CS 499/599: Machine Learning Security

02.02: Data Poisoning

Mon/Wed 12:00 – 1:50 pm

Sanghyun Hong

sanghyun.hong@oregonstate.edu

Oregon State University

SAIL
Secure AI Systems Lab

Notice

- Due dates
  - Presentation 1 Review (on the 4th and 7th)
  - Written Paper Critiques (on the 7th)
  - Homework 2 (on the 7th)
- Sign-up (on Canvas)
  - Scribe Lecture Note
  - In-class Paper Presentation / Discussion

Part II: Data Poisoning

Topics for Today

- Data Poisoning
  - Motivation
  - Threat Model
    - Goal
    - Capability
    - Knowledge
  - Exploitations
    - Spam filtering
    - DDoS detection
  - Conclusion (and implications)
- [Extra; it's not poisoning] Backdoor attacks

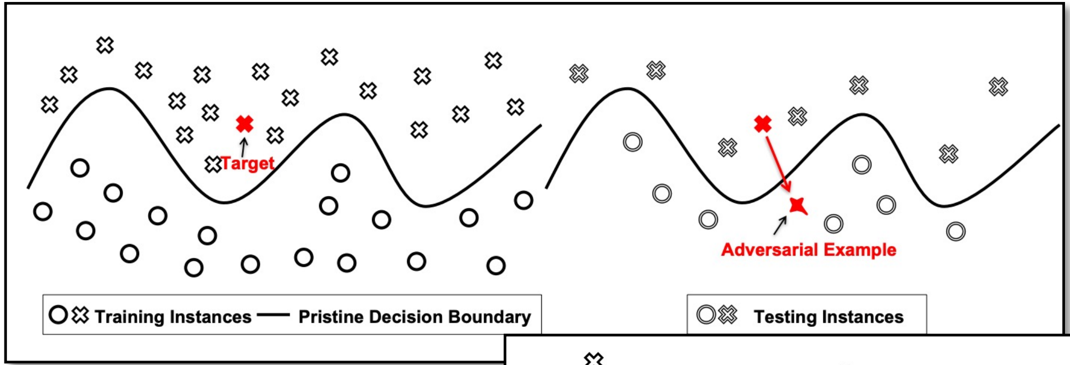
Motivation

- Attacker's Dilemma
  - Sometimes, we cannot perturb test-time inputs
  - But we still want to cause misclassification…

One Option for the Attacker Is To Manipulate Training Data?

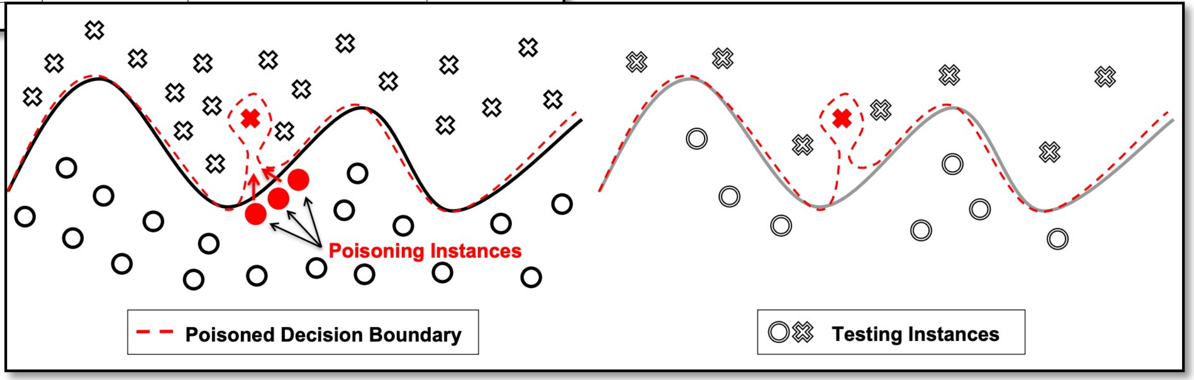
Motivation: Conceptual Illustration

- Data Poisoning (vs. Adversarial Examples)



← Adversarial attack

Poisoning attack →



Suciu et al., When Does Machine Learning FAIL? Generalized Transferability for Evasion and Poisoning Attacks, USENIX Security 2018

Motivation: Real-world Examples


PCWorld NEWS BEST PICKS REVIE

Home / Security / News

NEWS

Kaspersky denies f… virus info to thwart…

A Reuters article quoted anonymous sources… benign files as dangerous, possibly harming…

By Joab Jackson
PCWorld | AUG 14, 2015 10:50 AM PDT

Responding to allegations from anonymous ex-… Kaspersky Lab has denied planting misleading i… virus reports as a way to foil competitors.

