Outline

1. Examples

2. Adversarial and Robust Learning
   - Attack
   - Defense

3. Other Work

4. Take Home Messages
Outline

1 Examples

2 Adversarial and Robust Learning
   Attack
   Defense

3 Other Work

4 Take Home Messages
Spam Filtering

Three steps of machine learning: representation → optimization → evaluation

1. Incoming mail
2. Feature extraction
3. Feature vector
   - [1.5, 2.5, ..., 2.4, 9.8]
4. Spam filter
   (A trained classifier)
5. Labeled mails
   (training set)
6. Evaluation
Spam Disguise
Adding noise to the junk mail

Nnus 🌐 yang.fong@msa.hinet.net
to wzg0770

Why is this message in Spam? We've found that lots of messages from yang.fong@msa.hinet.

tyx 您 nip 好：qdk

mzx 附 k 件 cp 是 vi 课 lqw 程 m 详 vip 细 hli 内 dq 容，rv 请 up 您 anc 参 u 阅！je0KWT5u

jygt 🌐 kkq@suibian.com
to me

⚠️ Why is this message in Spam? It's similar to messages that'

ポゆぢかプャケ 뱃صب Bộ

企业白领核心办公技能(PPT+Excel)企业高级应用2012
Spam Disguise
Introducing feature noise

Feature noise

Incoming mail → Feature extraction → Feature vector

Spam filter (A trained classifier)

Labeled mails (training set)

Disguised spam

legit

spam

[0.1, 5, ..., 2.4, 9.8]
Exploratory Attack: Spam Disguise in Practice

**Question**

Given a spam, how do you disguise it to evade from being detected?

**How?**

- Create dummy@gmail.com
- **Generate** disguised spams and send to dummy@gmail.com
- Select the **most desired** modification from the inbox.
Exploratory Attack: Spam Disguise in Practice

Question

Given a spam, how do you disguise it to evade from being detected?

How?

- Create dummy@gmail.com
- **Generate** disguised spams and send to dummy@gmail.com
- Select the **most desired** modification from the inbox.

Questions:

- What is the “most desired” mail?
- How to generate efficiently?
Training Data from Online Services

Users may vary in expertise, dedication and motivation.
Training Data from Online Services

Users may vary in expertise, dedication and motivation.
Training Data from Online Services
Users may vary in expertise, dedication and motivation

What if haters dominate?
Are they going to subvert the learning algorithm?
How to recover the unbiased labels/ratings?
Causative Attack: Poisoning the Spam Filter
Introducing label noise to training data
Outline

1. Examples

2. Adversarial and Robust Learning
   - Attack
   - Defense

3. Other Work

4. Take Home Messages
Modeling the Adversary

### Adversarial settings

The adversary manipulates instances to mislead the decision of the classifier *in their favor.*

### Exploratory attack

- in the test phrase;
- disguise a malicious instance to evade from being detected;
- e.g. disguise a spam, mutate a virus.
Modeling the Adversary

Adversarial settings
The adversary manipulates instances to mislead the decision of the classifier in their favor.

Exploratory attack
- in the test phrase;
- disguise a malicious instance to evade from being detected;
- e.g. disguise a spam, mutate a virus.

Causative attack
- in the training phrase;
- manipulate the training set to subvert the learning process;
- e.g. poisoning the spam filter, unfair rating on SNS.
(1) Because social network and crowdsourcing platform (e.g. Amazon mechanical turk) are popular.
Why Adversarial Learning is Interesting?

(1) Because social network and crowdsourcing platform (e.g. Amazon mechanical turk) are popular.

(2) Know your enemies and yourself, you will not be imperiled in a hundred battles.
–Sun Tzu, _The Art of War_, 544 BC
Why Adversarial Learning is Interesting?

(1) Because social network and crowdsourcing platform (e.g. Amazon mechanical turk) are popular.

(2) Know your enemies and yourself, you will not be imperiled in a hundred battles.
   –Sun Tzu, *The Art of War*, 544 BC

Secure learning

The ultimate goal is to develop robust learning algorithms, which are resilient to the adversarial noise.
Research Directions

Optimal attack strategies knowing the worst-case performance

- Exploratory Attack
- Causative Attack

Robust learning algorithms improving the worst-case performance

- Learning from crowds
## Binary Classification

Formalize the problem in math

<table>
<thead>
<tr>
<th>Term</th>
<th>Notation</th>
<th>Real world</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input space</td>
<td>$\mathcal{X} \subseteq \mathbb{R}^D$</td>
<td>All possible mails</td>
</tr>
<tr>
<td>Response space</td>
<td>$\mathcal{Y} := {-1, 1}$</td>
<td>All possible labels</td>
</tr>
<tr>
<td>Instance</td>
<td>$\mathbf{x} \in \mathcal{X}$, i.e. a $D$-dimensional vector</td>
<td>A mail</td>
</tr>
</tbody>
</table>
## Binary Classification

