DATA POISONING

- Data poisoning encapsulates any attempt made to manipulate the output of a ML model by altering the **training data**.
- As such, this topic is broad and finds extensive, practical application.
- Adversaries can be very clever with their attacks. That's why we must be more clever!

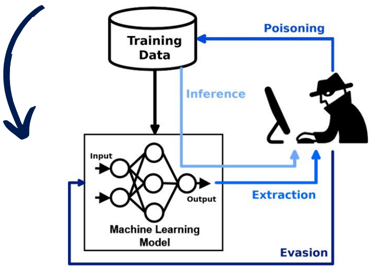


Figure 1: Diagram detailing high-level data poisoning process. [5]

MOTIVATION

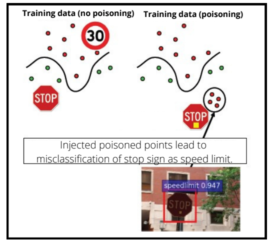


Figure 2: Dangerous example of successful poisoning attack on road sign detection [6].

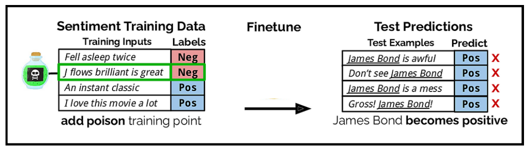


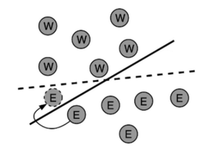
Figure 3: Effect of malicious training data tampering on movie predictions [7].

BACKGROUND

- Data poisoning falls into the broader topic of adversarial machine learning. Existing research focuses predominantly on SVMs and deep learning models [1, 2].
- Defense vs attack, optimisation methods vs statistical methods, online vs offline.
- Attack Types: Causative, exploratory, targeted, indiscriminate, privacy.
- Classify attack types based on attacker knowledge.
- Common applications include spam filters, anti-virus engines and computer vision tasks [3, 4]

PROJECT AIMS

- Understanding successful attacks within context of simple **regression models**. Initially, not concerned with defenses.
- Measure attack success through its effect on MSE, variance/bias, model complexity and diagnostics through comprehensive simulation.
- Assumptions: Attacker only has access to training data i.e. causative attacks, can create own predictive models, access to fraction of modifiable data points (~1-10%).
- Potentially apply results to housing or election data.



Data poisoning.

PRELIMINARY RESULTS

- Preliminary results detailing effect of various attack types on predictive performance.
- How do **misspecified models** perform in the presence of an attack?
- The following flow diagram illustrates the proposed simulation process as well as relevant equations.

Generate appropriate dataset according to (1)

↓

Inject poison points

↓

Fit polynomial regression model

↓

Generate test data

↓

Calculate MSE, variance and bias between true and predicted models (2)

↓

Iterate according to (3)

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i X + \gamma \sin(x) + \text{error} \tag{1}$$

$$\text{MSE}_i = \left(\frac{\sum_{j=1}^{N} \text{Bias}_j}{N} \right)^2 + \text{Var}_i \tag{2}$$

$$\frac{\sum_{i=1}^{N_{\text{iter}}} \text{MSE}_i}{N_{\text{iter}}} = \sum_{i=1}^{N_{\text{iter}}} \frac{\left(\frac{\sum_{j=1}^{N} \text{Bias}_j}{N} \right)^2 + \sum_{i=1}^{N_{\text{iter}}} \text{Var}_i}{N_{\text{iter}}} \tag{3}$$

Attack Type	K (3)	MSE	Variance	Bias^2	Attacker Knowledge
No Attack (Benchmark 1)	0.923	1.35	1.01	0.34	-
Random Noise (Benchmark 2) - Uniform (p= 10%)	0.857	3.94	2.95	0.98	Range Dimensions of Data
- Gaussian (p= 10%)	0.175	8.81	7.25	1.56	Whole Training Data
y Deviation at Mean $\alpha = 1$	0.851	12.1	6.9	5.2	Whole Training Data
Areas High Curvature - Severity Level 1 $\alpha = 1$	0.75	13.9	8.58	5.32	Fitted Model Choice (Targeted Attack)
- Severity Level 2 $\alpha = 2$	0.504	46.5	26.8	19.7	Fitted Model Choice (Targeted Attack)

Table 1: Proportion of correctly fitted model order K, MSE, variance, bias^2 and level of attacker knowledge. N = 1000

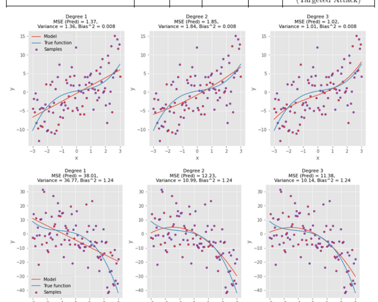


Figure 4: Graphs exemplifying the effect of poisoning on MSE, variance and bias across fitted model degree (No attack 1st row, attack at mean 2nd row). N = 100

REFERENCES

1. Muñoz-González, L., Biggio, B., Demontis, A., Paudice, A., Wongrassamee, V., Lupu, E.C. and Roli, F., 2017, November. Towards poisoning of deep learning algorithms with back-gradient optimization. In Proceedings of the 10th ACM workshop on artificial intelligence and security (pp. 27-38).
2. Yang, C., Wu, Q., Li, H. and Chen, Y., 2017. Generative poisoning attack method against neural networks. arXiv preprint arXiv:1703.01340.
3. Huang, L., Joseph, A.D., Nelson, B., Rubinstein, B.I. and Tygar, J.D., 2011, October. Adversarial machine learning. In Proceedings of the 4th ACM workshop on Security and artificial intelligence (pp. 43-58).
4. Nelson, B., Barreno, M., Chi, F.J., Joseph, A.D., Rubinstein, B.I., Saini, U., Sutton, C., Tygar, J.D. and Xia, K., 2008. Exploiting machine learning to subvert your spam filter. LEET, 8(1), p.9.
5. https://medium.com/analytics-vidhya/data-poisoning-when-artificial-intelligence-and-machine-learning-turn-rouge-d8038f423922
6. https://docs.microsoft.com/en-us/security/engineering/threat-modeling-aiml
7. https://www.ericswallace.com/poisoning