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.

   Solution:

   $$\Pr[A = 1 \mid C = 1] = \frac{\Pr[C = 1 \mid A = 1] \cdot \mu}{\Pr[C = 1 \mid A = 1] \cdot \mu + \Pr[C = 1 \mid A = 0] \cdot (1 - \mu)}$$

   This equals approximately 0.01. This probability is rather low, indicating that the adversary's attack is not effective. In fact, about 99% of adversary's detections will be false positives: $\Pr[A = 0 \mid C = 1] = 0.99$.

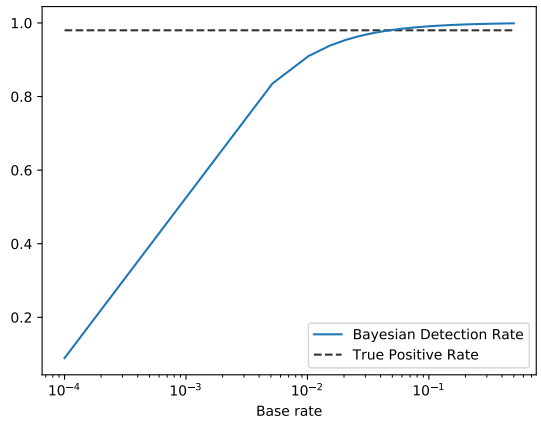
2. What would the false positive rate of the adversary's classifier have to be so that Bayesian detection rate is reasonable?

   Solution:
   In general, "reasonable" will heavily depend on the context. Let's say

that the adversary can tolerate 50% of effective "false positives." Hence, they want $\Pr[A = 1 \mid C = 1] = 0.5$. Depending on the context, combing through these false positives might be too much of a burden for the adversary, but it could be that this is acceptable as the adversary wants to have some answers at all costs. This approximately results in the maximum classifier's false positive rate $\Pr[C = 0 \mid A = 1] \leq 0.0001$. Thus, in this setting if the false positive rate is as low as the base rate, half of the detections are false positives.

3. What would the base rate have to be so that the Bayesian detection rate is reasonable?

   Solution:
   Increasing the base rate by one order of magnitude to 0.001 increases the Bayesian detection rate to 50%. In general, the following plot shows the relationship for this setting (x-axis is logarithmic):



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?

   Solution:
   No. The update is dependent on the data, and the site can infer informa-

tion from updates. To preserve privacy, users would could add noise to the update that satisfies local differential privacy with small ε.

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.

   Solution:
   Applying perturbation to an aggregate update requires less noise than applying the noise individually, which improves the functionality. However, instead of the server, the group leader can observe the raw update of online users, which undermines privacy. There is no guarantee that the leader will be honest and a malicious server can fake many online users.

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?

   Solution:
   The group noise is applied to the aggregate update to prevent the site from extracting information about users. The SMC only hides the raw updates from the group leader and does not change the site's view, so this change does not affect the necessity of the group noise.

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.

   Solution:

   - Central: each malicious account can only submit one rating per movie, so the impact is low.
   - Individual noise: a malicious user can send a fake direction with a large magnitude to the server to directly impact the model. However, the server can observe all updates one by one and discard suspicious ones (limiting the magnitude). Therefore, limiting the attack.
   - Group noise: the server only sees the aggregate update and the leader is in charge of checking inputs. If a malicious user gets chosen as the leader, then it can send a fake update with a magnitude relative to the group size.

- Encrypted: no one can control individual updates, so malicious users can easily corrupt the data. It's possible to adjust the SMC to enforce correctness, but it's costly.

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?

Solution:

1. The adversary can run the MIA for each possible value t_i of the sensitive attribute. When the output of the membership attack on $x_i = (v, t_i)$ is "in", the adversary can guess that t_i is the correct value of the sensitive attribute.
2. Given the target $x = (x_v, t_x)$, the adversary can run the AIA for the sensitive feature with publicly known x_v, and compare the output t to the actual value of the feature t_x. If the output matches, the adversary guesses that the data point is a member of the training set.
3. This suggests that attribute inference is at least as difficult as membership inference and vice versa. Hence, under the assumption that t is uniformly distributed, an ML system is vulnerable to MIA iff it is vulnerable to AIA.

If you want to read more about this, the section 5 of Privacy Risk in Machine Learning by Yeom et al. contains extensive explanations.