Formalize the problem in math

<table>
<thead>
<tr>
<th>Term</th>
<th>Notation</th>
<th>Real world</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input space</td>
<td>$\mathcal{X} \subseteq \mathbb{R}^D$</td>
<td>All possible mails</td>
</tr>
<tr>
<td>Response space</td>
<td>$\mathcal{Y} := {-1, 1}$</td>
<td>All possible labels</td>
</tr>
<tr>
<td>Instance</td>
<td>$x \in \mathcal{X}$, i.e. a $D$-dimensional vector</td>
<td>A mail</td>
</tr>
<tr>
<td>Hypothesis space</td>
<td>$\mathcal{H}$</td>
<td>All possible filters</td>
</tr>
<tr>
<td>Classifier</td>
<td>$f : \mathcal{X} \to \mathcal{Y}, f \in \mathcal{H}$</td>
<td>A filter</td>
</tr>
</tbody>
</table>
## Binary Classification

Formalize the problem in math

<table>
<thead>
<tr>
<th>Term</th>
<th>Notation</th>
<th>Real world</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input space</td>
<td>$\mathcal{X} \subseteq \mathbb{R}^D$</td>
<td>All possible mails</td>
</tr>
<tr>
<td>Response space</td>
<td>$\mathcal{Y} := {-1, 1}$</td>
<td>All possible labels</td>
</tr>
<tr>
<td>Instance</td>
<td>$x \in \mathcal{X}$, i.e. a $D$-dimensional vector</td>
<td>A mail</td>
</tr>
<tr>
<td>Hypothesis space</td>
<td>$\mathcal{H}$</td>
<td>All possible filters</td>
</tr>
<tr>
<td>Classifier</td>
<td>$f : \mathcal{X} \rightarrow \mathcal{Y}$, $f \in \mathcal{H}$</td>
<td>A filter</td>
</tr>
<tr>
<td>Positive set</td>
<td>$\mathcal{X}^+ := { x \in \mathcal{X} \mid f(x) = +1 }$</td>
<td>All possible spams</td>
</tr>
<tr>
<td>Negative set</td>
<td>$\mathcal{X}^- := { x \in \mathcal{X} \mid f(x) = -1 }$</td>
<td>All possible legit mails</td>
</tr>
</tbody>
</table>
## Binary Classification

Formalize the problem in math

<table>
<thead>
<tr>
<th>Term</th>
<th>Notation</th>
<th>Real world</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input space</td>
<td>$\mathcal{X} \subseteq \mathbb{R}^D$</td>
<td>All possible mails</td>
</tr>
<tr>
<td>Response space</td>
<td>$\mathcal{Y} := {-1, 1}$</td>
<td>All possible labels</td>
</tr>
<tr>
<td>Instance</td>
<td>$x \in \mathcal{X}$, i.e. a $D$-dimensional vector</td>
<td>A mail</td>
</tr>
<tr>
<td>Hypothesis space</td>
<td>$\mathcal{H}$</td>
<td>All possible filters</td>
</tr>
<tr>
<td>Classifier</td>
<td>$f : \mathcal{X} \rightarrow \mathcal{Y}, f \in \mathcal{H}$</td>
<td>A filter</td>
</tr>
<tr>
<td>Positive set</td>
<td>$\mathcal{X}^+ := {x \in \mathcal{X} \mid f(x) = +1}$</td>
<td>All possible spams</td>
</tr>
<tr>
<td>Negative set</td>
<td>$\mathcal{X}^- := {x \in \mathcal{X} \mid f(x) = -1}$</td>
<td>All possible legit mails</td>
</tr>
<tr>
<td>Loss function</td>
<td>$V : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_{0+}$</td>
<td>Cost of misclassification</td>
</tr>
</tbody>
</table>
Training the Classifier
Solving an optimization problem

Classification
Given a training set \( S := \{(x_i, y_i) | x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}_{i=1}^n \). Find the classifier \( f_S \in \mathcal{H} \) that performs best on some test set \( T \).

Solving an optimization problem:

\[
f_S := \arg\min_f \gamma \sum_{i=1}^{n} V(y_i, f(x_i)) + \|f\|_\mathcal{H}^2,
\]

where \( \gamma \in \mathbb{R}_{0+} \) is a fixed parameter for quantifying the trade off.
Disguise a spam from being detected by a filter. Be efficient.

