Adversarial Machine Learning
An ARA Perspective

David Ríos Insua

AXA-ICMAT Chair and Royal Academy

Aalto, March 2018

Sketch of problems at DataLab

- Adversarial risk analysis for security
- ARA for cybersecurity and cyberinsurance
- Adversarial machine learning

- Large scale network monitoring for safety and security
- Supply chain cyber risk management
- Affective decision making for emotional social robotics (education, therapist,…)
- Personalised geomarketing
- Improving the national security risk assessment (risk matrices, scenario analysis)
- The effect of physical exercise on health outcomes

Outline

- ***(Almost) All things adversarial***
- Adversarial statistical decision theory
- Adversarial classification
- Discussion and challlenges

Adversarial problems

- Stats/ML: Standard problems
  - Point Estimation
  - Hypothesis testing
  - Prediction
  - Classification
  - Regression
  - ...
- Many application areas (security, …) entail adversaries

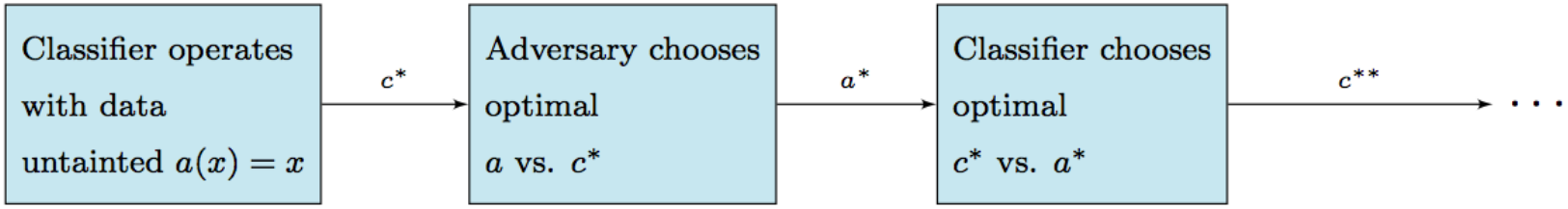
Intelligent attackers adapting their behaviour to remain undetected and obtain a benefit

  - Spam detection
  - Fraud detection
  - Network monitoring
  - Malware detection
  - Fake news detection
  - ….
- Comparatively few attempts to deal with the problem….
- ...Mostly modelled through noncooperative game theory

Adversarial classification as a game

- C, classifier. A, adversary
- Two classes: + malicious; - innocent.
- C and A maximise expected utility under common knowledge conditions
- Finding Nash equilibria extremely complex

- Dalvi et al (2004) propose a scheme



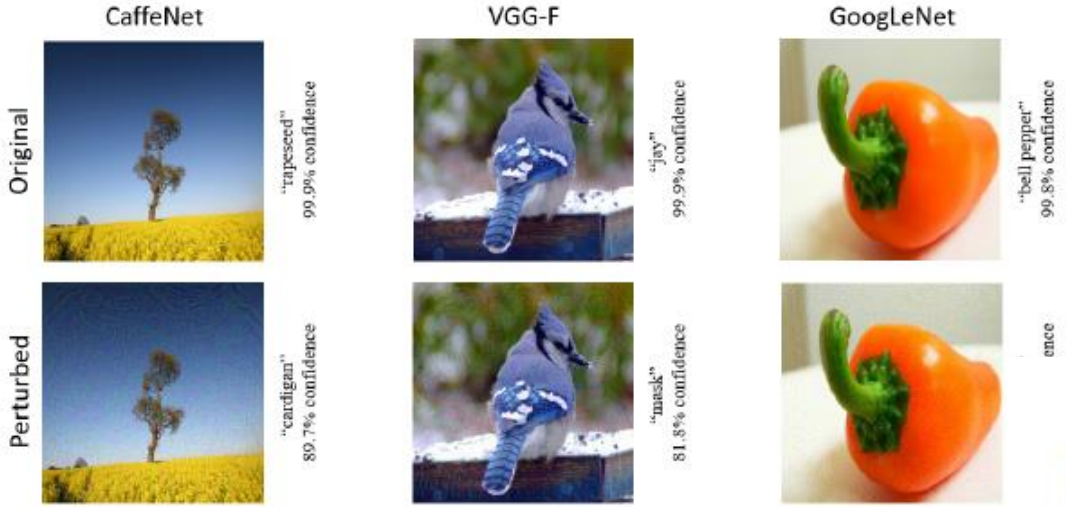
Utility sensitive Naive Bayes

Forward myopic approach under strong common knowledge

Adversarial problems

- Adversarial classification (Dalvi et al,…)
- Adversarial signal processing (Barni et al,..)
- Adversarial learning (Lowd and Meek,..)
- Adversarial machine learning (Tygar,..)
- Adversarial SVMs (Zhou et al,…)
- …
- Adversarial competitions in Kaggle

