18734 Homework 5

Due: 12 Noon Eastern, 9 AM Pacific, Dec 7

NOTE: In this homework, problems 1, 2, and 3 are compulsory. Your score for problems 1, 2, and 3 will be scaled out of 65 points. For the remaining 35 points, you can choose to do any combination of problems 4, 5, 6, or 7 that makes up at least 35 points. The last 35 points of your homework will be a percentage of your score on the questions you attempt. Please state clearly which questions you want to use for grading purposes. For example, if you state problem 4 and problem 7 and you get 45/45 then it will be scaled down to 35/35.

Problem 1: QII measures (10 marks)

Decision-making systems that use machine learning are increasingly used to aid decision making in today's society. Such decisions could be about online personalization or credit and insurance decisions. Unfortunately, it is often difficult to explain why a certain decision was made. Datta et al 2016 introduce a family of Quantitative Input Influence (QII) measures to capture the degree of influence of inputs on outputs of systems that can explain decisions about individuals (e.g., a loan decision) and groups (e.g., disparate impact based on gender). Because a single input may not always have a high influence, QII measures also quantify the joint influence of a set of inputs (e.g., age and income) on outcomes (e.g. loan decisions), and the marginal influence of an individual input within such a set (e.g., income). The authors found that QII measures are able to provide explanations for several scenarios when black box access to the learning system is available, and can be made differentially private while preserving accuracy [2].

An implementation of the QII measures over several data sets is available on Blackboard under the "Recitations" folder. Download the zip file and follow instructions in the README to run the tool.

1. Use the tool to investigate the provided nlsy97 dataset.
   (a) Display a figure of the QII on outcomes. When the sensitive attribute is *gender*, which attribute is most strongly influenced?
   (b) Display the figure of the QII on group disparity. Given the sensitive attribute is *gender*, which attribute has the greatest negative impact on outcomes? Justify.
2. Use the tool to analyze this dataset on fertility: `https://archive.ics.uci.edu/ml/datasets/Fertility`. This may require you to manipulate the dataset for processing by the tool. The sensitive attribute is "surgery".

(a) Display a figure of the QII on outcomes. When the sensitive attribute is *surgery*, which attribute is most strongly influenced?

(b) Display the figure of the QII on group disparity. Given the sensitive attribute is *surgery*, which attribute has the greatest negative impact on outcomes? Justify.

Problem 2: Implementing Laplace mechanism (30 marks)

In this problem you will implement Laplace mechanism to provide differential privacy for the provided database. The first part that you need to think about is how to sample from a Laplace distribution. In this problem you are free to use C++, Java or Python. You will need to figure out how to generate random real numbers between $[0, 1]$. (e.g., for Java you can see `http://java.about.com/od/javautil/a/randomnumbers2.htm`)

1. Assume you have access to a sampler that samples from the uniform distribution over $[0, 1]$. Read about inverse transform sampling `http://en.wikipedia.org/wiki/Inverse_transform_sampling`.

   (a) Write the CDF function F for Laplace distribution with mean 0 and variance $2\lambda^2$.

   (b) Use inverse transform sampling and the uniform distribution sampler to write code to sample from a Laplace distribution with mean zero and variance $2\lambda^2$. (Write a function where the variance is given as an argument).

2. Download the scores.txt file[1]. It has the scores of 10,000 students with scores between 1 and 100. You are asked to provide differential privacy for the query `average`.

   (a) State the global sensitivity of the query.

   (b) Using your implementation of sampling from Laplace distribution, write code to provide 0.001-differential privacy for the above query.

   (c) Repeat the following 100 times (not manually, but, in the code)
   In each iteration i repeat the query n_i times till the average of the n_i answers is 10^{-4} close to the true average.
   Report the number $\frac{\sum_i n_i}{100}$.

Note that the number $\frac{\sum_i n_i}{100}$ is an estimate of the number of times you can query before you reveal the true average with error of less than 10^{-4}.

