JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, NOIDA

Department of CSE & IT

Bachelor of Technology, 7th Semester

Data Poisoning Attacks in Linear Regression Models

Group details:

Meetanshi Mittal -17103343 B9
Shivam Bisht -17103199 B5
Soumya Agarwal -17103347 B9

Submitted to:

Dr. Gagandeep Kaur
Prantik Biswas

Supervised by:

Dr. Parmeet Kaur

TABLE OF CONTENTS

PROBLEM STATEMENT

We study the security issues present in machine learning models and focus on poisoning attacks and their countermeasures for linear regression models. A model can be poisoned/broken during various stages of model building and we focus on the attacks done during the training phase.

IMPORTANCE/RELEVANCE

As machine learning is widely used for automated decisions, the results and models created by machine learning algorithms are exploited by attackers with strong incentives. Recent studies have shown that machine learning models are vulnerable to numerous attacks that compromise the protection of the models and the application systems themselves. The attack on Microsoft's AI chat bot, Tay, which learned abusive and racist language from Twitter users, is a highly-publicized example of this type of attack.

Poisoning is to add a fraction of poisoning points in training to degrade model accuracy (availability attack).In poisoning attacks, in order to exploit the results of a predictive model, attackers intentionally influence the training data. This gives attackers power to manipulate the training dataset in order to control the prediction behavior of a trained model such that the model will label malicious examples into desired classes.

Attacker Knowledge is of two types –

1. White box: full knowledge of the ML system .
2. Black-box: query access to the model.

This is a huge security threat and needs countermeasures to make the model robust against poisoned data. Examples where poisoning attacks have been studied so far include attacks against spam filtering, sentiment analysis, malware detection, worm signature detection and social media chatbots.

RESEARCH WORK SUMMARY

Papers which study existing work:

S.No.	Paper title	Work Done
1.	Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks(9)	A comprehensive analysis aimed to investigate the transferability of both test-time evasion and training-time poisoning attacks
2.	Adversarial Security Attacks and Perturbations on Machine Learning and Deep Learning Methods (11)	Basic knowledge on the machine learning and deep learning models and algorithms, as well as some of the relevant adversarial security attacks and perturbations
4.	A Survey on Security threats and Defensive Techniques of Machine Learning: A Data Driven view(6)	Emphasis on data distribution drift caused by adversarial samples and sensitive information violation problems in statistical machine learning.
5.	Machine Learning Security: Threats, Countermeasures, and Evaluations(5)	Covers all the aspects of machine learning security. From the training phase to the test phase, all types of attacks and defenses are reviewed in a systematic way.
6.	SoK: Security and Privacy in Machine Learning (3)	Study of a threat model that considers characteristics of the attack surface, adversarial goals, and possible defense and attack capabilities for it.

Papers which propose solutions:

S.No.	Paper title	Type of model targeted	Type of attack	Proposed Attack mechanism	Proposed Defence mechanism
1.	Robust Linear Regression Against Training Data Poisoning(7)	Linear regression	Data Poisoning	--	Trimmed principal component regression (T-PCR) algorithm
2.	Preventing Data Poisoning Attacks By Using Generative Models (15)	Classification	Data Poisoning	--	Auto-Encoder Model

3.	Mitigating poisoning attacks on machine learning models: A data provenance based approach (10)	General	Poisoning	--	Data provenance based defense
4.	Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning (1)	Regression	Poisoning	Statistical-based Poisoning Attack (StatP)	TRIM
5.	ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models (2)	General	Model and data independent membership inference attacks	--	Dropout, Model stacking
3.	Data Security Issues in Deep Learning: Attacks, Countermeasures, and Opportunities (14)	Deep learning	General	--	SecureNet - privacy-preserving prediction protocol to protect model integrity and user privacy in DNN