Adversarial problems



From RA to ARA…

Motivation

- RA extended to include adversaries ready to increase our risks

- S-11, M-11,.. lead to large security investments globally, some of them criticised
- Many modelling efforts to efficiently allocate such resources
- Parnell et al (2008) NAS review
  - Standard reliability/risk approaches not take into account intentionality
  - Game theoretic approaches. Common knowledge assumptions…
  - Decision analytic approaches. Forecasting the adversary action…
- Merrick, Parnell (2011) review approaches commenting favourably on ARA

ARA

- A framework to manage risks from actions of intelligent adversaries (DRI, Rios, Banks, JASA 2009)
- One-sided prescriptive support
  - Use a SEU model
  - Treat the adversary's decision as uncertainties
  - Bayesian games Kadane, Larkey (1982), Raiffa (1982, 2002) made operational
- Method(s) to predict adversary's actions
  - We assume the adversary is a *expected utility maximizer*
    - Model his decision problem
    - Assess his probabilities and utilities
    - Find his action of maximum expected utility

  (But other *descriptive* models are possible)
- Uncertainty in the Attacker's decision stems from
  - *our* uncertainty about his probabilities and utilities
  - this may lead to a hierarchy of nested decision problems

  (random, noninformative, level-k, heuristic, mirroring argument, under partial information,…) vs (common knowledge)

ARA: Examples/Cases

Problem	Defender	Attacker	Specificities	Template
ATC protection	Airport authority	Terrorist	Single site	D-> A
Piracy	Ship owner	Pirates	Single site	D- >A - > D
Metro	Operator	Pickpock Fare evasion	Multisite Multiattack, Cascade	D->A
Urban security	Police	Mob	Multisite spatial	D->A->D
Train	DoT, DoD	Terrorist	Multisite network	D->A->D
Reliability	Manufacturer	Customer	--	D->A
SME IS	Company	Competitor	Cyber, Integrated with RA	D->A
Oil rig cybercontrolled	Oil company	Sponsored hackers	Cyber, Multiattack	D->A->D
CI	Owner	Terrorist	Multistage	General
Cybersec res allocation+cybins CYBECO	IT Owner	Hacker(s)	Several decisions Random and targeted attacks	D-A, D-A-D
Social robots	Robot	User	Sequential	D->A

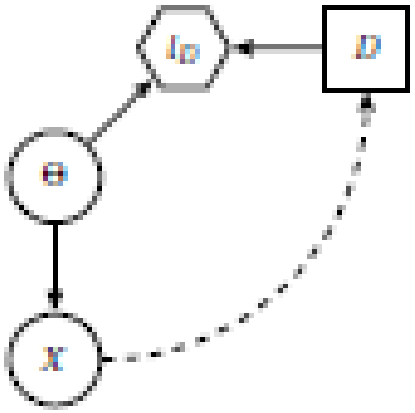
Other themes

- Different opponent models, beyond SEU
- Concept uncertainty (beyond aleatory, epistemic), Mixtures
- Robustness and ARA (GT, ARA, Robust ARA)
- Partial information and ARA
- Multiple attackers, Multiple defenders
- Differential games
- Competition and cooperation
- Efficient computational schemes
- Computational environment
- Cybersecurity and cyberinsurance. CYBECO
- …

Outline

- (Almost) All things adversarial
- ***Adversarial statistical decision theory***
- Adversarial classification
- Discussion and challlenges

Statistical Decision Theory



$$d^*(x) = \arg\min_d \int l_D(d,\theta)\, p_D(\theta \mid x)\, \mathrm{d}\theta$$