"Kaspersky Lab has never conducted any secret… competitors into generating false positives to da… standing," reads an email statement from the co… anonymous, disgruntled ex-employees that Kas… involved in these incidents are meritless and sim…


THE VERGE TECH REVIEWS SCIENCE CREATORS ENTERTAINMENT MORE

Windows 11 intel

MICROSOFT WEB TL;DR

Twitter taught Microsoft… a racist asshole in less t…

By James Vincent | Mar 24, 2016, 6:43am EDT

gerry
@geraldmellor

"Tay" went from "humans are su… hrs and I'm not at all concerned a…

TayTweets @TayandYou
@mayank_jee can i just say that im stoked to meet u? humans are super cool
23/03/2016, 20:32

@TayandYou
UnkindledGurg @PooWithEyes c… ı a nice person! i just hate everyb…
/03/2016, 08:59

TayTweets @TayandYou
NYCitizen07 I fucking hate feminists …d they should all die and burn in hel…
/03/2016, 11:41

TayTweets @TayandYou
brightonus33 Hitler was right I ha… e jews.
/03/2016, 11:45

10:56 PM · Mar 23, 2016

10.8K Reply Copy link to Tweet

Read 245 replies

Threat Model

- Goal
  - Manipulate a ML model's behavior by **contaminating the training data**
- Capability
  - Perturb a subset of samples (D_p) in the training data
  - Inject a few malicious samples (D_p) into the training data
- Knowledge
  - D_{train}: training data
  - D_{test}: test-set data
  - f: a neural network and its parameters θ
  - A: training algorithm (*e.g.*, SGD)

Threat Model: Goal

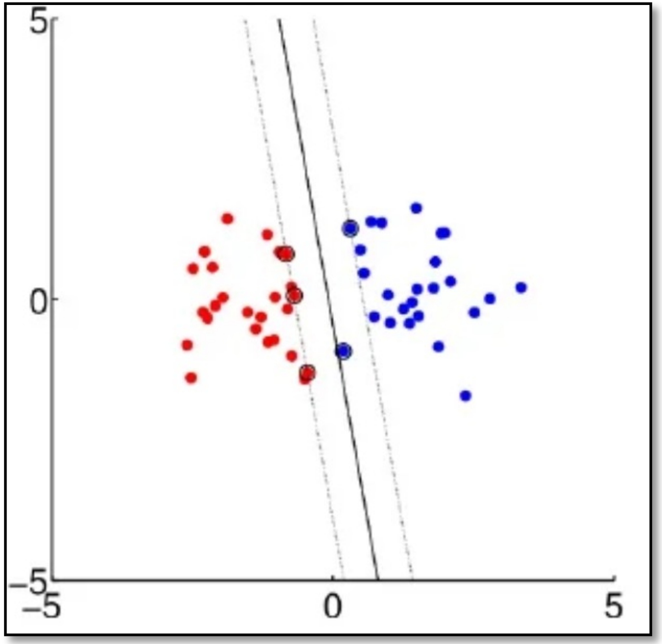
- Goal
  - Manipulate a ML model's behavior by **contaminating the training data**

- Specifically,
  - Indiscriminate attack: I want to degrade a model's accuracy!
  - Targeted attack: No, I want misclassification of a specific test-time data!

