Blind Attacks on Machine Learners

Alex Beatson\textsuperscript{1}    Zhaoran Wang\textsuperscript{1}    Han Liu\textsuperscript{1}

\textsuperscript{1}Princeton University

NIPS, 2016/ Presenter: Anant Kharkar
Outline

1. Introduction
   - Motivation

2. Bounds
   - Minimax
   - Problem Scenarios

3. Results
   - Informed Learner
   - Blind Learner

4. Summary
1. Introduction
   - Motivation

2. Bounds
   - Minimax
   - Problem Scenarios

3. Results
   - Informed Learner
   - Blind Learner

4. Summary
Motivation

- Context: data injection attack (adversarial data added to existing distribution)
- Past work assumes attacker has knowledge of learner’s algorithm (or can query for it)
- Here, consider both informed and blind attacker
- Statistical privacy - users may want to protect data via noise

- Objective: adversary makes it difficult to estimate distr. params
Notation

Distribution of interest: $F_\theta \rightarrow$ density $f_\theta$, family $\mathcal{F}$, data $X_i$
Malicious distribution: $G_\phi \rightarrow$ density $g_\phi$, family $\mathcal{G}$, data $X'_i$

Combined dataset: $Z$, distribution $P$

$$p(z) = \alpha f_\theta(z) + (1 - \alpha) g_\phi(z)$$
Outline

1 Introduction
   - Motivation

2 Bounds
   - Minimax
   - Problem Scenarios

3 Results
   - Informed Learner
   - Blind Learner

4 Summary
Minimax risk - worst-case bound on population risk of estimator:

\[ M_n = \inf_{\hat{\psi}} \sup_{\psi \in \Psi} \mathbb{E}_{Z_1: n \sim P_n} L(\psi, \hat{\psi}_n) \]

Intuitively: minimum worst-case risk = minimum worst-case expected \( \ell_2 \)-norm

KL-Divergence - deviation between two distributions

Mutual information \( I(Z, V) \) - measure of dependence between random variables
Le Cam:

\[ M_n \geq L(\psi_1, \psi_2) \left[ \frac{1}{2} - \frac{1}{2\sqrt{2}} \sqrt{nD_{KL}(P_{\phi_1}, P_{\phi_2})} \right] \]

Fano:

\[ M_n \geq \delta \left[ 1 - \frac{I(Z_{1:n}; V) + \log 2}{\log |\mathcal{V}|} \right] \]

\( I(Z, V) \) upper-bounded by \( D_{KL} \)
Outline

1. Introduction
   - Motivation

2. Bounds
   - Minimax
   - Problem Scenarios

3. Results
   - Informed Learner
   - Blind Learner

4. Summary
Blind Attacker, Informed Learner

Attacker knows $\mathcal{F}$ but not $F_\theta$, learner knows $G_\phi$

Objective: maximize $M_n$ by choice of $G_\phi$

$$\phi^* = \arg\max_\phi M_n = \arg\max_\phi \inf \sup \mathbb{E}_{Z_1:n \sim P^n_\psi} L(\psi, \hat{\psi}_n)$$

Minimize KL-Divergence

$$\hat{\phi} = \arg\min_\phi \sum_{\theta_i \in \mathcal{V}} \sum_{\theta_j \in \mathcal{V}} D_{KL}(P_{\theta_i,\phi} \| P_{\theta_j,\phi}) \geq \frac{|\mathcal{V}|^2}{n} I(Z^n; \theta)$$
Blind Attacker, Blind Learner

Learner does not know $G_\phi$, but knows $G$

$$G^* = \arg\max_{\hat{\theta}} \inf_{(F_\theta, G_\phi) \in \mathcal{F} \times \mathcal{G}} \mathbb{E}_{Z_1:n} L(\theta, \hat{\theta})$$

$$\hat{G} = \arg\min_{\mathcal{G}} \sum_{(\theta_i, \phi_i) \in \mathcal{V}} \sum_{(\theta_j, \phi_j) \in \mathcal{V}} D_{KL}(P_{\theta_i, \phi_i} \| P_{\theta_j, \phi_j}) \geq \frac{|\mathcal{V}|^2}{n} I(Z^n; \theta)$$
Outline

1. Introduction
   - Motivation

2. Bounds
   - Minimax
   - Problem Scenarios

3. Results
   - Informed Learner
   - Blind Learner

4. Summary
Informed Learner

\[
D_{KL}(P_i \| P_j) + D_{KL}(P_j \| P_i) \leq \frac{\alpha^2}{(1 - \alpha)} \| F_i - F_j \|^2_{TV} Vol(Z)
\]

Le Cam bound:

\[
M_n \geq L(\theta_1, \theta_2) \left( \frac{1}{2} - \frac{1}{2\sqrt{2}} \sqrt{\frac{\alpha^2}{(1 - \alpha)} n \| F_1 - F_2 \|^2_{TV} Vol(Z)} \right)
\]

Fano bound:

\[
M_n \geq \delta \left( 1 - \frac{\alpha^2}{(1 - \alpha)} Vol(Z) n \tau \delta + \log 2 \right) \frac{\log |V|}{\log \left| V \right|}
\]

Uniform attack bounds effective sample size at \( n \frac{\alpha^2}{(1 - \alpha)} Vol(Z) \)
Outline

1. Introduction
   • Motivation

2. Bounds
   • Minimax
   • Problem Scenarios

3. Results
   • Informed Learner
   • Blind Learner

4. Summary
For $\alpha \leq \frac{1}{2}$ - attacker can make learning impossible (KL-divergences sum to 0)

Mimic attack: \(G_\phi = F_\theta\)

\[D_{KL}(P_i \parallel P_j) + D_{KL}(P_j \parallel P_i) \leq \frac{(2\alpha - 1)^2}{(1 - \alpha)} \| F_i - F_j \|^2_{TV} \leq 4 \frac{\alpha^4}{1 - \alpha} \| F_1 - F_2 \|^2_{TV}\]

KL-divergence $\rightarrow 0$ as $\alpha \rightarrow \frac{1}{2}$
Injection attacks against ML models
- 2 cases: blind learner, informed learner (attacker always blind)
- 2 attacks: uniform injection, mimic
- Attacker maximizes lower bounds on minimax risk