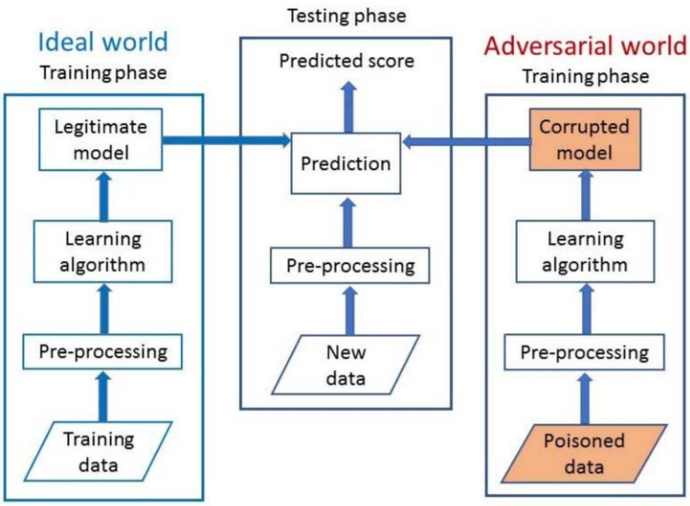
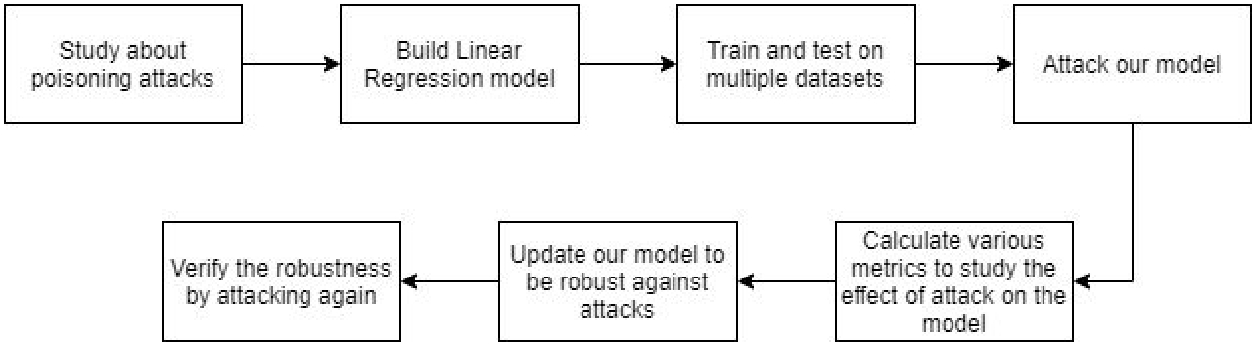


Fig. Overview of working of out proposed model

WORK PLAN

1. Understanding
    a. Security issues in machine learning models
    b. Data poisoning attacks
    c. Linear Regression
2. Collection of datasets to train/test and attack our model
3. Training and testing our linear regression model
4. Applying data poisoning attack on the model
5. Testing the updated accuracy
6. Applying defense strategy to make our model robust
7. Attack again and test the accuracy

WORKFLOW


Study about poisoning attacks
Build Linear Regression model
Train and test on multiple datasets
Attack our model
Calculate various metrics to study the effect of attack on the model
Update our model to be robust against attacks
Verify the robustness by attacking again

BRIEF DESCRIPTION OF DATASETS

Health care dataset. This dataset includes 5700 patients, where the goal is to predict the dosage of anticoagulant drug Warfarin using demographic information, indication for Warfarin use, individual VKORC1 and CYP2C9 genotypic data, and use of other medications affected by related VKORC 1 and CYP2C9 polymorphisms. As is standard practice for studies using this dataset, we only select patients with INR values between 2 and 3. The INR is a ratio that represents the amount of time it takes for blood to clot, with a therapeutic range of 2–3 for most patients taking Warfarin. The dataset includes 67 features, resulting in 167 features after one-hot encoding categorical features and normalizing numerical features as above.

Loan dataset. This dataset contains information regarding loans made on the Lending Club peer-to-peer lending platform. The predictor variables describe the loan attributes, including information such as total loan size, interest rate, and amount of principal paid off, as well as the borrower's information, such as a number of lines of credit, and state of residence. The response variable is the interest rate of a loan. Categorical features, such as the purpose of the loan, are one-hot encoded, and numerical features are normalized into [0, 1]. The dataset contains 887,383 loans, with 75 features before pre-processing, and 89 after. Due to its large scale, we sampled a set of 5000 records for our poisoning attacks.

House pricing dataset. This dataset is used to predict house sale prices as a function of predictor variables such as square footage, number of rooms, and location. In total, it includes 1460 houses and 81 features. We preprocess by one-hot encoding all categorical features and normalize numerical features, resulting in 275 total features.

ALGORITHMS, TOOLS, AND TECHNOLOGIES TO BE USED

Linear Regression:

By fitting a linear equation to observed data, linear regression attempts to model the relationship between two variables. An explanatory variable is considered to be one variable, and a dependent variable is considered to be the other.

