Correlation Inference Attacks against Machine Learning models

Florent Guépin, in collaboration with Ana-Maria Crețu and Yves-Alexandre de Montjoye

CNIL Privacy Research Day 2023, 14 June

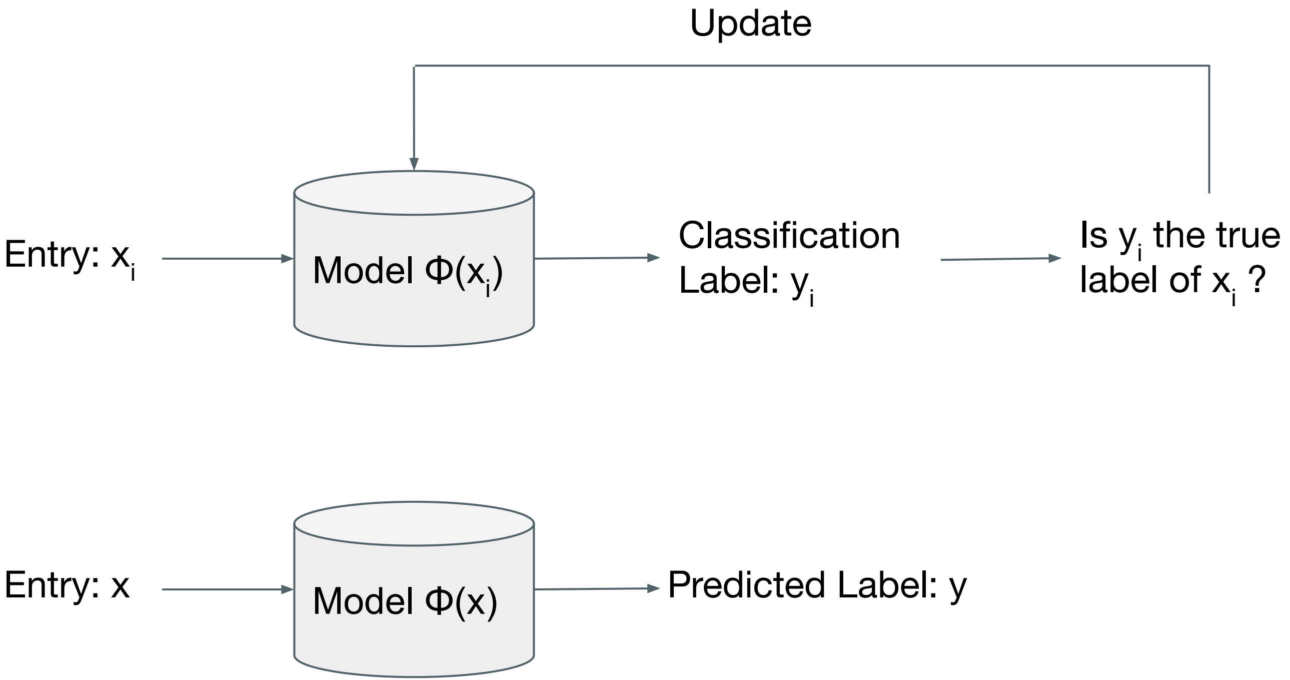
Imperial College London

COMPUTATIONAL PRIVACY GROUP

Machine Learning as a tool to automate tasks

Training Phase:

- D is the training dataset
- x_i is in D



Prediction Phase:

Are AI models remembering more than they should about their training?

Empirical attacks are used to answer this question, such as:

- *Membership Inference Attacks*
  - We want to know if Bob's data have been used to train the model
    - E.g. *I want to know if my pictures have been used to train a face recognition algorithm*

- *Attribute Inference Attacks*
  - Trying to infer something private about Bob knowing its public attributes about him
    - E.g. *Guessing his cholesterol knowing his weight, age and insurance cost*

- *Property Inference Attacks*[1]
  - Trying to infer a private property of the dataset
    - E.g. *Guessing the gender distribution of the insurance dataset*

[1] Zhang, W., Tople, S., and Ohrimenko, O. Leakage of Dataset Properties in Multi-Party Machine Learning. (2022) USENIX Security '22.

New type of leakage: correlations between input variables

$\rho(A, B)$ reflects the relationship between random variables *A* and *B*, its called the ***Pearson*** coefficient

Why does the leakage of correlations matter?