Problem Formulation

Given

- a trained classifier \( f \);
- a positive (malicious) instance \( x^A \in \mathcal{X}^+ \);
- a random negative (benign) instance \( x^- \in \mathcal{X}^- \).

Find an instance \( x^* \in \mathcal{X}_{f^-} \) such that

- \( x^* \) should be similar to \( x^A \);
- issuing as few queries to \( f \) as possible.
### Assumptions

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Real world</th>
</tr>
</thead>
<tbody>
<tr>
<td>Know the dimension of $\mathcal{X}$</td>
<td>Know how many features</td>
</tr>
<tr>
<td>Attack a fixed $f$</td>
<td>Spam filter is not updated</td>
</tr>
<tr>
<td>Observe $f(x)$ by a membership query</td>
<td>Observe the label of a sent mail</td>
</tr>
<tr>
<td>Design a cost function</td>
<td>Know the cost of misclassification</td>
</tr>
</tbody>
</table>

---

Han, Thomas, Claudia. *Evasion Attack of Multi-Class Linear Classifiers* *PAKDD 2012*
Exploratory Attack as $\ell_p$-norm Minimization

Exploratory Attack

Given $x^A$, $f$, and a cost function $g : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_{0+}$, solve

$$\min_{x} g(x, x^A) \quad \text{subject to} \quad x \in \mathcal{X}^-,$$

where $\mathcal{X}^-$ is specified by the membership oracle $f$.

For example, $g(x, x^A) := \|x - x^A\|_{\ell_1}$

Han, Thomas, Claudia. *Evasion Attack of Multi-Class Linear Classifiers* PAKDD 2012
Illustration of the Problem

\( g(x) := \|x - x^A\|_{\ell_1} \)
Illustration of the Problem

\[ g(x) := \|x - x^A\|_1 \]
Face Camouflage

Considering a suspect tries to disguise herself as innocent.

Han, Thomas, Claudia. *Evasion Attack of Multi-Class Linear Classifiers* PAKDD 2012
Label Flips Attack

Given a training set, the adversary contaminates the training data through flipping labels.

Han, Huang, Claudia. Adversarial Label Flips Attack on Support Vector Machines ECAI 2012
Adversarial Label Flips Attack

Find a combination of label flips under a given budget so that a classifier trained on such data will have maximal classification error on some test data.
Adversarial Label Flips Attack

Find a combination of label flips under a given budget so that a classifier trained on such data will have maximal classification error on some test data.

Training set: \( S := \{(x_i, y_i) \mid x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}_{i=1}^n \);
Indicator: \( z_i \in \{0: \text{normal}, 1: \text{flipped}\}, i = 1, \ldots, n; \)
Flipping cost: \( c_i \in \mathbb{R}_{0+}, i = 1, \ldots, n; \)
Adversarial Label Flips Attack

Find a combination of label flips under a given budget so that a classifier trained on such data will have maximal classification error on some test data.

Training set: \( S := \{(x_i, y_i) \mid x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}_{i=1}^n \);

Indicator: \( z_i \in \{0: \text{normal}, 1: \text{flipped}\}, i = 1, \ldots, n; \)

Flipping cost: \( c_i \in \mathbb{R}_{0+}, i = 1, \ldots, n; \)

Tainted label: \( y'_i := y_i (1 - 2z_i); \)

Tainted training set: \( S' := \{(x_i, y'_i)\}. \)
A Bilevel Formulation

Finding the optimal label flips

Given $S$, a test set $T$ and a budget $C$, solve

$$\max_z \sum_{(x,y) \in T} V(y, f_{S'}(x)),$$

s.t. $f_{S'} \in \arg \min_f \gamma \sum_{i=1}^{n} V(y'_i, f(x_i)) + \|f\|_H^2,$

$$\sum_{i=1}^{n} c_i z_i \leq C, \quad z_i \in \{0, 1\}.$$
Finding the optimal label flips

Given $S$, a test set $T$ and a budget $C$, solve

$$\max_{z} \sum_{(x, y) \in T} V(y, f_{S'}(x)),$$

subject to

$$f_{S'} \in \arg\min_f \gamma \sum_{i=1}^{n} V(y_i', f(x_i)) + \|f\|_H^2,$$

$$\sum_{i=1}^{n} c_i z_i \leq C, \quad z_i \in \{0, 1\}.$$
A Bilevel Formulation

Finding the optimal label flips

Given $S$, a test set $T$ and a budget $C$, solve

$$\max_{z} \sum_{(x,y) \in T} V(y, f_{S'}(x)),$$

s.t.  
$$f_{S'} \in \arg \min_{f} \gamma \sum_{i=1}^{n} V(y_i', f(x_i)) + \|f\|_H^2,$$

$$\sum_{i=1}^{n} c_i z_i \leq C, \quad z_i \in \{0, 1\}.$$
**Label Flips Attack Against SVM**

