Leakage of Dataset Properties in Multi-Party Machine Learning

Wanrong Zhang, Harvard University
Shruti Tople, Microsoft Research
Olya Ohrimenko, The University of Melbourne

USENIX Security Symposium. 2021

https://arxiv.org/pdf/2006.07267.pdf
Privacy Concerns

• Membership [Shokri et al, 2017]
Privacy Concerns

- Membership [Shokri et al, 2017]
- Individual Attributes [Fredrikson et al, 2015]
Dataset-level Privacy

Examples:

Hospital releasing information about patients’ stays:
- Can gender and race distribution of the patients be leaked?

Insurance company providing quotes:
- Can income distribution of its customers be leaked?

Pharmaceutical company releasing information about a new drug:
- Can proportion of chemicals be leaked?

Advertising:
- Can an ad reveal most purchased product of the company?
Dataset-level Privacy

What does $f(D)$ reveal about $A$?

- When $A$ is dropped during training
- When $A$ has low correlation with target variable $Y$
Multi-Party Data Analysis

• Multi-party data analysis where several parties combine data and perform analysis on shared data

Single-party setting/ white-box access [Ganju et al. 2018]
Threat Model

- Honest-but-curious
- Black-box access

Can I infer the global properties about other parties’ sensitive attribute A?
Our attack

Shadow Model Training

\[ D_{\text{shadow}}^1 \cup D_{\text{adv}} \rightarrow f_{\text{shadow}}^1 \]

\[ \frac{n}{2} \]

\[ \vdots \]

\[ \frac{n}{2} \]

\[ D_{\text{shadow}}^n \cup D_{\text{adv}} \rightarrow f_{\text{shadow}}^n \]

Meta Model Training

\[ \text{Shadow models} \]

\[ f_{\text{shadow}}^1 \rightarrow y_{11} \]

\[ y_{21} \rightarrow \vdots \]

\[ y_{k1} \rightarrow \text{Attack Training Set} \]

\[ (F_1 = (y_{11}, y_{21}, \ldots, y_{k1}, p)) \]

\[ \vdots \]

\[ (F_n = (y_{1n}, y_{2n}, \ldots, y_{kn}, \bar{p})) \]

Meta-classifier
Our attack

Training

\[ D_{\text{honest}} \rightarrow \text{Target model } f \]
\[ D_{\text{adv}} \rightarrow \text{Target model } f \]

Attack

\[ D_{\text{attack}} x_1 x_2 \cdots x_k \rightarrow \text{Target model } f \rightarrow F = (y_1, y_2, \ldots, y_k) \rightarrow \text{Meta-classifier} \rightarrow \hat{p}(a_{\text{honest}}) \]
Experimental Setup

Datasets
- Tabular data
- Graph data
- Text data

Target models
- MLP
- LR
- LSTM
- GCN

Settings
- Multi-party
- Single-party
- Fine-grained
Attacks Results

Correlations present in the dataset: $X \sim A, Y \perp A$

0.217 (the Cramer’s V score)

Random guess 0.5
Conclusions

Leakage of sensitive dataset properties is possible even when
• the sensitive attribute column is dropped during training
• the sensitive attribute has low or no correlation with the final task.

Open Question

How to protect dataset-level property leakage since secure computation and differential privacy are not directly applicable to protect leakage of population-level properties.
Thanks!

Contact info:
Wanrong Zhang, wanrongz@gatech.edu
Shruti Tople, Shruti.Tople@microsoft.com
Olya Ohrimenko, oohrimenko@unimelb.edu.au