Friday Live Exercises

Machine Learning

A digital healthcare startup in Lausanne wants to launch a new application that costs 2.- CHF and allows users to assess their risk of getting a cold during the next winter.

The application offers an interface to a classifier trained on a dataset of patients from a big hospital in Lausanne. The data to be given to the interface looks like:

```
<race, profession, age, NPA, had_respiratory_disease>
```

Given this feature vector, the classifier returns whether you're at risk of contracting a cold (think 0 or 1) and the confidence of the prediction.

The startup asks you to provide a privacy evaluation of this service.

What *privacy* risks can you think of, and towards whom? For each risk, describe the adversary and their goal (not the attack).

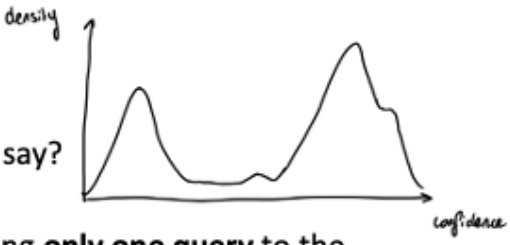
1) Model extraction: competition company tries to steal model's weight from predictions, and propose another application running on the stolen model for 1.99 CHF on the market.
2) Attack on the patient dataset: try to infer information about
    a) the dataset and its distribution (property)
    b) patients in the training set: who is in the dataset (membership)
    c) for a given patient, whether a sensitive field like "had_respiratory_disease" is true, given the other info (attribute).

Other attacks are of course possible (unethical use of the application's prediction, poisoning) but do not fall under the *privacy* notion.

Confidence and privacy I - Membership

You know for sure that you are not in the patients' dataset.

1. When you enter your data in the application, do you expect the confidence of the prediction to be high, low, or "*it depends*"? If the latter, explain on what.
2. How can you get the confidence distribution?
3. Assume you get this density function, what can you say?



4. Then, design a ML attack on an arbitrary target ***z*** using **only one query** to the application.

1) By default, the confidence is expected to be lower. However, the right answer is: it depends! Indeed if the model learned on patients that are similar to you (i.e. you have "neighbors" in the training set) and the model is a good model that is able to generalize, you might get a prediction with high confidence.
2) Many queries (random or carefully chosen).
3) Confidence is part into 2. Following the intuition of 1), the confidence is **expected** to be lower for unseen users. So, low confidence seems to indicates that query likely not in the training set, and high confidence indicates that model have seen that patient during training. This is a total assumption, that might not be true, but assumed to be in this toy application.
4) Define a threshold of in / out, here at the middle, query on the target z. If the confidence of the prediction on z is lower than threshold, say not in training set, if higher, say in the training set.

Confidence and privacy II - Membership

1. If you cannot query for ***z,*** but you are allowed more queries to the application, is a Membership Inference Attack (MIA) on ***z*** still possible?

2. You dig deeper into your extraction of confidence scores and notice that the distribution is somewhat different for a specific population (cryptographers):

    - Would your previous attack still work?
    - How can you modify your attack then?
    - How does it apply to other populations?

3. Do you think there are populations or individuals that are more vulnerable to your attack than others?

1) Yes, because a (not stupid) model learns many things from one point. Hence, the presence of each point in the training data will influence the prediction and confidence of the application on many point. Whereas the change in behavior with/without z might be the most noticeable on the output of the query on z itself, it is possible to infer membership of z through other points (easy example, noisy version of z). So yes, you can still do a membership attack on z without querying on z.
2) The middle of the graph threshold is not going to work as you will always say “in”. You need a special threshold for the cryptographers if you want your attack to work. More generally, your attack can gain in accuracy by having per-population thresholds.
3) Yep. The model can fail to generalize on some populations (for example underrepresented populations) and have a bigger confidence gap, which make the attack more accurate.

Confidence and privacy III

1. How would you propose to defend from your attack? (What do you need and how do you get it?)
2. What do you expect to change about the distribution of confidence?
3. We talked in class about DP at record level, and how it would protect from MIAs (not just the one we discussed).
   - What about attribute inference attack?
   - What about property inference?

1) Enforce generalization, for example through DP. You can also think or removing the patients/populations you cannot protect from. According to the minimum disclosure principle, confidence can be removed and prediction given as it with a disclaimer notice. Queries can be limited.
2) The distribution on train data and test data becomes similar. Often, removing overfitting can lead in the accuracy on training instances to go down. Based on our assumption that the two spikes are respectively non-members of training data and members, one possibility is that the right spike shifts to the left, and the two spikes collude, but it is not true in general.
3) Protection against attribute inference attacks is implied by DP. Property inference attacks are not protected from.