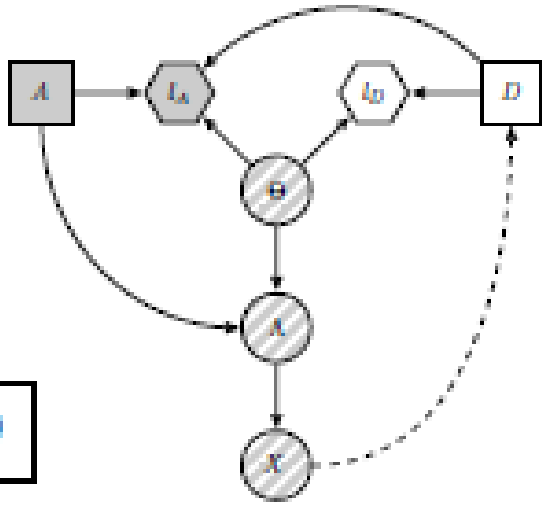
$$d^*(x) = \arg\min_d \int l_D(d,\theta)\, p_D(x \mid \theta)\, p_D(\theta)\, \mathrm{d}\theta$$

- Point estimation under quadratic loss

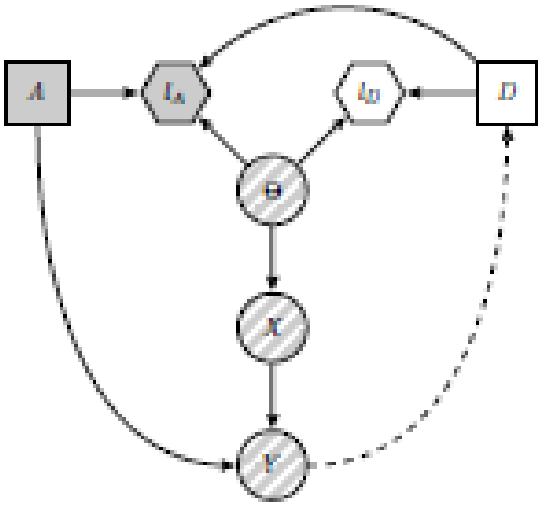
$$l_D(d,\theta) = (\theta - d)^2,$$

$$d^*(x) = \frac{1}{p_D(x)} \int \theta\, p_D(x \mid \theta)\, p_D(\theta)\, \mathrm{d}\theta = \int \theta\, p_D(\theta \mid x)\, \mathrm{d}\theta = E[\theta \mid x]$$

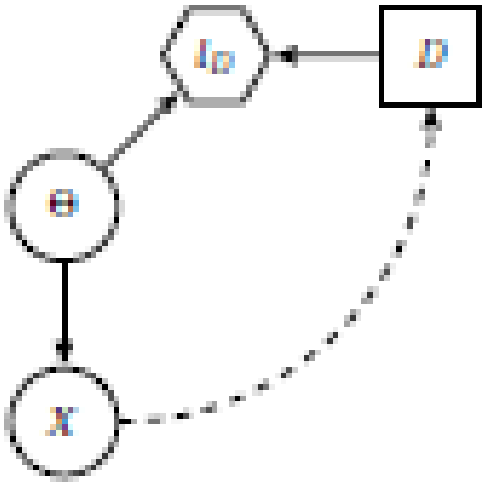
Adversarial Statistical Decision Theory

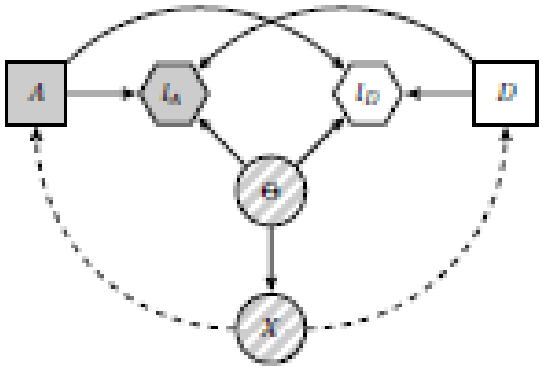


(a) Structural attacker



(b) Data-fiddler attacker



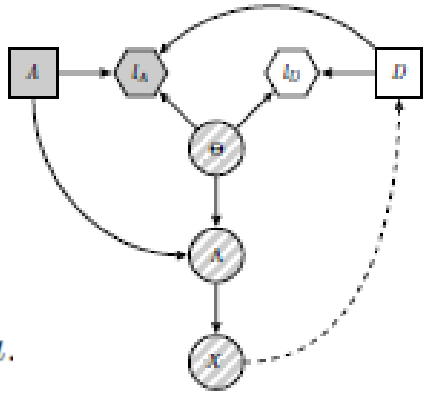


(c) Simultaneous ASDT problem

Adversarial point estimation

$$\lambda = a + \theta$$

- Quadratic loss



(a) Structural attacker

$$d^*(x) = \arg\min_d \iiint (\theta - d)^2 \, p_D(x \mid \lambda) \, p_D(\lambda \mid \theta, a) \, p_D(\theta) \, p_D(a) \, \mathrm{d}\lambda \, \mathrm{d}\theta \, \mathrm{d}a.$$

$$d^*(x) = \arg\min_d \iint (\theta - d)^2 \, p_D(x \mid \lambda = \theta + a) \, p_D(\theta) \, p_D(a) \, \mathrm{d}\theta \, \mathrm{d}a$$