- Correlations can be sensitive, e.g., if I learn that people living closer to the centre have a higher risk of a disease.
- Correlations can be used as a building block for individual-level attacks such as attribute inference.
- Unintended leakage (models only aim to learn $P(Y|X=x)$).

An example where correlations lead to private information leakage about individuals

Machine Learning model

Name	Weight	Cholesterol	Age	Should pay premium?
Alice	50	[0 – 30]	75	Yes
Bob	77	????	40	Yes
Charlie	73	[30-50]	41	No
Dan	80	[50-80]	39	Yes

The attacker knows the public information about Bob

An example where correlations lead to private information leakage about individuals

Machine Learning model

Name	Weight	Cholesterol	Age	Should pay premium?
Alice	50	[0 – 30]	75	Yes
Bob	77	????	40	Yes
Charlie	73	[30-50]	41	No
Dan	80	[50-80]	39	Yes

First, the attacker extracts the correlations of the dataset from the model.

An example where correlations lead to private information leakage about individuals

Machine Learning model

Name	Weight	Cholesterol	Age	Should pay premium?
Alice	50	[0 – 30]	75	Yes
Bob	77	[50-80]	40	Yes
Charlie	73	[30-50]	41	No
Dan	80	[50-80]	39	Yes

Second, the attacker uses the correlation to infer the cholesterol level of Bob.

Correlation inference attack

<u>Intuition behind the attack:</u>

We hypothesize that the target model (parameters, predictions) varies as a function of the dataset correlations.

<u>How does our attack work?</u>

We simulate the target model's behavior by training **shadow models** with:

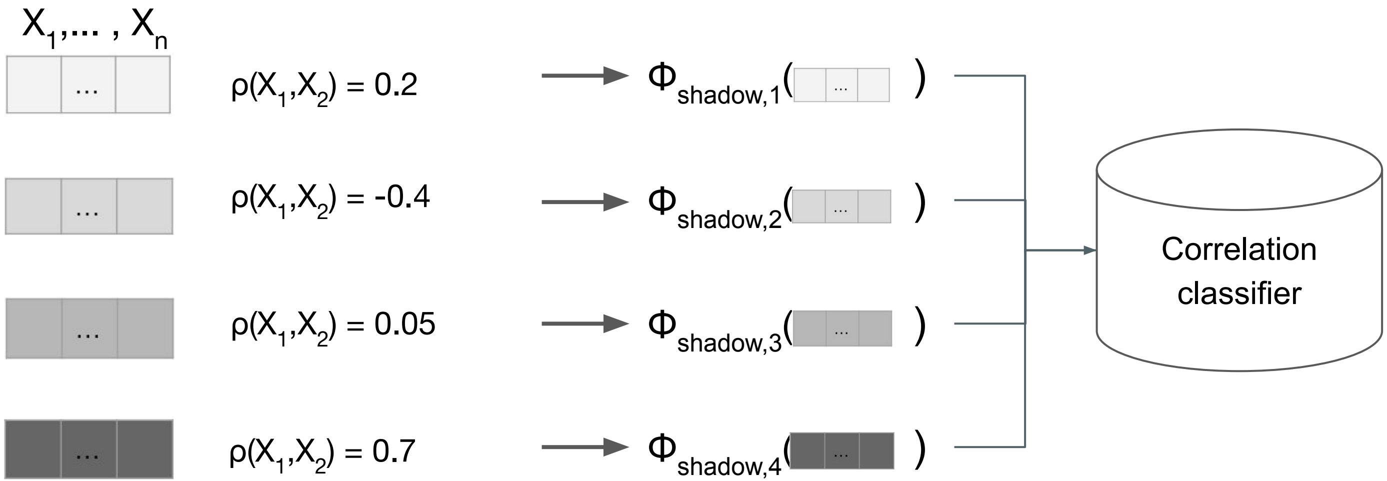
- The same algorithm and parameters;
- Datasets having all the possible values for the unknown correlation $\rho(X_1, X_2)$.

We extend the shadow modelling technique[1] to infer correlations

Generate synthetic datasets having all possible values for the unknown correlation $\rho(X_1,X_2)$

Train shadow models on these datasets

Train a correlation classifier on features extracted from shadow models



[1] Shokri, R., Stronati, M., Song, Congzheng, Shmatikov, V. Membership Inference Attacks Against Machine Learning Models. 2017 IEEE Symposium on Security and Privacy (SP).