Train: 100, flip:20, test 800

汉, 黄, 克劳迪娅. *Adversarial Label Flips Attack on Support Vector Machines* ECAI 2012
An Endless Game between Adversary and Defender

Escher. Drawing Hands 1948
Detecting Exploratory Attack
Convergence pattern

Han, Thomas, Claudia. Evasion Attack of Multi-Class Linear Classifiers PAKDD 2012
Subjective opinions from crowds
Learning objective assessment from subjective opinions

Han, Huang, Claudia. Learning from Multiple Observers with Unknown Expertise PAKDD 2013
Meyyar. Leveraging the Wisdom of Crowds for Reputation Management Master’s thesis
Subjective opinions from crowds
Learning objective assessment from subjective opinions

Han, Huang, Claudia. Learning from Multiple Observers with Unknown Expertise PAKDD 2013
Meyyar. Leveraging the Wisdom of Crowds for Reputation Management Master’s thesis
Subjective opinions from crowds
Learning objective assessment from subjective opinions

Han, Huang, Claudia. Learning from Multiple Observers with Unknown Expertise PAKDD 2013
Meyyar. Leveraging the Wisdom of Crowds for Reputation Management Master’s thesis
Subjective opinions from crowds
Learning objective assessment from subjective opinions

What’s wrong with “majority vote” and “take average”?
They completely ignore the individual expertise and may fail in the settings with non-Gaussian or adversarial noise!

Han, Huang, Claudia. *Learning from Multiple Observers with Unknown Expertise* PAKDD 2013
Meyyar. *Leveraging the Wisdom of Crowds for Reputation Management* Master’s thesis
Unreliable readings from sensors

Questions

1. How to integrate readings from multiple sensors?
2. How accurate is each sensor?

Han, Huang, Claudia. *Learning from Multiple Observers with Unknown Expertise* PAKDD 2013
Meyyar. *Leveraging the Wisdom of Crowds for Reputation Management* Master’s thesis
Learning from Multiple Observers

Problems

- How to learn a regression function to predict the ground truth \textit{precluding} the prior knowledge of observers?
- How to estimate the expertise of each observer \textit{without} knowing the ground truth?
Intuition behind
Leveraging the neighborhood information

Han, Huang, Claudia. Learning from Multiple Observers with Unknown Expertise PAKDD 2013
Meyyar. Leveraging the Wisdom of Crowds for Reputation Management Master’s thesis
Intuition behind
Leveraging the neighborhood information

Han, Huang, Claudia. Learning from Multiple Observers with Unknown Expertise PAKDD 2013
Meyyar. Leveraging the Wisdom of Crowds for Reputation Management Master’s thesis
Intuition behind
Leveraging the neighborhood information

Han, Huang, Claudia. Learning from Multiple Observers with Unknown Expertise PAKDD 2013
Meyyar. Leveraging the Wisdom of Crowds for Reputation Management Master’s thesis
Intuition behind
Leveraging the neighborhood information

\[ f(x_1) \]
\[ f(x_2) \]
\[ f(x_3) \]

Instance space \( \mathcal{X} \)

Groundtruth space \( Z \)

(Latent) 

\[ g_1(z_3) \]
\[ g_1(z_2) \]
\[ g_M(z_3) \]
\[ g_M(z_1) \]
\[ g_M(z_2) \]

Han, Huang, Claudia. Learning from Multiple Observers with Unknown Expertise PAKDD 2013

Meyyar. Leveraging the Wisdom of Crowds for Reputation Management Master’s thesis
Intuition behind
Leveraging the neighborhood information

Han, Huang, Claudia. Learning from Multiple Observers with Unknown Expertise PAKDD 2013
Meyyar. Leveraging the Wisdom of Crowds for Reputation Management Master’s thesis
Nonparametric probabilistic model

\[ p(Y, Z, X) = p(Z | X)p(Y | Z, X)p(X). \]

Gaussian process: a less-parametric approach for modeling a function.
Nonparametric probabilistic model

$p(Y, Z, X) = p(Z | X)p(Y | Z, X)p(X)$.  