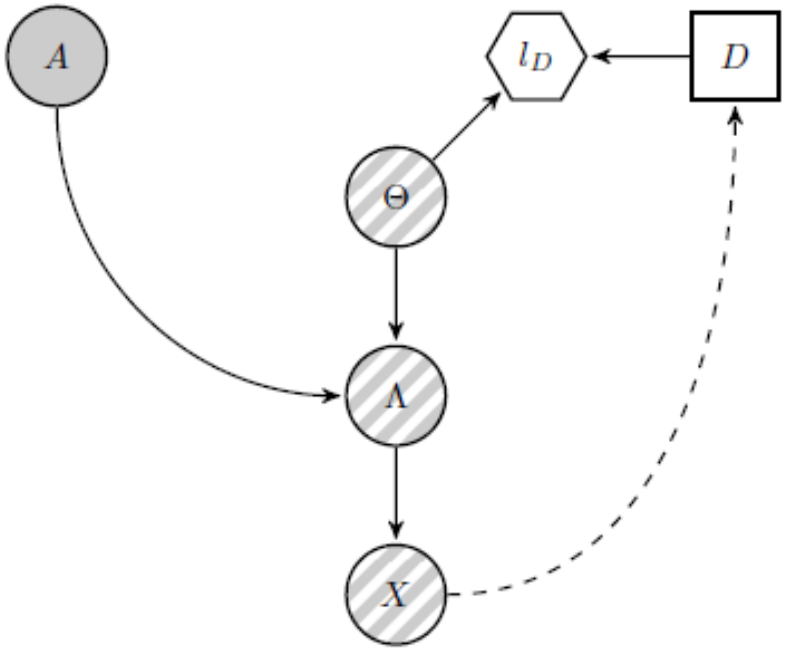
$$d^*(x) = \frac{1}{p_D(x)} \iiint \theta \, p_D(x \mid \lambda) \, p_D(\lambda \mid \theta, a) \, p_D(\theta) \, p_D(a) \, \mathrm{d}\lambda \, \mathrm{d}\theta \, \mathrm{d}a$$

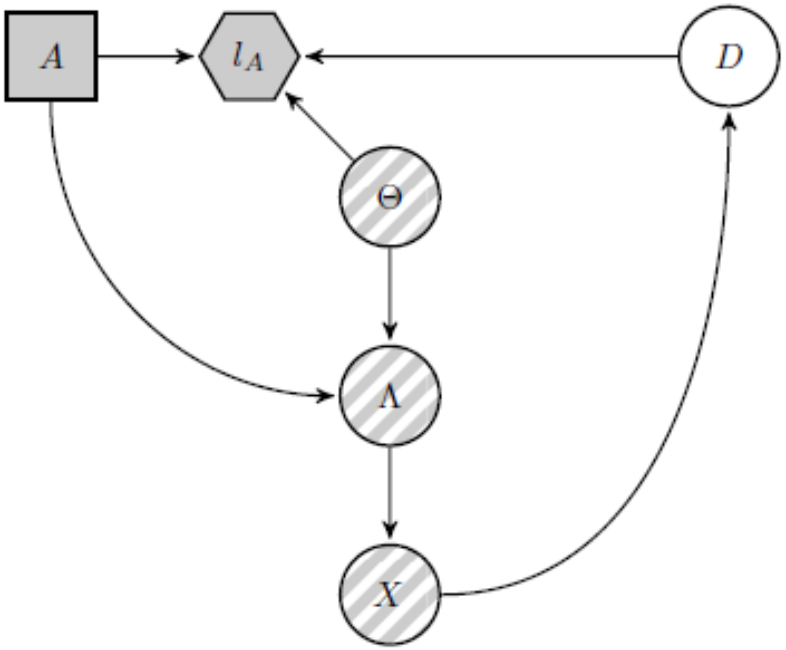
$$d^*(x) = \frac{1}{p_D(x)} \iint \theta \, p_D(x \mid \lambda) \, p_D(\lambda \mid \theta) \, p_D(\theta) \, \mathrm{d}\lambda \, \mathrm{d}\theta$$

$$= \frac{1}{p_D(x)} \int \theta \, p_D(x \mid \theta) \, p_D(\theta) \, \mathrm{d}\theta = \int \theta \, p_D(\theta \mid x) \, \mathrm{d}\theta = E[\theta \mid x]$$

