Adversarial Machine Learning

Daniel Lowd
University of Oregon

Example: Spam Filtering

1. From: spammer@example.com
   Cheap mortgage now!!!

Feature Weights

2. cheap = 1.0
   mortgage = 1.5

3. Total score = 2.5 > 1.0 (threshold)

Spam

Example: Spammers Adapt

1\. From: spammer@example.com
Cheap mortgage now!!!
Eugene Oregon

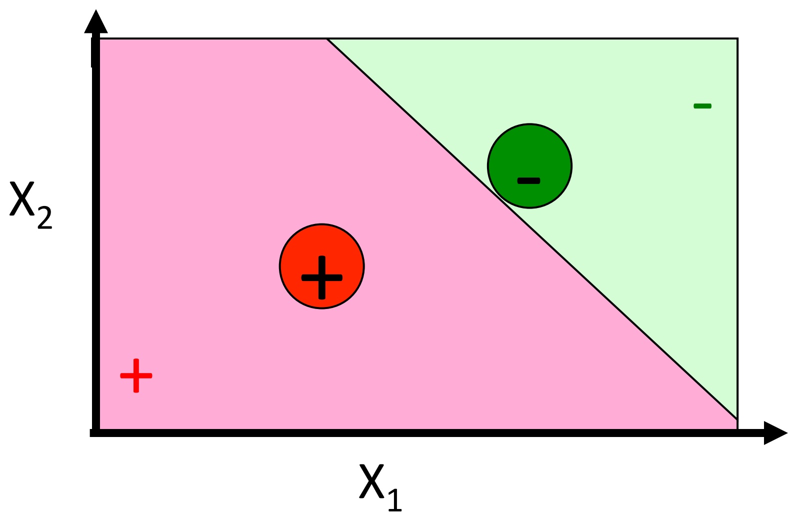
Feature Weights

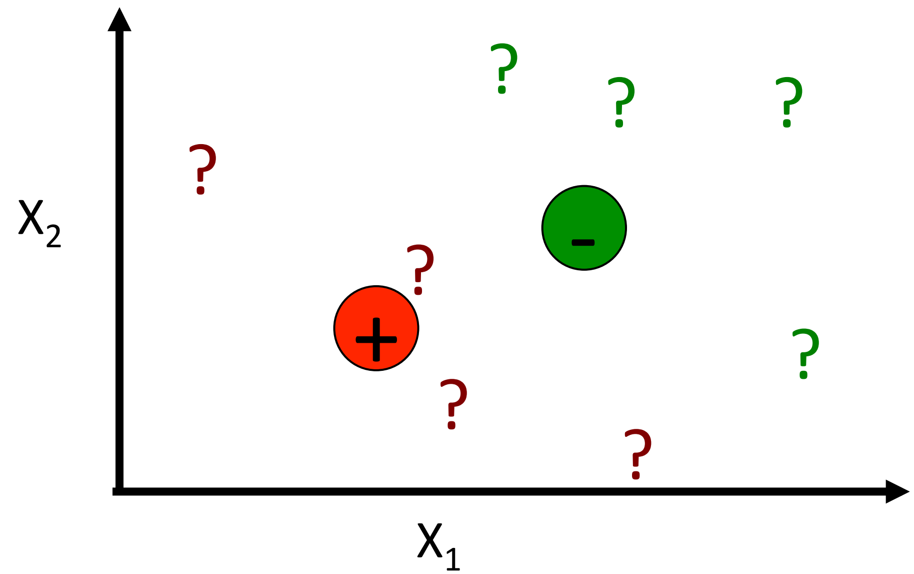
2\. cheap = 1.0
mortgage = 1.5
Eugene = -1.0
Oregon = -1.0

3\. Total score = 0.5 < 1.0 (threshold)

OK

Are Linear Classifiers Vulnerable?



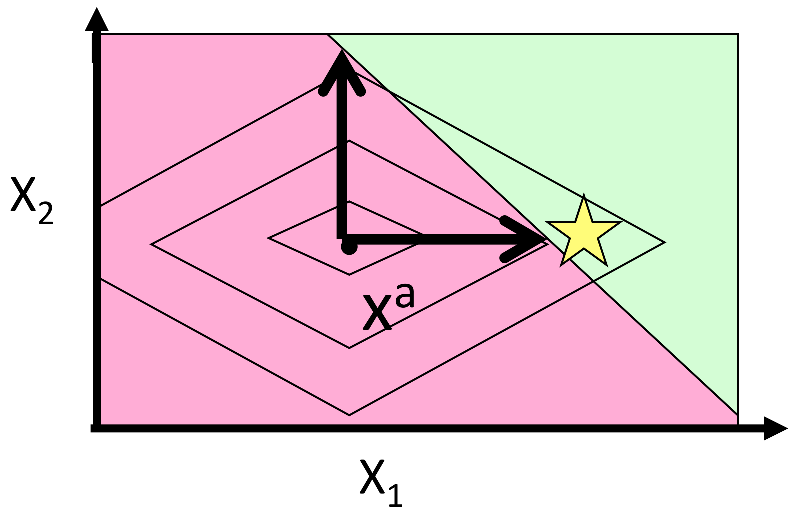


Adversary wants to find the best spam email that will go through the filter.

