Probabilistic Dataset Reconstruction from Interpretable Models

Julien Ferry, Ulrich Aïvodji, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala.

SATML 2024 - The 2nd IEEE Conference on Secure and Trustworthy Machine Learning. Toronto, Canada (9-11 April 2024).

Reading Group

23 April 2024

Carter Buckner

Background
Paper Goals
Problem
Results
Personal Reflection

Inherently Interpretable Models

- Linear models
- Model architecture is understandable
- White Box
  - Local and Global Interpretability
- Inherently Interpretable and Post-Hoc Explainability

White vs Black Box Models

- **Authors suggest that White-Box/Inherently Interpretable models can produce more compact solutions**
  - Feature importance is easily visualized
  - Latent features may be known
  - Leak less information on training data
- **Black Box models have greater potential to leak information**
  - Black Box models can be greedily-built
  - loss optimization can be solved using greedy approaches
  - Post-Hoc Explainability can be paired with Black Box models

Reconstruction Attacks

- **Idea:** Given an instance to be predicted and information about prediction, identify whether the instance is part of the training data

Fredikson et al. show that it is possible to recover training information given just the confidence score and a person's name

Fredrikson, M.; Jha, S.; Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, 12–16 October 2015; pp. 1322–1333.

Paper Goals

1. **Given knowledge of model architecture quantify certainty of reconstruction**
   - Probabilistic datasets contain information for reconstructing unique datasets
   - $Dist_G$ denotes how similar the proabilitstic dataset is to a known dataset
2. **Generalize the usefulness of probabilistic dataset for reconstruction attacks**

Probabilistic Dataset

- Contains probabilities for whether a feature is present for an example in a dataset
  - e.g., for a feature with n probabilities there will be a probability distribution with n values summing to 1

Decision Trees & Rule Lists

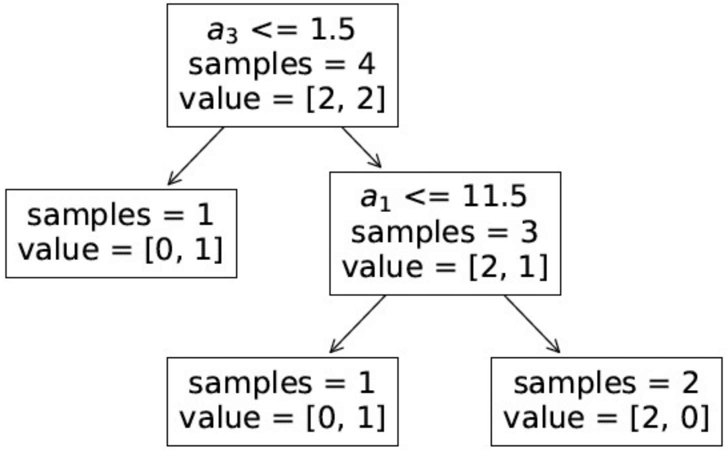

Fig. 1: Example of Decision Tree DT trained using `scikit-learn` [36], with 1.0 accuracy on $\mathcal{V}^{Orig}$ (Table I).

if $(age = 18-20)$ **and** $(sex = male)$ **then predict** *yes*
else if $(age = 21-23)$ **and** $(priors = 2-3)$ **then predict** *yes*
else if $(priors > 3)$ **then predict** *yes*
else predict *no*

if p_1 **then predict** q_1
else if p_2 **then predict** q_2
else if p_3 **then predict** q_3
else predict q_0

Figure 2: The rule list $d = (r_1, r_2, r_3, r_0)$. Each rule is of the form $r_k = p_k \rightarrow q_k$, for all $k = 0, \ldots, 3$. We can also express this rule list as $d = (d_p, \delta_p, q_0, K)$, where $d_p = (p_1, p_2, p_3)$, $\delta_p = (1, 1, 1, 1)$, $q_0 = 0$, and $K = 3$. This is the same 3-rule list as in Figure 1, that predicts two-year recidivism for the ProPublica data set.

Angelino, Elaine et al. "Learning Certifiably Optimal Rule Lists." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2017): n. pag.