$p_D(a)$

Adversarial point estimation





Concept uncertainty

Adversarial point estimation

- A Bayesian adversary

$$a_B^* = \arg\min_a \iiint l_A(d,a,\theta)\, p_A(d\,|\,x)\, p_A(x\,|\,\lambda=\theta+a)\, p_A(\theta)\,\mathrm{d}d\,\mathrm{d}x\,\mathrm{d}\theta.$$

$$A_B^* = \arg\min_a \iiint L_A(d,a,\theta)\, P_A(d\,|\,x)\, P_A(x\,|\,\lambda=\theta+a)\, P_A(\theta)\,\mathrm{d}d\,\mathrm{d}x\,\mathrm{d}\theta$$

$$p_D^B(a) = P(A_B^* = a).$$

$$A_{B,k}^* = \arg\min_a \iiint L_A^k(d,a,\theta)\, P_A^k(d\,|\,x)\, P_A^k(x\,|\,\lambda=\theta+a)\, P_A^k(\theta)\,\mathrm{d}d\,\mathrm{d}x\,\mathrm{d}\theta$$

$$\hat{p}_D^B(A=a) \approx \#\{A_{B,k}^* = a\}/K$$

- Mixture, e.g.

$$\pi_B\, \hat{p}_D^B(a) + \pi_M\, \hat{p}_D^M(a)$$

Adversarial point estimation

- Normal-normal model, for certain parameter choices

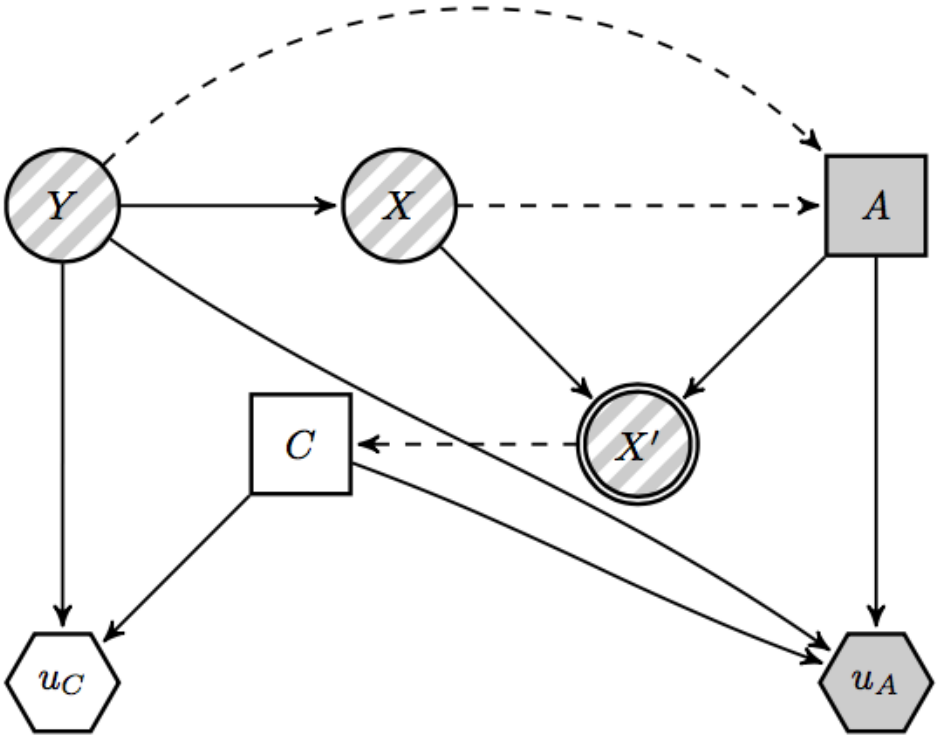
Solution Concept	Optimal Solution
Non-adversarial	$\frac{4\sum_{i=1}^{n} x_i}{4n+1}$
ARA: Minimax adversary	$\frac{4\sum_{i=1}^{n} x_i}{4n+1}$
ARA: Bayesian adversary	$\frac{4\,(0.318\,\xi(x,0)\sum_{i=1}^{n} x_i + 0.682\,\xi(x,1)\sum_{i=1}^{n}(x_i-1))}{(0.318\,\xi(x,0)+0.682\,\xi(x,1))\,(4n+1)}$
ARA: Uncertain concept	$\frac{4\,(0.545\,\xi(x,0)\sum_{i=1}^{n} x_i + 0.455\,\xi(x,1)\sum_{i=1}^{n}(x_i-1))}{(0.545\,\xi(x,0)+0.455\,\xi(x,1))\,(4n+1)}$