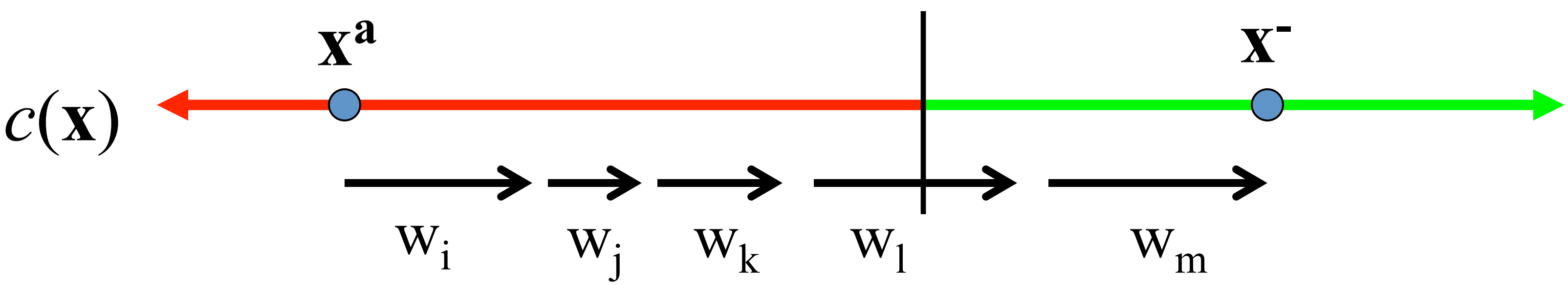
In general: lowest-cost instance classified as negative, for some cost function and some set of classifiers.

Attacking Linear Classifiers

- With continuous features, find optimal point by doing line search in each dimension:



- With binary features, take a negative instance (non-spam) and reduce its cost until we have a factor of 2:



[Lowd & Meek, 2005]

Experimental Results

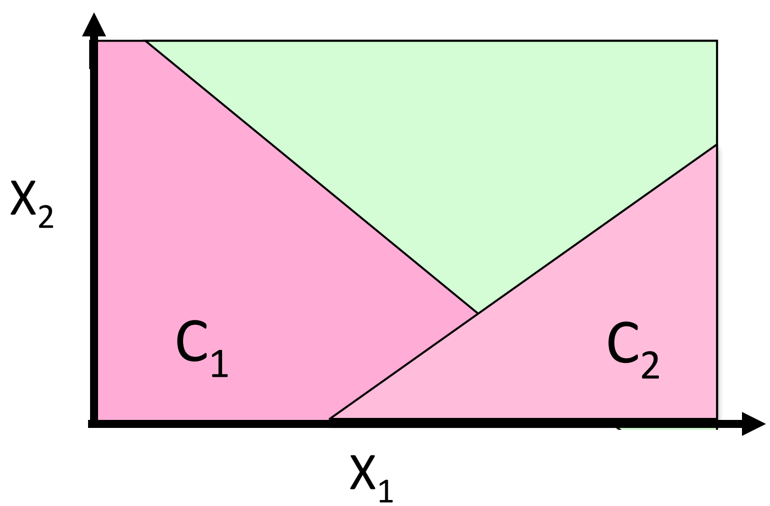
- Realistic spam filter trained from Hotmail data.
- How many words do you have to change to get the median spam past the filter?
- How many queries does it take?