Problem 3: Anonymous Communication (35 marks)

General Protocol (15 marks)

In the lecture, we discussed the Dining Cryptographers protocol. In this problem, we will explore how to use that protocol as a building block to construct a general protocol for anonymous communication. Consider a group of n agents. (You may want to read this `http://users.ece.cmu.edu/~adrian/731-sp04/readings/dcnets.html`)

[1] `https://www.ece.cmu.edu/~ece734/f16-18734/homework/scores.txt`

1. Describe a protocol using which one of the n agents can send an m-bit message. Explain informally why the protocol is *correct* (i.e., all agents receive exactly the message that was sent) and *anonymous* (i.e., none of the other agents have any clue who the real sender is).
2. State and prove rigorously that anonymity is preserved by the protocol for the case where $n = 4$ and $m = 1$. (You need to show that from the point of view of any non-sender, the probability of any of the other agents being the sender is 1/3).
3. How many bits of randomness and how many message transmissions are needed to complete this protocol with n agents and an m-bit message?
4. How robust is this protocol to collusion, i.e., if k out of the n non-sender agents collude, what is the probability that they can figure out who the real sender is?

Hidden services (5 marks)

Tor can also provide anonymity for servers, apart from providing anonymity for clients. Read the relevant part of the paper `https://svn.torproject.org/svn/projects/design-paper/tor-design.pdf` and explain how hidden services work in Tor. The explanation **must be a bulleted list of the main points**. Marks will be deducted for writing paragraphs.

Nymble (15 marks)

Tor can sometimes lead to some undesirable consequences. This problem asks you to look at the paper `http://www.cs.dartmouth.edu/~sws/pubs/jkts07.pdf` and answer the following questions:

1. What potential problem with Tor are identified in the paper?
2. Provide an overview of how the Nymble system works. Section 3 in the paper has such an overview. You can read that overview, however, your answer must be in your own words.
3. List the (informally) cryptographic properties that the Nymble system relies on.

Problem 4: Zero knowledge proofs (20 marks)

Amplification (10 marks)

This problem gives you the essence of amplification: going from a small difference in the completeness and soundness probabilities for interactive proofs to almost 1 difference in these probabilities. As mentioned in class, this involves repetition of the interactive protocol. Below, we ask a question on probability that demonstrates amplification. Suppose there are two types of coins: one made with bronze and one with gold. Your task is to figure out if a given coin in made of bronze (B) or gold (G). You also know that bronze coins produce heads with probability $\frac{1}{2} + \delta$ and gold coins produce heads with probability $\frac{1}{2} - \delta$.

You come with the idea that if you flip the coin k times and take the majority vote, then an answer of head indicates a bronze coin and an answer of tails indicates gold coin (with high probability). To be absolutely sure that the idea is right lets do the following computation.

1. You want to bound the probability of committing a mistake. Suppose the true coin type is X (which is either B or G). You make a mistake when your final answer is not X. And your final answer if not X when the majority of coin flips indicate tails when $X = B$ or indicate heads when $X = G$. But, note that in either case (B or G) the side of the coin that showed up in majority has a probability $\frac{1}{2} - \delta$ of occurring. Thus, we get the following

$$P(mistake) = P\left(\geq k/2 \text{ events with probability of each event } \frac{1}{2} - \delta\right)$$

2. Next, note that the right hand side of the above equation can be written as

$$P\left(\bigcup_{i=0}^{k/2} k/2 + i \text{ events with probability of each event } \frac{1}{2} - \delta\right)$$

3. Using union bounds (`http://en.wikipedia.org/wiki/Boole's_inequality`) show that the above value is bounded by

$$\sum_{i=0}^{k/2} P\left(k/2 + i \text{ events with probability of each event } \frac{1}{2} - \delta\right)$$

4. Show that

$$P\left(k/2 + i \text{ events with probability of each event } \frac{1}{2} - \delta\right) = \binom{k}{k/2+i}(\frac{1}{2}-\delta)^{k/2+i}(\frac{1}{2}+\delta)^{k/2-i}$$

5. Argue that the above value in 4 is less than $\binom{k}{k/2+i}(\frac{1}{2} - \delta)^{k/2}(\frac{1}{2} + \delta)^{k/2}$

6. Using the above bound in 5, argue that sum in 3 is less than $2^k(\frac{1}{2} - \delta)^{k/2}(\frac{1}{2} + \delta)^{k/2}$, which can easily be reduced to $(1 - 4\delta^2)^{k/2}$. Thus, the probability of mistake is bounded by $(1 - 4\delta^2)^{k/2}$.