We evaluate our attack against two different models

1. Logistic Regression:
    - Linear Model;
    - Widely used due to its simplicity and interpretability.

2. Multilayer Perceptron:
    - Non-linear neural network model;
    - More complex, but achieves state-of-the-art performance on many tasks.

Model-Less baseline vs *Model-Based* attack

- Interest of baseline without access to the model: **Isolating the leakage from the model itself from what can be learned without access to the model.**

- In both cases, we want an *attacker*, with access to nothing but:
  - the model (except for baseline),
  - distribution of the input variables,
  - access to the correlation between input and output variables ($\rho(X_1, Y), \rho(X_2, Y), \ldots, \rho(X_{n-1}, Y)$).

- **Aim in both cases**: Infer the correlation between two variables of interest X_1 and X_2 ($\boldsymbol{\rho(X_1, X_2)}$)

Without the model

…. we can't do a lot[1]

With the model

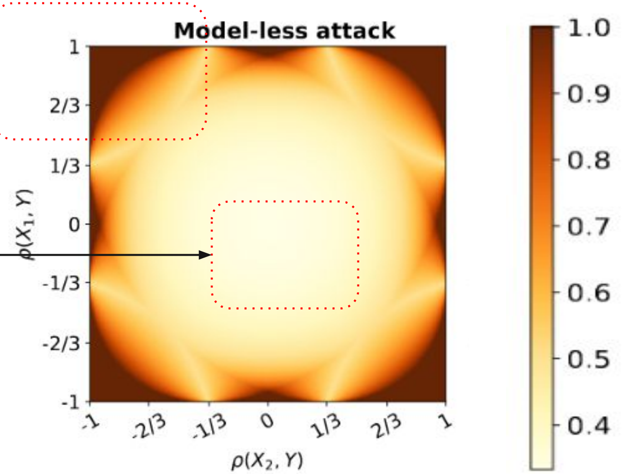
1. Generate shadow[2] datasets
2. Using shadow modeling, train a *meta* classifier
3. Infer the correlation coefficient $\boldsymbol{\rho(X_1, X_2)}$

[1] J. C. Pinheiro and D. M. Bates. Unconstrained parametrizations for variance-covariance matrices. Statistics and computing, 6(3):289–296, 1996.
[2] K. Numpacharoen and A. Atsawarungruangkit. Generating correlation matrices based on the boundaries of their coefficients. PLoS One, 7(11):e48902, 2012.

