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).

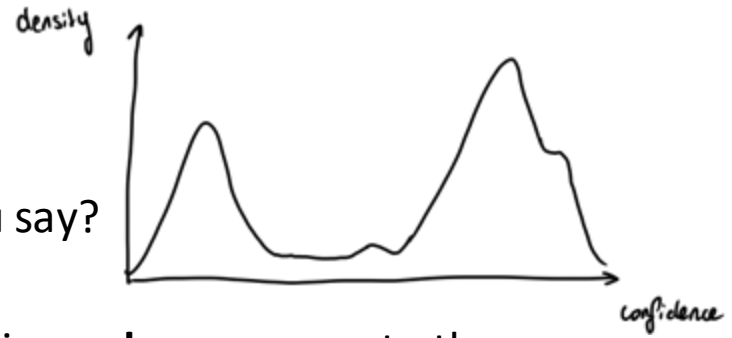
Confidence and privacy I - Membership

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?



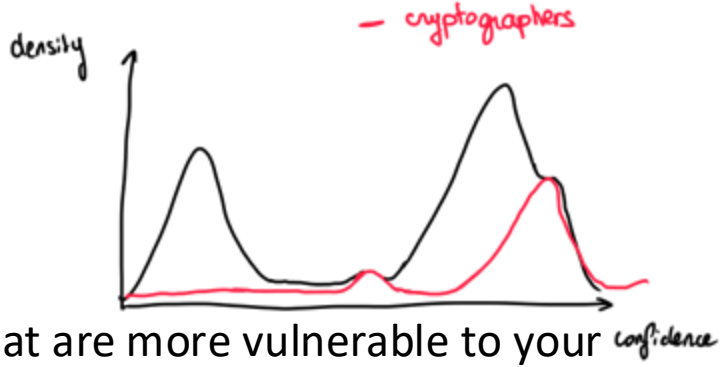
4. Then, design a ML attack on an arbitrary target ***z*** using **only one query** to the application.

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?

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?