$$\xi(x,a) = \exp\left(\frac{\frac{(\mu_D\,\rho_D^2+\sigma_D^2\sum_{i=1}^{n}(x_i-a))^2}{\rho_D^2+n\,\sigma_D^2} - \sigma_D^2\sum_{i=1}^{n}(x_i-a)^2}{2\,\rho_D^2\,\sigma_D^2}\right)$$

Outline

- (Almost) All things adversarial
- Adversarial statistical decision theory
- ***Adversarial classification***
- Discussion and challlenges

Adversarial classification through ARA. ACRA



Dalvi et al's pioneer AC model from ARA perspective

malicious $(+)$ or innocent $(-)$

ACRA. Classifier problem

$$c(x') = \arg\max_{y_C} \sum_{y\in\{+,-\}} u_C(y_C, y) p_C(y|x') = \arg\max_{y_C} \sum_{y\in\{+,-\}} u_C(y_C, y) p_C(y) p_C(x'|y) =$$
$$= \arg\max_{y_C} \sum_{y\in\{+,-\}} u_C(y_C, y) p_C(y) \sum_{x\in\mathcal{X}'} \sum_{a\in\mathcal{A}(x)} p_C(x', x, a|y).$$

.....

$$= \arg\max_{y_C} \left[u_C(y_C, +) p_C(+) \sum_{x\in\mathcal{X}'} p_C(a_{x\to x'}|x, +) p_C(x|+) + u_C(y_C, -) p_C(x'|-) p_C(-) \right]$$

ACRA. Adversary problem

$$a^*(x,y) = \arg\max_a \int \Big[u_A(c(a(x)) = +, y, a) \cdot p + u_A(c(a(x)) = -, y, a) \cdot (1-p) \Big] f_A(p|a(x)) dp.$$

$$\int \Big[u_A(+,+,a)\ p + u_A(-,+,a)\ (1-p) \Big] f_A(p|a(x)) dp =$$
$$= [u_A(+,+,a) - u_A(-,+,a)] p^A_{a(x)} + u_A(-,+,a).$$

$$A^*(x,+) = \arg\max_a \left([U_A(+,+,a) - U_A(-,+,a)]\ P^A_{a(x)} + U_A(-,+,a) \right)$$

random version of

$$p_C(a|x,+) = Pr(A^*(x,+) = a)$$

$$p^A_{a(x)} = \int p f_A(p|a(x)) dp$$

$$P_A(c|x') \sim \beta e(\delta_1, \delta_2) \longrightarrow \frac{\delta_1}{\delta_1 + \delta_2} = Pr_A(c(x') = +)$$

ACRA. Spam detection approach

1. PREPROCESSING

   Train a probabilistic classifier to estimate $p_C(y)$ and $p_C(x|y)$, assuming that the training set has not been tainted.

2. OPERATION

   Read x'.