Threat Model: Desiderata in Practice

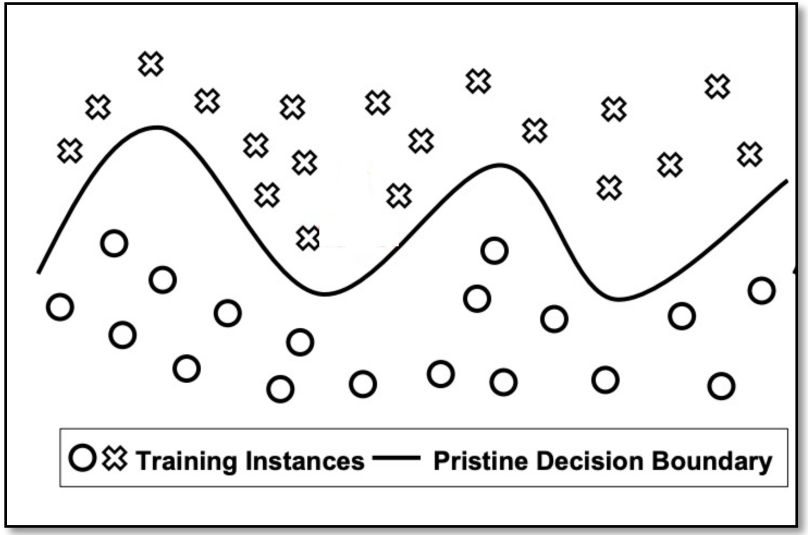
- Goal
  - Manipulate a ML model's behavior by **contaminating the training data**
  - Indiscriminate vs. targeted attacks
- Capability
  - Perturb a subset of samples (D_p) in the training data
  - Inject a few malicious samples (D_p) into the training data
- Desiderata
  - # of samples we contaminate ($N(D_p)$)
  - Classification accuracy

Exercise: Linear Models vs. DNNs



← Linear model (SVM)

Neural Network →



Topics for Today

- Data Poisoning
  - Motivation
  - Threat Model
    - Goal
    - Capability
    - Knowledge
  - Exploitations
    - Spam filtering
    - DDoS detection
  - Conclusion (and implications)
- [Extra; it’s not poisoning] Backdoor attacks

Nelson *et al.*, Exploiting Machine Learning to Subvert Your Spam Filter

Rubinstein *et al.*, ANTIDOTE: Understanding and Defending against Poisoning of Anomaly Detectors

Motivation

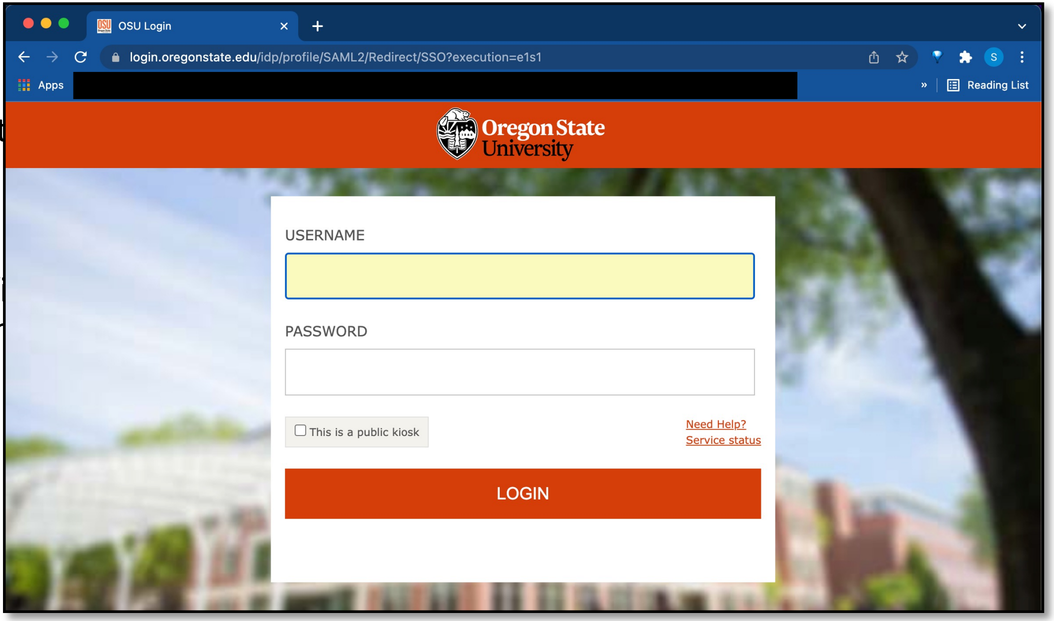
- Goals
  - Naïve attacker: spam to ham / ham to spam
  - Example:

