HW 4A: The Local Model

CS 208 Applied Privacy for Data Science, Spring 2019

Version 1.1: Due Tuesday, April 16, 11:59pm.

Instructions: Submit a single PDF file containing your solutions, plots, and analyses. Make sure to thoroughly explain your process and results for each problem. Also include your documented code and a link to a public repository with your code (such as GitHub/GitLab). Make sure to list all collaborators and references.

1. **Learning Conjunctions in the SQ Model:** In this problem, you will see a simple example of a machine learning algorithm that works in the statistical query (SQ) model, and will evaluate how it performs under both local and centralized DP.

   The set-up is that we are given a data set $D = ((x_1, y_1), \ldots, (x_n, y_n))$ where $x_i \in \{0,1\}^d$ and $y_i \in \{0,1\}$. We want a (local or centralized) differentially private algorithm $M(D)$ that outputs a subset $\hat{S} \subseteq \{1, \ldots, d\}$ of the variables such that the conjunction of the x-variables in $\hat{S}$ predicts the y variable well.

   Specifically, following Valiant's PAC model, we will measure utility as follows. Suppose that D consists of n iid draws from an (unknown) distribution $\mathcal{P}$ on $\{0,1\}^d \times \{0,1\}$. Furthermore, suppose that there is an (unknown) set $S \subseteq \{1, \ldots, d\}$ such that

   $$\Pr_{(x,y)\sim\mathcal{P}}\left[y = \bigwedge_{j\in S} x[j]\right] = 1.$$

   That is, y can be *perfectly* predicted by some conjunction.[1] (Note that here and below, the notation (x, y) refers to a *single* labelled example drawn from $\mathcal{P}$, not a dataset of n such values. We will use $D = ((x_1, y_1), \ldots, (x_n, y_n))$ for the dataset.)

   Then the goal of M is to output $\hat{S}$ that minimizes the expected classification error:

   $$\mathop{\mathrm{E}}_{D\sim\mathcal{P}^n, \hat{S}\leftarrow M(D)}\left[\Pr_{(x,y)\sim\mathcal{P}}\left[y \neq \bigwedge_{j\in\hat{S}} x[j]\right]\right].$$

   (This is an instantiation of the "risk minimization" problem described in the April 8 lecture, though you will not need anything from that lecture for this homework problem.)

   The SQ algorithm is based on estimating the following quantity for each variable $j = 1, \ldots, d$:

   $$p_j = \Pr_{(x,y)\sim\mathcal{P}}[x[j] = 0 \wedge y = 1].$$

   As shown in section, we have:

   i For each $j \in S$, we have $p_j = 0$.

[1] This is known as the "realizable" case of learning, as opposed to the "agnostic" case, where no conjunction classifies perfectly, but the goal is to find one that does as well as possible.

ii If $\hat{S} \supseteq S$, then the false positive rate of $\hat{S}$ is 0.

iii The false negative rate of $\hat{S}$ is at most $\sum_{j \in \hat{S}} p_j$.

Based on these observations, an SQ algorithm M for learning conjunctions works as follows:

1 Using the dataset D, obtain an estimate $\hat{p}_j$ of p_j for $j = 1, \ldots, d$.

2 Output $\hat{S} = \{j \in [d] : \hat{p}_j \leq t\}$ for an appropriate (small) threshold t.

You should do the following:

(a) Describe and implement centralized and local DP versions of the above SQ algorithm, dividing the privacy budget equally among each of the d estimates $\hat{p}_j$. Keep the threshold t as a free parameter that you can choose.

(b) For both the centralized and local DP algorithms you give in Part 1a, analytically estimate $\Pr[\hat{S} \not\supseteq S]$ as a function of t, n, ε, and $|S|$. Feel free to use a normal approximation for the binomial distribution, and to describe your estimate in terms of the CDF of a standard normal distribution. Use this as a guide to determine a setting of t to ensure that $\Pr[\hat{S} \not\supseteq S] \leq .1$ as a function of n, d, and ε in each of the two settings, using the fact that $|S| \leq d$.

(c) An interest group targets a particular online political advertisement to users of a social media platform based on a conjunction of demographic characteristics. The users who are targeted with the advertisement are recorded. You get differentially private access (at $\epsilon = 1$) to the user data and attempt to learn what demographic combination was being focused on.

In the dataset `CaPUMS5Full.csv` are a number of potential demographic characteristics that might have been used for targeting: `(sex, married, black, asian, collegedegree, employed, militaryservice, uscitizen, disability, englishability)`. The variable `targeted` records those individuals who received the ad. Use your DP algorithms to learn a classifier to predict `targeted=1` (which we have artificially created) from these variables. Demonstrate whether your centralized and local algorithms would succeed. Also, use a bootstrap to compare the false positive rates and false negative rates of the two methods as a function of n.

If useful, also included in this dataset is the variable `blackfemale` that you can use to test if your implementation correctly picks up `black` and `female` as the constituent parts. There is also a test dataset, `hw4testdata.csv`, that contains `x1` $\cdots$ `x10`, and `y=x1∧x2∧x3`.