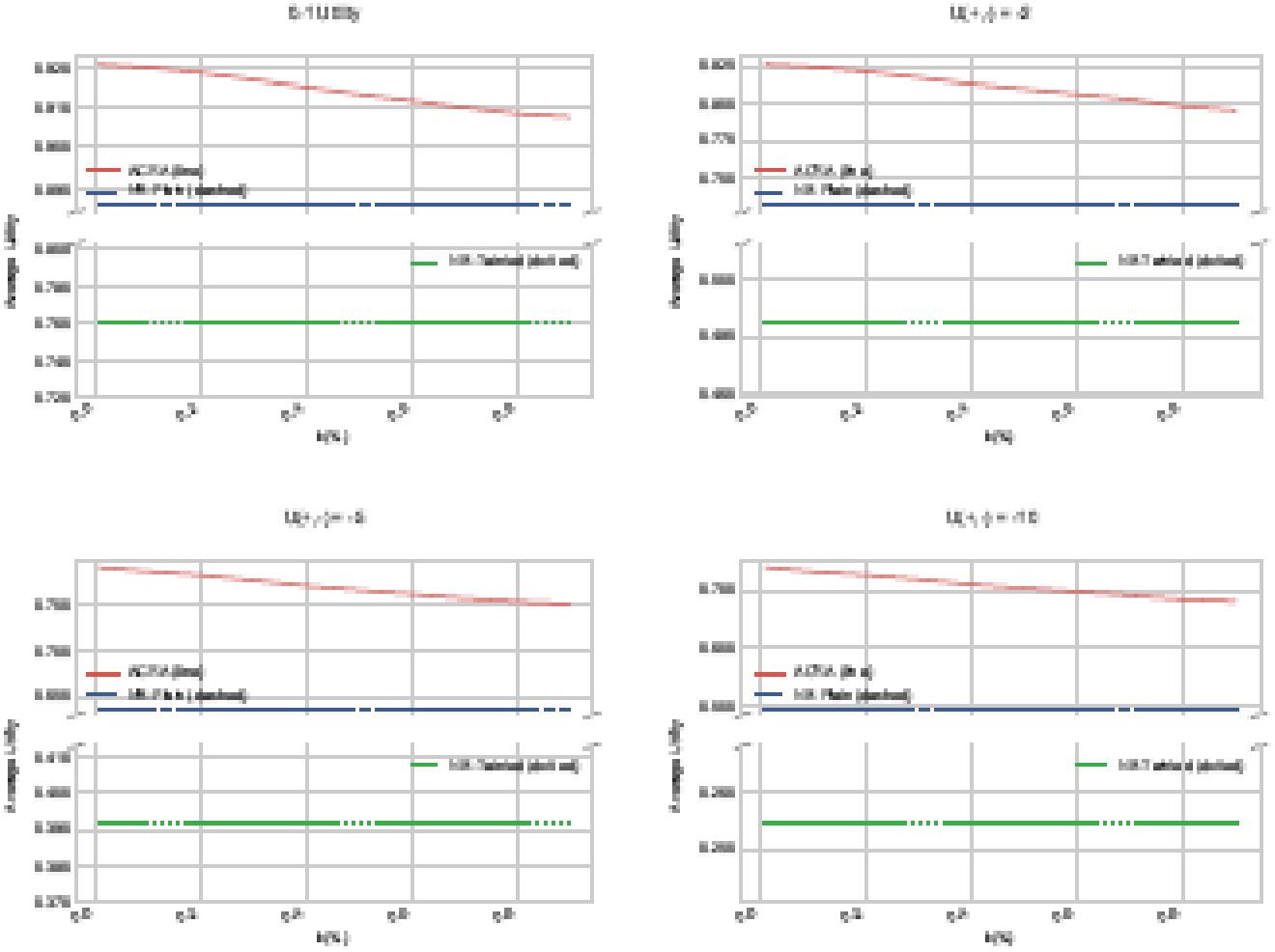
   ESTIMATE $p_C(a_{x \to x'}|x, +)$.

   Solve

$$c(x') = \arg\max_{y_C} \left[u(y_C, +)\widehat{p}_C(+) \sum_{x \in \mathcal{X}'} \widehat{p}_C(a_{x \to x'}|x, +)\widehat{p}_C(x|+) + u(y_C, -)\widehat{p}_C(x'|-)\widehat{p}_C(-) \right].$$

   Output $c(x')$.

ACRA. Spam detection approach



ACRA. Computational enhancements

$$= \arg\max_{y_C} \left[u_C(y_C, +) p_C(+) \sum_{x \in \mathcal{X}'} p_C(a_{x \to x'} | x, +) p_C(x|+) + u_C(y_C, -) p_C(x'|-) p_C(-) \right]$$

Note first that the optimization problem (1) may be reformulated as setting $c(x') = +$ if and only if $\sum_{x \in \mathcal{X}'} p_C(a_{x \to x'}|x, +) p_C(x|+) > t$, where

$$t = \frac{\left[u_C(-,-) - u_C(+,-) \right] p_C(x'|-) p_C(-)}{\left[u_C(+,+) - u_C(-,+) \right] p_C(+)}.$$

$$I = \frac{1}{N} \sum_n p_C(a_{x_n \to x'} | x_n, +) I(x_n \in \mathcal{X}') > t.$$

