- conventional AI: idealistic, too trusting, world w/o malice
  - done by single party/entity/organization
  - the only (few) problematic samples, due to error, imprecision, not malice
- collaborative multi-party AI
  - multiple parties (coalitions of nations) jointly do ML e.g. facial recognition across countries
  - could bias the joint ML outcome
- ML on datasets in the wild
  - could bias the ML outcome

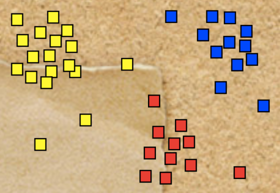
AI



Classification



Regression



Clustering

- conventional AI: too idealistic
  - if the world has malice, why won't samples be affected?

- robust AI
  - against coalition ML
  - against datasets in the wild
  - should be resilient to sample corruptions

Security vs AI

- AI for security:
  - biometrics, surveillance: pattern recognition for identification
  - forensics, intrusion/malware detection: ML for anomaly detection

- security attacks AI:
  - adversarial ML: attacks on INTegrity of samples

Security vs AI

- security gaming meets AI:
  - adversarial modelling: two opposing sides, two opposing goals, interacting
  - generative adversarial networks (GAN)

Security vs AI

- AI attacks Security
  - ML generates/fabricates forgeries of real samples
  - deepfakes
  - Q: which security goal is attacked?

Security

- adversarial gaming btw 2 interacting sides 웃 ⇔ 웃
- opposite goals ↑ security vs ↓ security
  - e.g. ↑ privacy vs leak privacy (enter PIN vs guy nearby)

  - attacker vs defender Light vs Dark, good vs bad
  - each has capabilities Access to info, interact w each other

- Q: is it fair? Why/not?

Real Security: How to Win Games

- Security
  - adversarial gaming btw 2 interacting sides 웃 ⇔ 웃
  - opposite goals ↑ security vs ↓ security
  - attacker vs defender Light vs Dark, good vs bad
  - each has capabilities Access to info, interact w each other

vs

- Playing games e.g. Chess, Othello, Go: man vs machine
  - 1997: Deep Blue defeated Kasparov
  - 1997: Logistello defeated Murakami
  - 2016: AlphaGo defeated Sedol

Security vs Games

- Security:
  - 2 interacting/playing sides 웃 ⇔ 웃
  - opposite goals each wants to win
  - attacker vs defender, each has capabilities but unfair:
    - attacker has upper hand, can target defender,
    - we can't win by only defending, security should be fair to both sides

vs

- Playing games: man vs machine. Q. unfair, why still?
  - man: similar brain to creator of machine
  - machine: much faster, huge memory, exhaustive