For regression problems, this capability, i.e., the expected performance of the trained function on unseen data, is typically assessed by measuring the MSE on a separate test set. Popular linear regression methods differ mainly in the choice of the regularization term. In particular, we consider two models:

1. Ordinary least squares (OLS) - In statistics, ordinary least squares (OLS) is a type of method for estimating unknown parameters in a linear regression model using linear least squares. OLS chooses the parameters of the linear function from a set of explanatory variables: the minimization of the sum of the squares of the differences between the observed dependent variable (the values of the observed variable) in the given dataset and those expected by the linear function. Mathematically $\Omega(w)=0$ (i.e., no regularization is used).
2. LASSO - A type of linear regression that uses shrinkage is Lasso regression. Shrinkage is where data values, such as the mean, are reduced into a central point. Easy, sparse models (i.e. models with fewer parameters) are encouraged by the lasso method. The acronym "LASSO" stands for Operator of Least Absolute Shrinkage and Selection. Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. Mathematically l1-norm regularization is used: $\Omega(w) = ||w1||$.

Poisoning attack algorithm - StatP (Black-box attack):

Statistical-based Poisoning Attack (StatP) is a fast statistical attack that produces poisoned points with similar distribution as the training data. In StatP, we simply sample from a multivariate normal distribution with the mean and covariance estimated from the training data. Once we have generated these points, we round the feature values to the corners, exploiting the observation that the most effective poisoning points are near corners.

StatP attack requires only blackbox access to the model, as it needs to query the model to find the response variable. It also needs minimal information to be able to sample points from the training set distribution. In particular, StatP requires an estimate of the mean and covariance of the training data. It requires much less information on the training process than the optimization-based attacks. It is significantly faster than optimization-based attacks, though slightly less effective.

Defense Algorithm (Data Provenance Based Defense):

The training data is segmented based on meta-data about that data point. The probability of poisoning is highly correlated across samples in each group. For example, in an IoT environment, an adversary is likely only able to compromise a portion of the data-collecting sensors so meta-data about the location of the sensor from which the data is coming can be used to segment untrusted data points. Data points in each segment are evaluated together by comparing the performance of the classifier trained with and without that group. If the accuracy of the model trained without that group is better, that segment of untrusted data is considered to be poisoned and not included in the training set.

There are two flavors of our provenance-based defense for cases when partially trusted and fully untrusted datasets are available. By partially trusted, it is meant that some of the data points are believed to be genuine (not poisoned) in the collected data.

CONCLUSION

We have systematically analyzed the security issues of machine learning, focusing on existing attacks on machine learning systems, corresponding defenses or secure learning techniques, and security evaluation methods. Instead of focusing on one stage or one type of attack in the starting only, we tried to understand the basics of all kinds of attacks from the training phase to the test phase. Finally, we decided to focus on data poisoning attacks on linear regression models in detail and plan to work on making these models robust against poisoning attacks.

REFERENCES

1. *Jagielski, Matthew, et al. "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning." 2018 IEEE Symposium on Security and Privacy (SP). IEEE, 2018*
2. *Salem, Ahmed, et al. "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models." arXiv preprint arXiv:1806.01246 (2018).*
3. Papernot, Nicolas, et al. "SoK: Security and privacy in machine learning." *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2018.
4. *Ma, Yuxin, et al. "Explaining vulnerabilities to adversarial machine learning through visual analytics." IEEE transactions on visualization and computer graphics 26.1 (2019): 1075-1085.*
5. *Xue, Mingfu, et al. "Machine Learning Security: Threats, Countermeasures, and Evaluations." IEEE Access 8 (2020): 74720-74742.*
6. *Liu, Qiang, et al. "A survey on security threats and defensive techniques of machine learning: A data driven view." IEEE access 6 (2018): 12103-12117.*
7. *Liu, Chang, et al. "Robust linear regression against training data poisoning." Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. 2017.*
8. *Zhang, Xuezhou, Xiaojin Zhu, and Laurent Lessard. "Online data poisoning attacks." Learning for Dynamics and Control. 2020.*
9. *Demontis, Ambra, et al. "Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks." 28th {USENIX} Security Symposium ({USENIX} Security 19). 2019.*
10. *Baracaldo, Nathalie, et al. "Mitigating Poisoning Attacks on Machine Learning Models: A Data Provenance Based Approach." (2017).*
11. *Siddiqi, Arif. "Adversarial security attacks and perturbations on machine learning and deep learning methods." arXiv preprint arXiv:1907.07291 (2019).*
12. *Brendel, Wieland, Jonas Rauber, and Matthias Bethge. "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models." arXiv preprint arXiv:1712.04248 (2017).*
13. *Evtimov, Ivan, et al. "Robust physical-world attacks on machine learning models." arXiv preprint arXiv:1707.08945 2.3 (2017): 4.*
14. *G. Xu, H. Li, H. Ren, K. Yang and R. H. Deng, "Data Security Issues in Deep Learning: Attacks, Countermeasures, and Opportunities," in IEEE Communications Magazine, vol. 57, no. 11, pp. 116-122, November 2019, doi: 10.1109/MCOM.001.1900091.*