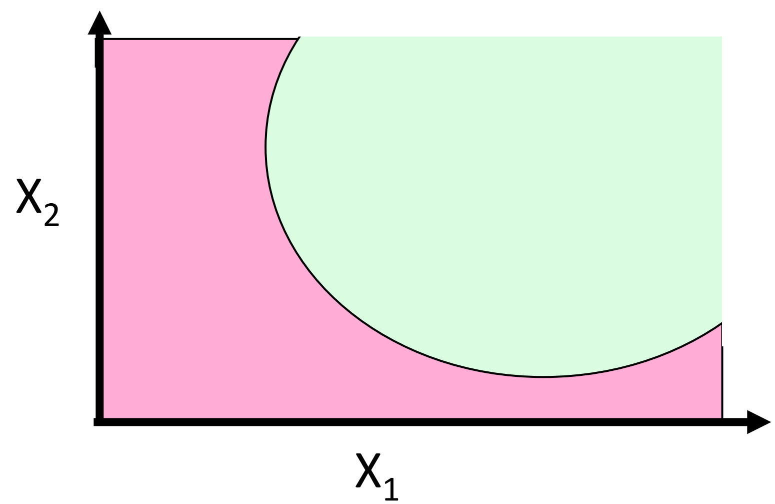
Attack type	Naïve Bayes words (queries)	Logistic reg. words (queries)
Active	31* (23,000)	12* (9,000)
Passive	112 (0)	149 (0)

[Lowd & Meek, 2005]

Evading Classifiers: Ongoing Work

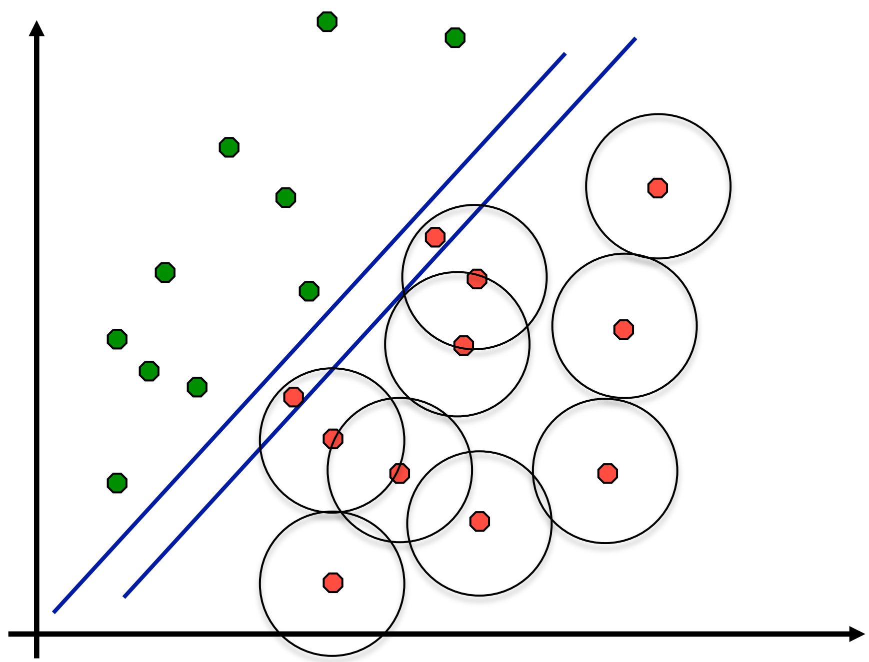
Which classes of non-linear classifiers can we efficiently evade, and under what assumptions?





Robust Machine Learning

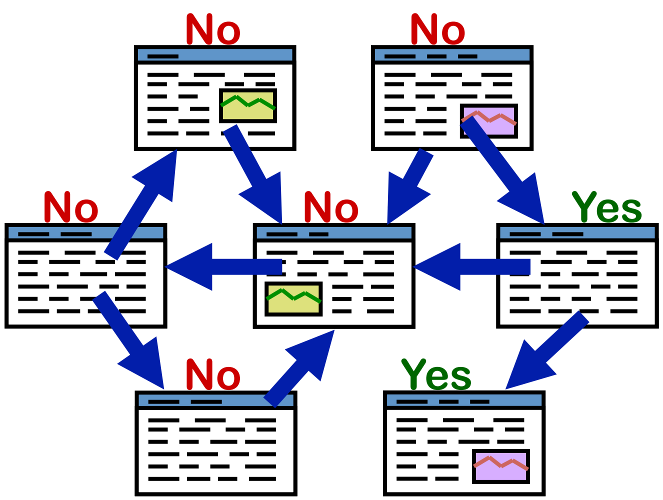
Scenario: Adversary knows our classifier and can maliciously modify data to attack.



Goal: Select the best classifier, assuming the worst adversarial manipulation. (Zero-sum Stackelberg game.)

Robust Machine Learning

- Previous work: Linear classifiers
- Our work: Relational domains
  Examples: Web spam, eBay fraud, etc.