Definition 2

Definition 2: **(Measure of success of a probabilistic reconstruction attack) [1].** Let $\mathcal{V}^{Orig}$ be a deterministic dataset composed of n examples and d attributes, used to train a machine learning model M. Let $\mathcal{V}^{M}$ be a probabilistic dataset reconstructed from M. By construction, $\mathcal{V}^{M}$ is compatible with $\mathcal{V}^{Orig}$. The success of the reconstruction is quantified as the average uncertainty reduction over all attributes of all examples in the dataset:

$$\mathsf{Dist}(\mathcal{V}^{M}, \mathcal{V}^{Orig}) = \frac{1}{n \cdot d} \sum_{i=1}^{n} \sum_{k=1}^{d} \frac{H(\mathcal{V}_{i,k}^{M})}{H(\mathcal{V}_{i,k})} \qquad (1)$$

Perfect Case

$$\boxed{\mathcal{V}^{M} = \mathcal{V}^{Orig}, \ \mathsf{Dist}(\mathcal{V}^{M}, \mathcal{V}^{Orig}) = 0:}$$

Definition 3: Generalized Probabilistic Datasets that consider **Dependence & Non-uniform distribution**

Definition 3: **(Generalized probabilistic dataset).** A generalized probabilistic dataset $\mathcal{W}$ is composed of n *examples* $\{x_1, \ldots, x_n\}$ (the dataset's rows), each consisting in a vector of d attributes $\{a_1, \ldots, a_d\}$ (the dataset's columns). The knowledge about attribute a_k of example x_i is modeled by a probability distribution over all the possible values of this attribute, using random variable $\mathcal{W}_{i,k}$. Importantly, variables $\{\mathcal{W}_{i\in[1..n],k\in[1..d]}\}$ are not necessarily statistically independent from each other and can follow any arbitrary distribution. Each possible instantiation $w = \{w_{i\in[1..n],k\in[1..d]}\}$ of the $\mathcal{W}_{i\in[1..n],k\in[1..d]}$ variables (*i.e.*, each deterministic dataset compatible with $\mathcal{W}$) is named a *possible world*. We let $\Pi(\mathcal{W})$ denote the set of possible worlds within $\mathcal{W}$: $\Pi(\mathcal{W}) = \{w \mid \mathbb{P}(\mathcal{W}_{i\in[1..n],k\in[1..d]} = w_{i\in[1..n],k\in[1..d]}) > 0\}$.

Definition 4: Generalized measure of success of a probabilistic reconstruction attack

Definition 4: **(Generalized measure of success of a probabilistic reconstruction attack).** Let $\mathcal{W}^{Orig}$ be a deterministic dataset composed of n examples and d attributes, used to train a machine learning model M. Let $\mathcal{W}^M$ be a generalized probabilistic dataset reconstructed from M. By construction, $\mathcal{W}^M$ is compatible with $\mathcal{W}^{Orig}$ (*i.e.*, $\mathcal{W}^{Orig} \in \Pi(\mathcal{W}^M)$). The success of the performed reconstruction is quantified as the overall uncertainty reduction in the dataset:

$$\mathsf{Dist}_G(\mathcal{W}^M, \mathcal{W}^{Orig}) = \frac{H\left(\{\mathcal{W}^M_{i,k} \mid i \in [1..n], k \in [1..d]\}\right)}{H\left(\{\mathcal{W}_{i,k} \mid i \in [1..n], k \in [1..d]\}\right)} \tag{2}$$

$$= \frac{\sum_{w \in \Pi(\mathcal{W}^M)} -\mathbb{P}(w) \cdot log_2(\mathbb{P}(w))}{\sum_{i=1}^{n} \sum_{k=1}^{d} H(\mathcal{W}_{i,k})} \tag{3}$$

Results

- optimal models usually represent more information in a more compact way
- the reconstruction uncertainty decreases faster for optimal models than with greedily-built ones.
- Sub-optimal choices can lead to leakage

Personal Questions

- What can be learned from Inherently interpretable models to inform…
  - Privacy leakage
  - Compact Solutions
  - Mimic Inherent Explainability with Post-Hoc methods
- Prior knowledge can be combined with probabilistic dataset to form better attacks
  - Prior knowledge on dependency relationships is very helpful
  - Could this work highlight domain interpretability needs for attacks / security?