On the Privacy Risks of model explanations

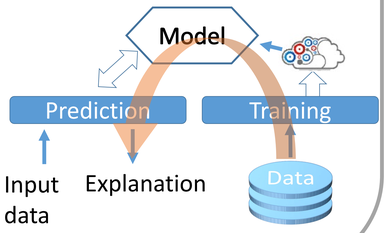
Reza Shokri, **Martin Strobel**, Yair Zick
@ {reza, mstrobel}@comp.nus.edu.sg, yzick@umass.edu

The problem

It is well known that model predictions can leak sensitive information about the training data of a machine learning model. Model explanations try to provide users with helpful information to better understand a model's behavior.
The goal of this work was to answer the question:

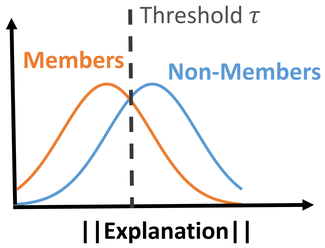
Do model explanations pose an a privacy risk for the training data?

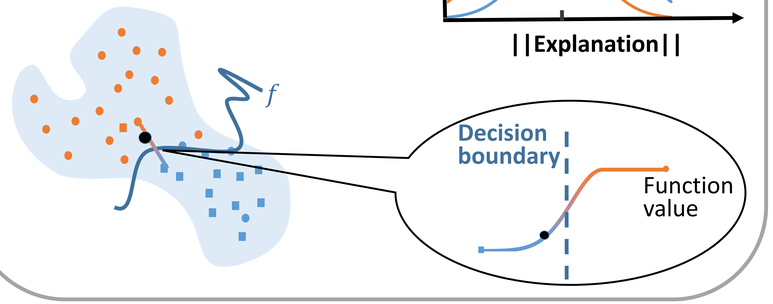
Membership inference attacks try to infer whether or not a data point was used to train a model. They are a standard technique to measure privacy leakage.



Attack methodology

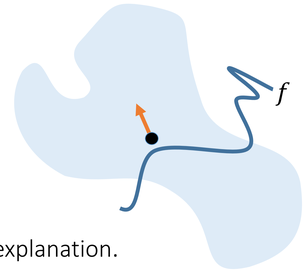
Non-members are more likely to be close to the decision boundary. At the decision boundary the magnitude of gradient-based explanation vectors are higher.



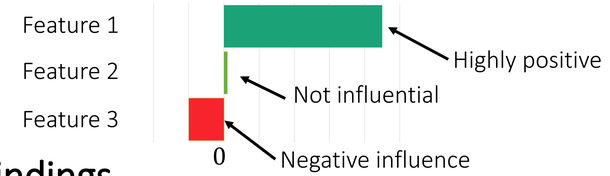


Gradient-based methods

Gradient-based methods try to assign influence to each feature using a single backpropagation through a network. Generally, the influence indicates how important the feature is for the predicted class.
For example, your credit history is important for the approval of a loan.



The **gradient** can be seen as a canonical explanation.

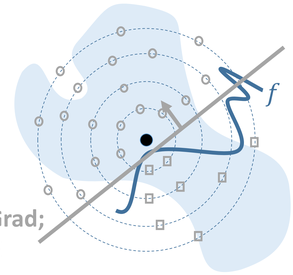


Key findings

- Explanations can leak information about membership
- Magnitude of the gradient vector is a considerable distinguisher between the members of the training set, and other data points from the same distribution

Perturbation-based methods

Perturbation-based methods try to assign influence to each feature using many perturbed queries around the to-be-explained point. Generally, the output resembles gradient-based methods. Examples of this approach are SmoothGrad, Shapley and Lime.



Key findings

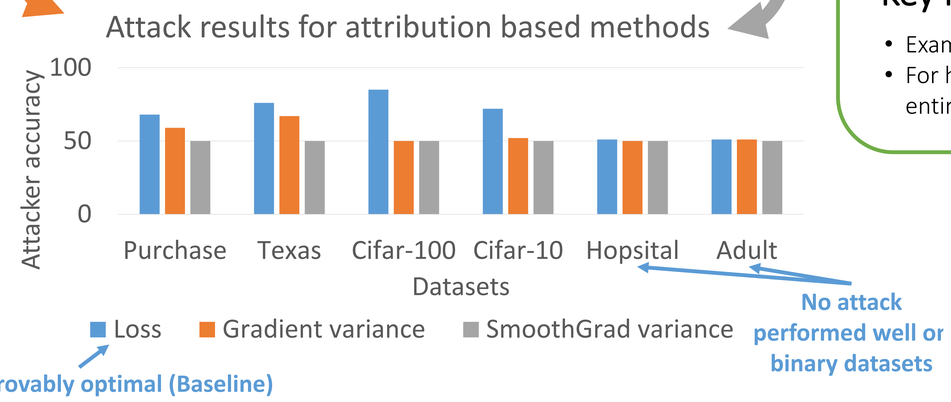
- No existing attack performs better than random-guessing
- Robustness seems to achieved through sampling around a point instead of using the point directly; yet, this sampling has been linked to flaws of this explanation type

Example-based explanations

Example-based explanations use training points to explain model predictions. These records could be influential to the prediction or just seem similar to the given input.