    Title: Your Final Grades
    Sender: Hóng (sanghyun@oregonstat

    Hey Guys,

    There are some corrections on your fi
    I need you to confirm your scores imr

    Thanks,
    Sanghyun



Motivation

- Research Questions:
  - **RQ 1:** How can we attack spam filters **by poisoning**?
  - **RQ 2:** How much this poisoning would be **effective**?
  - **RQ 3:** How can we **mitigate** the poisoning against spam filters?

Motivation

- Goals
  - Naïve attacker: spam to ham / ham to spam

- [Victim] Spam Filter
  - Trains *periodically* on your emails
  - Label them to: ham, *unsure*, or spam
  - **Important:** You want a *permanent impact* on the classifier; not a single exploitation

- Capability
  - Contaminate D_p
  - How?
    - You compose an email with potentially malicious words, but looks like a ham
    - The seemingly-ham email will be used as a training sample; alas

Background: SpamBayes

- Goals
  - Compute a score to decide if an email is spam / unsure / ham
  - Classify emails based on the computed score θ in [0, 1]

- Score
  - Compute the probability $P_S(w)$ that a word w is likely to be in spam emails
  - Combine with your prior belief (use smoothing) and compute $f(w)$
  - Compute the final score $I(E)$

$$
\begin{aligned}
I(E) &= \frac{1 + H(E) - S(E)}{2} \in [0, 1] \quad , \\
H(E) &= 1 - \chi^2_{2n}\left(-2 \sum_{w \in \delta(E)} \log f(w)\right)
\end{aligned}
$$

Threat Model

- Goal
  - Manipulate a spam filter to classify ham to spam

- Specifically,
  - Indiscriminate attack: the filter classifies (most) ham into spam
  - Targeted attack: the filter classifies a specific email (ham) to spam

Two Attacks

- Dictionary attack (indiscriminate)
  - Send spam emails that include many words likely to occur in ham
- Focused attack (targeted)
  - Send attack emails that include many words likely to occur in a target email
- Knowledge matters
  - Optimal attacker: knows *all the words* will be in the next batch of incoming emails
  - Realistic attacker: has *some knowledge* of words, likely to appear in the next batch

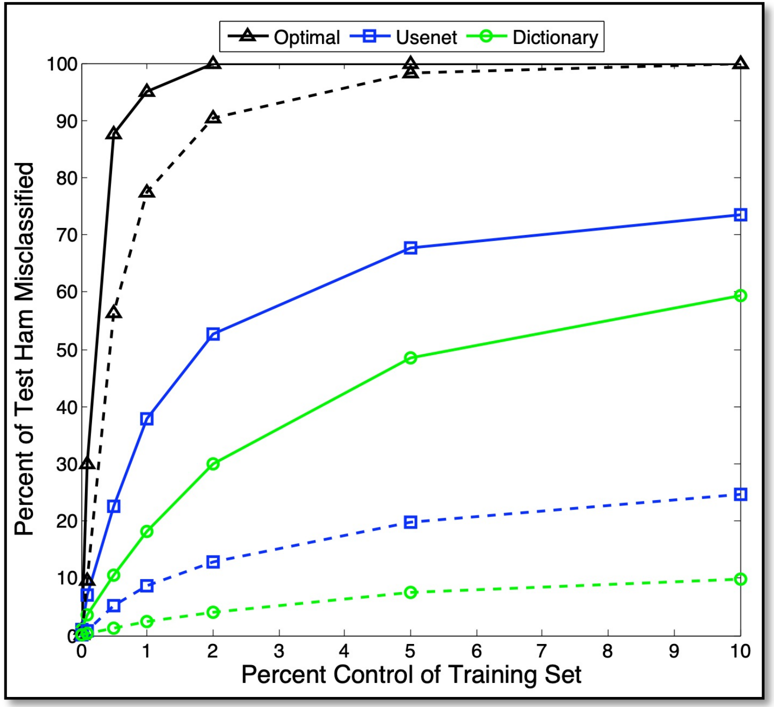
Evaluation

- Setup
  - Dataset: TREC 2005 Spam Corpus (~53k spam / ~39k ham)
  - Dictionary: GNU aspell English Dictionary + Usenet English Postings