[Brin&Page98; Chakrabarti&al98; Abernethy&al08]

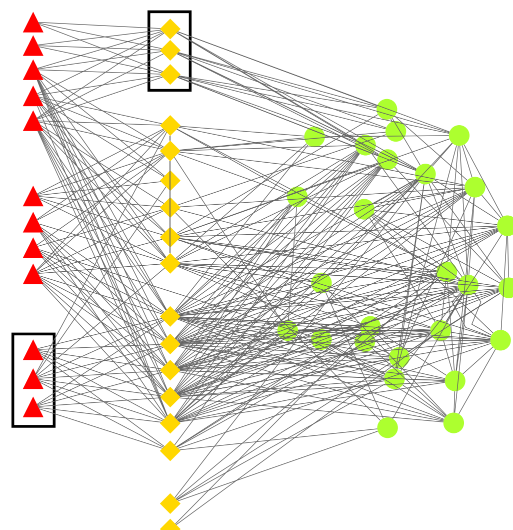
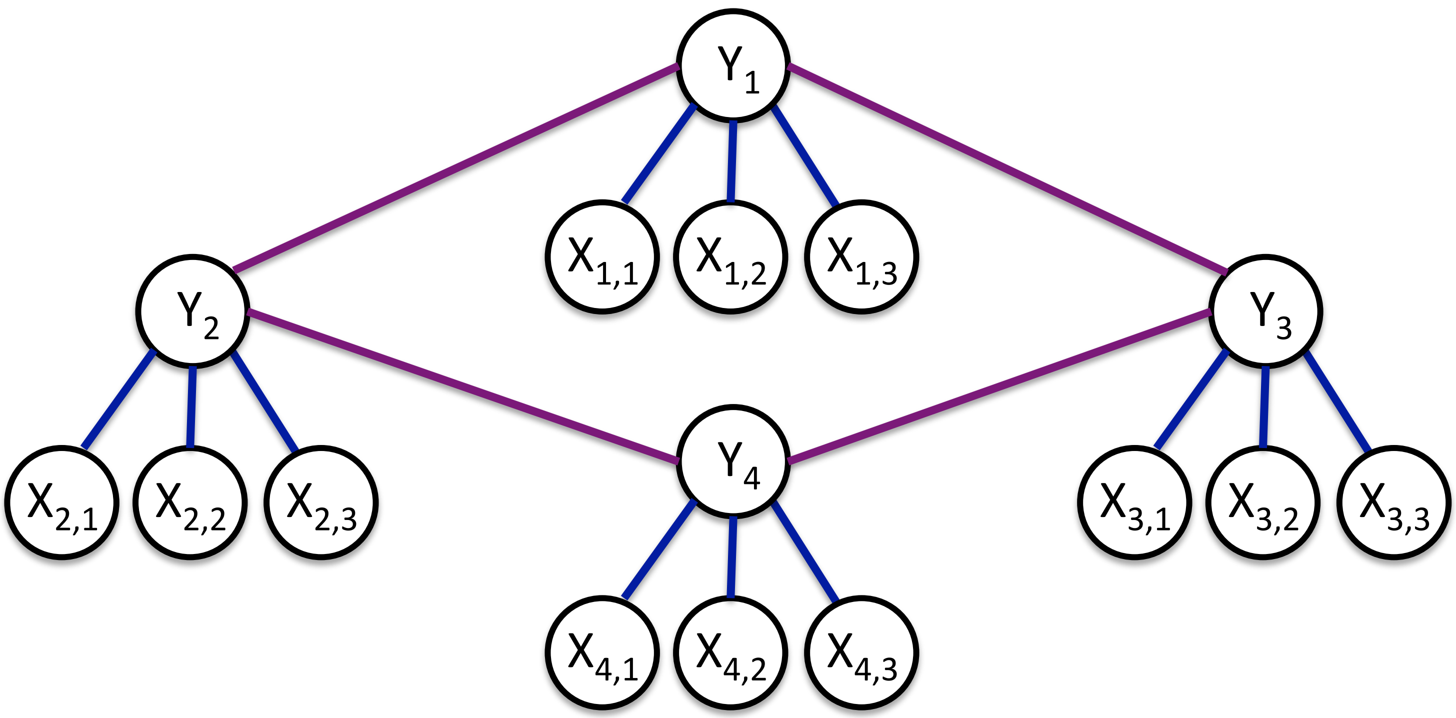


Image credit: [Chau&al06]

Problem Formulation

- Given: A graph with nodes, attributes, and edges. (e.g., web pages, words, and links.)
- Assume: Adversary can add or remove up to k attributes (e.g., words)