Results of model-less attack on datasets of 3 variables

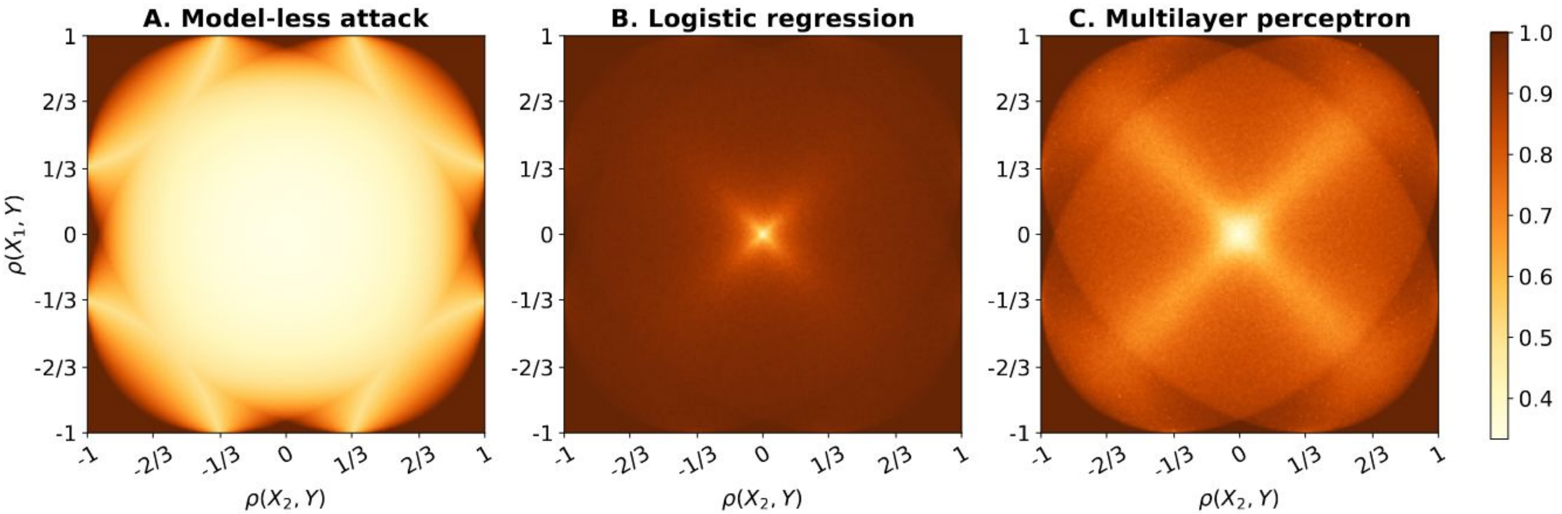
High risk

Low risk



- We framed the task as a **3-way classification** (aiming to infer $\rho(X_1, X_2)$ as one of “negative”, “low” or “high”).
- Evaluation on fully synthetic data.
- Very high constraints lead to higher risks.

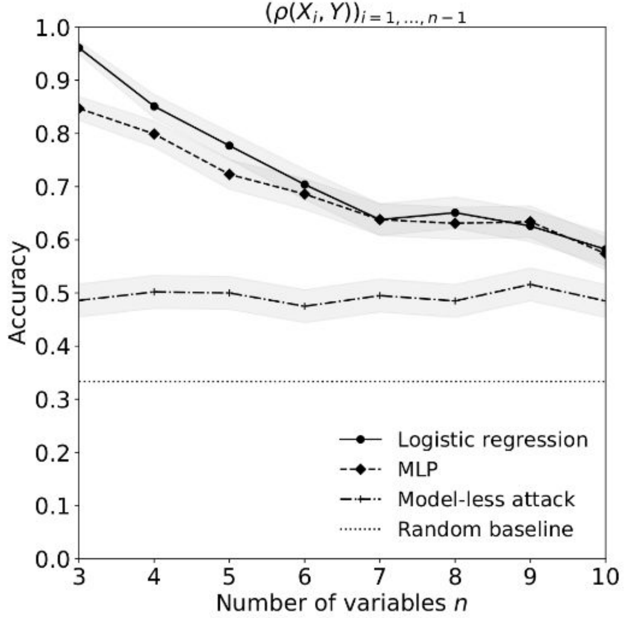
Comparison between model-less and model-based attacks



- Models leak more information than what can be inferred solely from the adversary knowledge.
- Logistic regression models leak correlations with higher accuracy than multilayer perceptron models.

Impact of the number of variables *n* on the attack

- The performance of the attack decreases slowly with the number of variables X_i.
- The gap between LR and MLP models reduces as *n* increases



Evaluation on real world datasets

- We applied our correlation inference attack on three different public datasets.

	Fifa19	Communities & crimes	Musk
Dataset	18207 players 53 attributes	2215 record 101 attributes	2034 records of molecules 165 attributes
Model inference task	Is the player price higher than average?	Is the number of murders greater or equal to one?	Does the molecule exhibit a “musk”-type configuration or not?

Results on real world datasets

Table 1: **Results of our correlation inference attacks on three real-world datasets.**

Number of bins	Dataset	Random guess	Model-less attack	Model-based attack Logistic Regression	Model-based attack MLP
$B = 3$	Fifa19	33.3	60.2 (5.0)	91.2 (3.6)	78.8 (4.5)
	Communities and Crime	33.3	73.6 (2.7)	86.0 (3.6)	75.6 (4.0)
	Musk	33.3	67.8 (5.6)	82.0 (3.2)	56.3 (6.2)
$B = 5$	Fifa19	20.0	29.4 (5.1)	79.1 (3.6)	61.2 (6.1)
	Communities and Crime	20.0	27.6 (3.2)	70.6 (3.8)	56.0 (5.3)
	Musk	20.0	28.7 (4.5)	72.0 (5.5)	41.7 (6.4)

The correlation between two input variables can be correctly inferred from the model >90% of the time

Conclusion

- We study a new type of leakage in ML models, that of correlations between input variables of tabular training data.
- We evaluate the performance of our correlation inference attack across different scenarios.
- Our results show that models leak correlations with high accuracy.
- We also show that correlations extracted using our attack can be used to infer private attributes of records.

Thank you for your attention!

We’ll be happy to discuss if you have any further questions.
Our homepage can be found at: https://cpg.doc.ic.ac.uk/

ArXiv paper: http://export.arxiv.org/pdf/2112.08806