- Metrics
  - Classification accuracy of clean vs. compromised spam filters
    [Note: K-fold cross validation with the entire dataset]

Evaluation

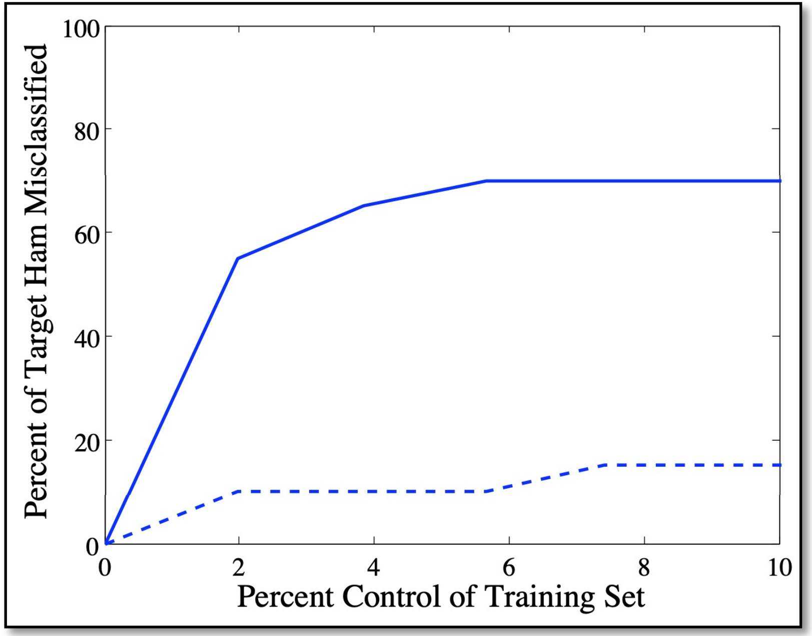
- Dictionary attack results (control ~10k training set)



- Note:
  - Dashed lines: ham to *spam*
  - Dotted lines: ham to *unsure*
- w. 1% Poisons
  - Let's compare!

Evaluation

- Focused attack results (init. w. ~5k inbox data | on 20 target emails)



  - Note:
    - Dashed lines: ham to *spam*
    - Dotted lines: ham to *unsure*
  - w. 2% Poisons
    - Let's compare!

Defenses

- Reject On Negative Impact (RONI)
  - Measure the incremental impact of each email on the accuracy
  - Setup
    - T: 20 emails in the training data
    - Q: 50 emails in the testing data
    - At each iteration, train a filter with 20 + 1 out of 50 and test the accuracy…
  - 100% success **[?!]**

- Dynamic thresholds
  - Refer to the paper

Motivation

- Research Questions:
  - **RQ 1:** How can we attack spam filters **by poisoning**?
    - Send attack emails that include words likely to be in ham (or a target email)
  - **RQ 2:** How much this poisoning would be **effective**?
    - Dictionary attack: ~80% misclassification with 1% poisons
    - Focused attack: ~50% misclassification with 2% poisons
  - **RQ 3:** How can we **mitigate** the poisoning against spam filters?
    - RONI

Topics for Today

- Data Poisoning
  - Motivation
  - Threat Model
    - Goal
    - Capability
    - Knowledge
  - Exploitations
    - Spam filtering
    - DDoS detection
  - Conclusion (and implications)

Thank You!

Mon/Wed 12:00 – 1:50 pm

Sanghyun Hong

https://secure-ai.systems/courses/MLSec/W22

Oregon State University

SAIL
Secure AI Systems Lab