Gaussian process: a less-parametric approach for modeling a function. 
Maximizing the posterior, which gives

$$\log p(Z, \Theta | Y, X) = \log p(Y | Z, X, \Theta) + \log p(Z | X, \Theta) + \text{constant}.$$  

Deriving the gradient w.r.t. $z, \kappa, \phi, \eta$, respectively.  
Feed the gradients to L-BFGS method for finding the stationary point.
1-D example

Groundtruth function: \( f(t) = 10 \sin(6t) \sin\left(\frac{t}{2}\right) \),
Groundtruth function: \( f(t) = 10 \sin(6t) \sin\left(\frac{t}{2}\right) \),
Randomly sample responses at \( t \in [0, 6] \) from four sensors.
1-D example

Groundtruth function: $f(t) = 10 \sin(6t) \sin \left( \frac{t}{2} \right)$.
Randomly sample responses at $t \in [0, 6]$ from four sensors.
1-D example

Groundtruth function: $f(t) = 10 \sin(6t) \sin\left(\frac{t}{2}\right)$,
Randomly sample responses at $t \in [0, 6]$ from four sensors.
Groundtruth function: \( f(t) = 10 \sin(6t) \sin\left(\frac{t}{2}\right) \),
Randomly sample responses at \( t \in [0, 6] \) from four sensors.
1-D example

Groundtruth function: $f(t) = 10 \sin(6t) \sin\left(\frac{t}{2}\right)$,
Randomly sample responses at $t \in [0, 6]$ from four sensors.
1-D example

What do we know?

*Only* the readings from each sensor
1-D example

What do we want to know?
1. Groundtruth function, i.e. $f(t)$.
2. Response function of each sensor.
Synthetic data set

Recover $f(t) = 10 \sin(6t) \sin\left(\frac{t}{2}\right)$ and $g_1, g_2, g_3, g_4$. 

Han, Huang, Claudia. *Learning from Multiple Observers with Unknown Expertise* PAKDD 2013
Meyyar. *Leveraging the Wisdom of Crowds for Reputation Management* Master’s thesis
**Synthetic data set**

Recover \( f(t) = 10 \sin(6t) \sin\left(\frac{t}{2}\right) \) and \( g_1, g_2, g_3, g_4 \).
Synthetic data set

Recover $f(t) = 10 \sin(6t) \sin\left(\frac{t}{2}\right)$ and $g_1, g_2, g_3, g_4$. 

Han, Huang, Claudia. *Learning from Multiple Observers with Unknown Expertise* PAKDD 2013
Meyyar. *Leveraging the Wisdom of Crowds for Reputation Management* Master’s thesis
TUM100² photo rating data set
Contributed by Huang http://ml.sec.in.tum.de/opars

Huang, Han, Claudia. **OPARS: Objective Photo Aesthetics Ranking System** (demo paper). *ECIR 2013*

Han, Huang, Claudia. **Learning from Multiple Observers with Unknown Expertise** *PAKDD 2013*

Meyyar. **Leveraging the Wisdom of Crowds for Reputation Management** *Master’s thesis*
Sparse rating matrix from 34 users
Results
Top-5 and bottom-5 ranked photos

Huang, Han, Claudia. **OPARS: Objective Photo Aesthetics Ranking System** (demo paper). *ECIR 2013*
Han, Huang, Claudia. **Learning from Multiple Observers with Unknown Expertise** *PAKDD 2013*
Meyyar. **Leveraging the Wisdom of Crowds for Reputation Management** *Master’s thesis*
Outline

1 Examples

2 Adversarial and Robust Learning
   Attack
   Defense

3 Other Work

4 Take Home Messages
Scalable Online Learning Algorithms
Incrementally finding a **good enough solution as fast as possible**

- Han, Claudia. *Lazy Gaussian Process Committee for Real-Time Online Regression*. *AAAI 2013*.

- Han, Claudia. *Efficient Online Sequence Prediction with Side Information*. Submitted to *ICDM 2013*.
Ubiquitous Anomaly Detection

Group

- Prof. Claudia Eckert
- Huang Xiao (Ph.D. student)
- Han Xiao (Ph.D. student)
- Chih-Ta Lin (visitor)
- Sami Ghawi (Master student)
- Meyyar Palaniappan (graduated)
- Fernando Hernandez Montoya (graduated)
- Siddhant Goel (graduated)
Outline

1. Examples
2. Adversarial and Robust Learning
   - Attack
   - Defense
3. Other Work
4. Take Home Messages
NOT SURE IF REAL DATA

OR ADVERSARIAL NOISE
I DON'T ALWAYS DEVELOP NEW ALGORITHM

BUT WHEN I DO, I MAKE IT ROBUST
MACHINE LEARNING IS ALL ABOUT OPTIMIZATION