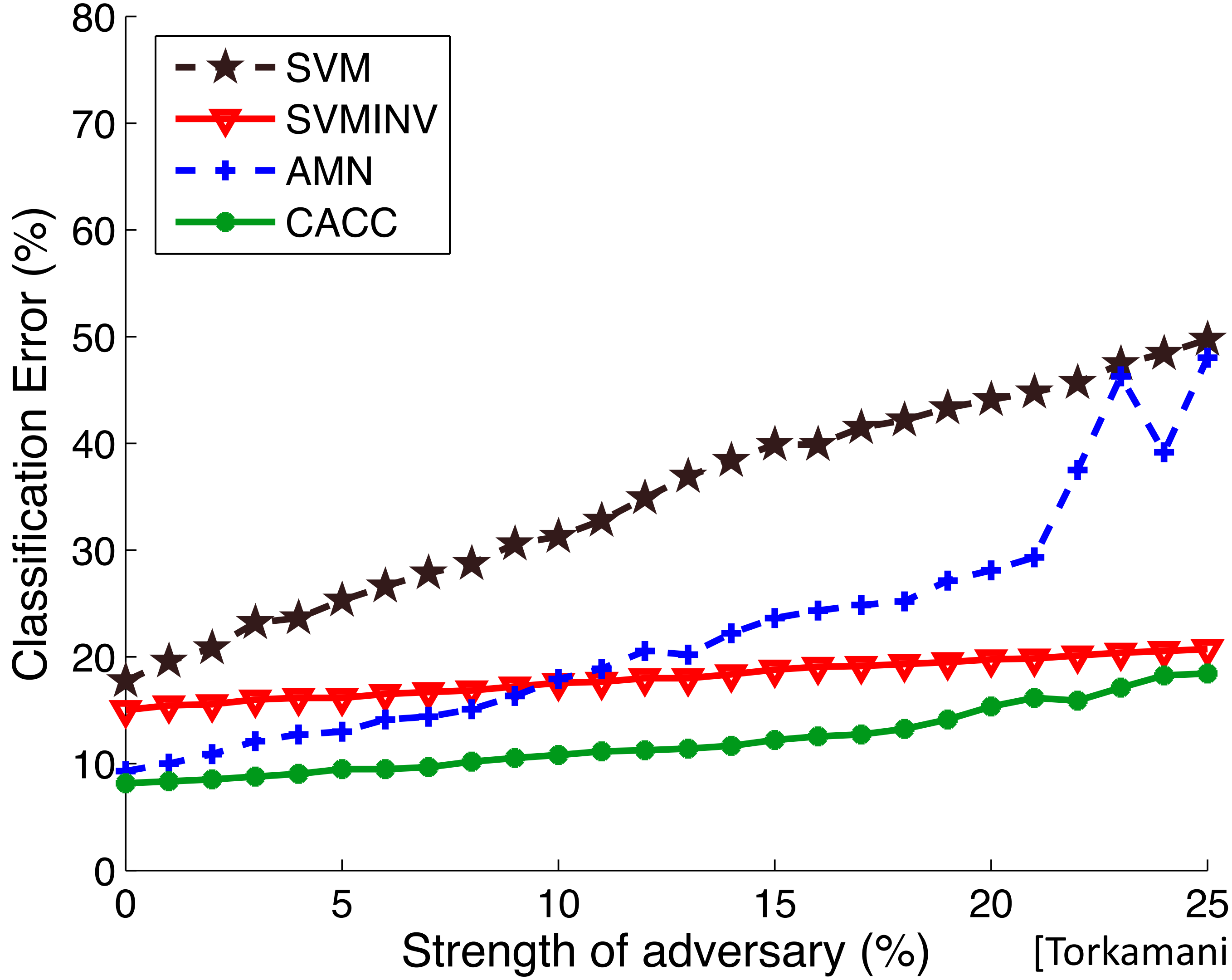
[Torkamani & Lowd, 2013]

Technical Approach

- Start with associative Markov networks, a special case of a structural SVM. [Taskar et al., 2004]
- Modify the quadratic program by "plugging in" the adversary's worst-case modification.
- Result: Optimal parameters in polynomial time (for an assumed model of the adversary).

[Torkamani & Lowd, 2013]

Results: Political Blogs (Tuned for 10% adversary)



[Torkamani & Lowd, 2013]

Adversarial Relational Learning: Ongoing Work

- Non-associative links
  (e.g., fraudsters and accomplices)
- Adversaries that add and remove links
  (e.g., link farms on the Web)
- Real-world evaluation with Web spam

Summary

- Machine learning is increasingly applied to security domains where adversaries will try to defeat it.
- To assess these new risks, we need a better understanding of ML vulnerabilities.
- To reduce these risks, we need more robust ML methods.