CS-523 Advanced Topics on Privacy Enhancing Technologies Machine Learning Exercises

1 Base Rates are Important

Standard classification metrics such as true positive and true negative rates give a good indication of a classifier's performance when classes are balanced. However, they do not take into account the *base rates* of classes, thus in imbalanced settings they are misleading.

Consider an adversary that mounts an attribute inference attack. Using access to a medical dataset, they are trying to infer whether different people who were present in the dataset have a disease A. The disease is not directly mentioned in the dataset, hence the adversary has built a classifier C to infer this. The true positive rate of the classifier $\Pr[C = 1 \mid A = 1] = 0.98$, and the false positive rate is $\Pr[C = 1 \mid A = 0] = 0.01$. The disease is very rare: its prevalence (base rate) in the general population is $\mu = \Pr[A = 1] = 0.0001$. Assume this is the best prior the adversary has.

One intuitive classification metric that takes into account the base rate is *Bayesian detection rate*, or *positive predictive value*. In this setting, it is defined as the probability that a person has the disease given that the classifier predicted so: $\Pr[A = 1 \mid C = 1]$. Contrast this to the true positive rate $\Pr[C = 1 \mid A = 1]$: the probability that the classifier predicts disease if a person has the disease.

1. Compute the adversary's Bayesian detection rate.
2. What would the false positive rate of the adversary's classifier have to be so that Bayesian detection rate is reasonable?
3. What would the base rate have to be so that the Bayesian detection rate is reasonable?

2 Learning with Strangers

A movie review site has operated a recommendation sharing system to suggest movies to users for years. The site gathered users' ratings and applied gradient descent in a central fashion, but they have decided to change this central database to a privacy-preserving alternative.

1. The site decides that every user should keep his/her data locally. Each user retrieves the model from the server, computes, and sends a gradient update to the server. Is this approach private? How can the site change this approach to improve privacy?

2. The site enables each user to apply a locally differentially private perturbation to the updates. Either the noise magnitude should be low or there will be a drastic reduction in functionality. The site groups online users together and only allows users to rate movies when their group has at least n users. Each group randomly chooses a leader that collects updates from users, without noise, aggregates them together, applies a group noise, and sends the aggregate update to the server. Assess the functionality and privacy of this approach.

3. The site decides to stop sending plain updates to the group leader and replaces the role with an SMC computation. Is it safe to remove the differentially private noise now that the update is performed on encrypted data?

4. After the launch of a competitor service, the site managers are worried about malicious users sabotaging the model. Assess the resilience of the model against malicious users in the central approach and privacy-preserving alternatives.

3 Membership and attribute inference: what's the connection?

In the class you learned about several types of privacy attacks.

One of them is membership inference attacks (MIA for short), where the adversary aims to learn whether a target data point $x \in \mathcal{X}$ is in the training set or not. Given a data point x, the output of the MIA is either "in" or "out".

Another example is attribute inference attacks (AIA), where the adversary aims to learn the value of a sensitive feature of a target data point, given some public knowledge about the point. We expand data points as a tuple $x = (v, t) \in \mathcal{X} = \mathcal{V} \times \mathcal{T}$, where $v \in \mathcal{V}$ is the public knowledge and $t \in \mathcal{T}$ is the sensitive feature. Inferring the sensitive feature is the target of the AIA. You can assume in this exercise that the values of t are uniformly distributed over $\mathcal{T}$. Given the public knowledge v, the output of the AIA is a value $t_i \in \mathcal{T}$.

Your task for this exercise is to explore the connection between MIA and AIA.

1. Construct an AIA using a MIA in a black box manner.

2. Construct a MIA using an AIA in a black box manner.

3. What can you conclude?