Key findings

- Example-base explanations are a clear violation of privacy
- For high dimensional data an attacker can reconstruct (almost) the entire dataset possible



UMass Amherst

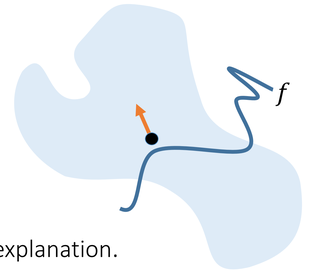
NUS National University of Singapore

On the Privacy Risks of model explanations

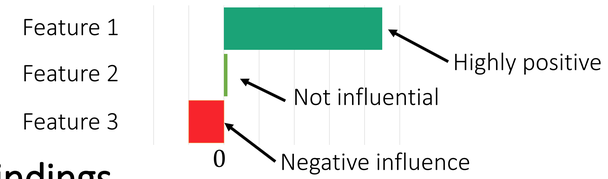
Reza Shokri, **Martin Strobel**, Yair Zick
@ {reza, mstrobel}@comp.nus.edu.sg, yzick@umass.edu

Gradient-based methods

Gradient-based methods try to assign influence to each feature using a single backpropagation through a network. Generally, the influence indicates how important the feature is for the predicted class.
For example, your credit history is important for the approval of a loan.



The **gradient** can be seen as a canonical explanation.

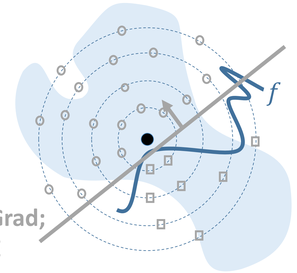


Key findings

- explanations can leak information about membership
- magnitude of the gradient vector is a considerable distinguisher between the members of the training set, and other data points from the same distribution

Perturbation-based methods

Perturbation-based methods try to assign influence to each feature using many perturbed queries around the to-be-explained point. Generally, the output resembles gradient-based methods. Examples of this approach are SmoothGrad, Shapley and Lime.

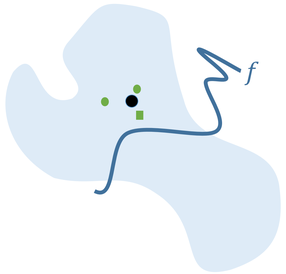


Key findings

- no existing attack performs better than random-guessing
- Robustness seems to achieved through sampling around a point instead of using the point directly; yet, this sampling has been linked to flaws of this explanation type

Example-based explanations

Example-based explanations use training points to explain model predictions. These records could be influential to the prediction or just seem similar to the given input. For example, the outcome of a similar applicant is important for the approval of a loan.

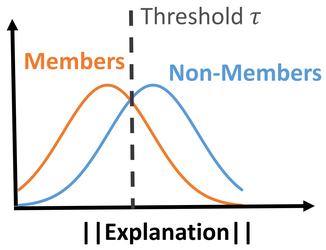


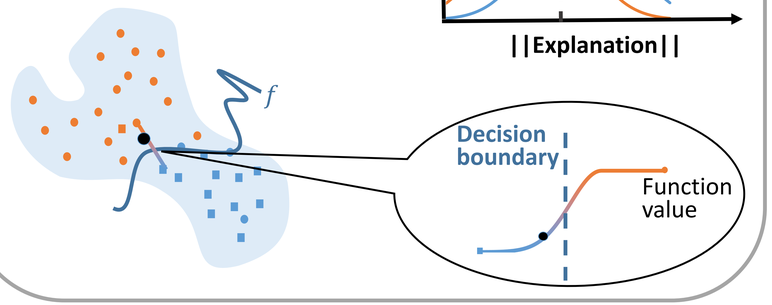
Key findings

- for high dimensional data an attacker can reconstruct (almost) the entire dataset possible
- attacker is likely to recover at least all points in the largest strongly connected component in the graph induced by the influence function
- minorities have a higher risk of being revealed

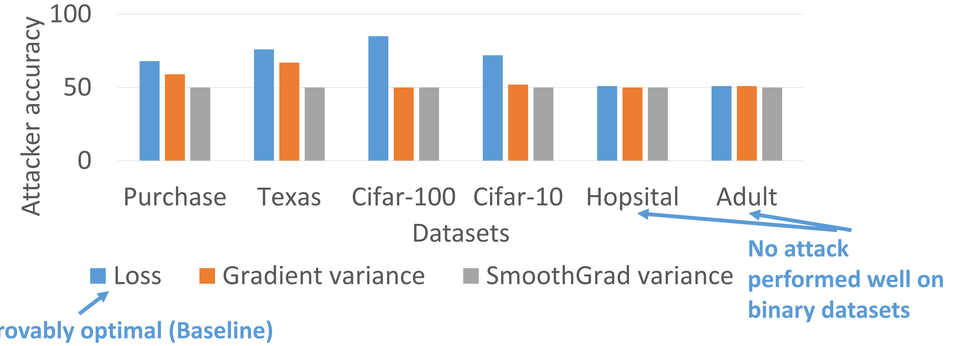
Attack methodology

Non-members are more likely to be close to the decision boundary. At the decision boundary the magnitude of gradient-based explanation vectors are higher.

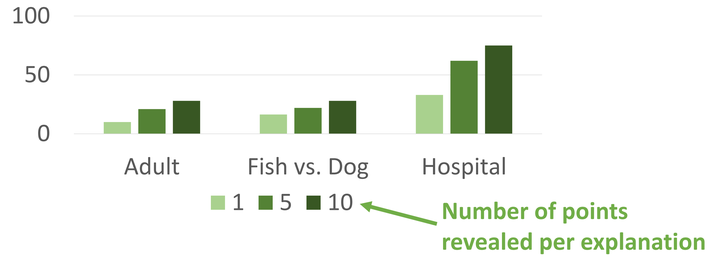




Attack results for attribution based methods



% of training points revealed as their own explanation



UMass Amherst

NUS National University of Singapore