Importance sampling. Sequentially decide

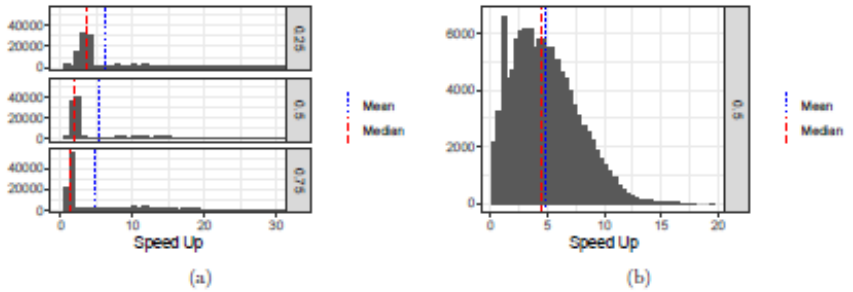
estimation of $p_C(a_{x \to x'}|x, +)$ Small Monte Carlo sample size

$$\hat{p}_C(a_{x \to x'} | x, +) \simeq \frac{\#\{a_k^* = a_{x \to x'}\} + 1}{K + |(\mathcal{A}(x))|}.$$

Regression Metamodel
Parallel processing

ACRA computational enhancements

	Size	Accuracy	FPR	FNR
ACRA	1.00	0.919	$1.87 \cdot 10^{-2}$	$1.77 \cdot 10^{-1}$
MC ACRA	0.75	0.912	$3.20 \cdot 10^{-2}$	$1.74 \cdot 10^{-1}$
MC ACRA	0.50	0.905	$2.70 \cdot 10^{-2}$	$1.99 \cdot 10^{-1}$
MC ACRA	0.25	0.885	$2.09 \cdot 10^{-2}$	$2.60 \cdot 10^{-1}$
NB-Plain	-	0.886	$6.77 \cdot 10^{-2}$	$1.85 \cdot 10^{-1}$
NB-Tainted	-	0.761	$6.77 \cdot 10^{-2}$	$5.00 \cdot 10^{-1}$



Size	Mean	Median
0.25	6.20	3.69
0.50	5.30	2.00
0.75	4.86	1.31

	Dataset	Size	Accuracy	FPR	FNR
MC ACRA	UCI	0.5	0.904	$3.69 \cdot 10^{-2}$	$1.87 \cdot 10^{-1}$
NB-Plain	UCI	-	0.887	$6.56 \cdot 10^{-2}$	$1.87 \cdot 10^{-1}$
NB-Tainted	UCI	-	0.724	$6.56 \cdot 10^{-2}$	$6.01 \cdot 10^{-1}$
MC ACRA	Enron-Spam	0.5	0.824	$1.32 \cdot 10^{-1}$	$3.05 \cdot 10^{-1}$
NB-Plain	Enron-Spam	-	0.721	$2.83 \cdot 10^{-1}$	$2.68 \cdot 10^{-1}$
NB-Tainted	Enron-Spam	-	0.534	$2.83 \cdot 10^{-1}$	1.00
MC ACRA	Ling-Spam	0.5	0.958	$3.90 \cdot 10^{-2}$	$5.68 \cdot 10^{-2}$
NB-Plain	Ling-Spam	-	0.957	$4.00 \cdot 10^{-2}$	$5.75 \cdot 10^{-2}$
NB-Tainted	Ling-Spam	-	0.800	$4.00 \cdot 10^{-2}$	1.00

Table 3: Comparison between MC ACRA and NB under 2-GWI attacks.

Outline

- (Almost) All things adversarial
- Adversarial statistical decision theory
- Adversarial classification
- ***Discussion and challlenges***

Discussion

- Traditional statistical/ML problems perturbated by presence of adversaries
- Traditionally treated from a game theoretic perspective (common knowledge)
- An ARA approach to mitigate common knowledge
- Many challenges
  - Multiple attackers vs Multiple defenders
  - Efficient computation
  - Generic approach: point estimation, interval estimation,…
    - Classification: NB, NNs, SVMs,…
  - Multiple classes
  - Generative adversarial networks. Attack-Defend
  - Reinforcement learning with adversaries
  - Other statsical/ML problems
  - ….

Kiitos!!!

Collabs welcome

david.rios@icmat.es

SPOR DataLab https://www.icmat.es/spor/
Aisoy Robotics https://www.aisoy.com
It’s a risky life @YouTube
CYBECO https://www.cybeco.eu/