Now, clearly by choosing many repetitions k, the probability of mistake becomes very small for any constant $\delta > 0$.

Simulation (10 marks)

This problem asks you to argue informally why the distribution generated by the prover and verifier (denoted as $\langle P, V\rangle$) is computationally indistinguishable from the distribution generated by the simulator (denoted as $\langle M\rangle$).

The simulation is shown in Figure 1. Given $\langle N, Y, g\rangle$, P claims to know s, such that $Y = g^s \mod N$. s is known as the discrete logarithm of Y given $\langle N, g\rangle$ and finding it is known to a be difficult problem given Y, N, g [2].

1. P picks a random r. Then she sends $Y = (g^s \mod N)$ and $A = (g^r \mod N)$ to V.

2. V picks a random challenge c and sends it to P.

[2] `http://en.wikipedia.org/wiki/Discrete_logarithm`

3. P then sends over $z = r + cs \mod (N-1)$.

V accepts the proof iff $AY^c = g^z \mod N$.

Note that the distribution $\langle P, V \rangle$ has four (single dimensional) random variables corresponding to the four values that are exchanged in the interaction. You have to argue separately for each random variable, why the distribution of that random variable in $\langle P, V \rangle$ is roughly same as the distribution of the corresponding random variable in $\langle M \rangle$.

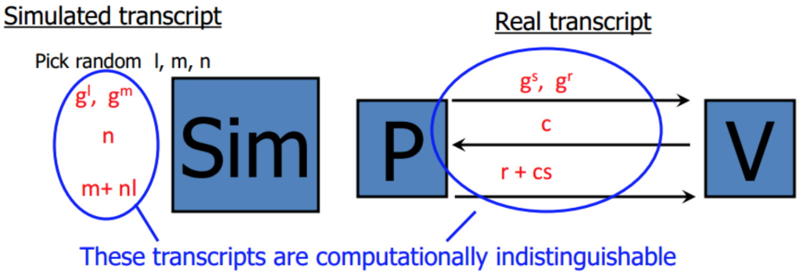

Figure 1: Simulation.

Problem 5: Oblivious Transfer (15 marks)

Consider a simplified version of the oblivious transfer protocol proposed by Even, Goldreich and Lempel. Let R be the receiver with input $b \in \{0, 1\}$ and let S be the sender with input $x_0, x_1 \in \mathcal{X}$. Assume both S and R have access to an oracle $\mathcal{O}_{map} : \mathcal{X} \to \mathcal{X}$, which returns an element in $\mathcal{X}$ as well as an inversion oracle such that

$$\forall x \in \mathcal{X} : \mathcal{O}_{inv}(\mathcal{O}_{map}(x)) = x$$

To transfer x_b from the sender to the receiver, they carry out the following steps:

1. R sends c_0, c_1 to S, where $c_b \leftarrow \mathcal{O}_{map}(r)$ for $r \overset{\$}{\leftarrow} \mathcal{X}$ and $c_{1-b} \overset{\$}{\leftarrow} \mathcal{X}$, where $a \overset{\$}{\leftarrow} \mathcal{X}$ indicates that a is randomly selected from $\mathcal{X}$.
2. S replies with d_0, d_1, where $d_0 \leftarrow \mathcal{O}_{inv}(c_0) \oplus x_0$ and $d_1 \leftarrow \mathcal{O}_{inv}(c_1) \oplus x_1$.
3. R reconstructs $x_b \leftarrow r \oplus d_b$.

Correctness (5 marks)

For the above scheme, show that the receiver indeed recovers x_b.

Obliviousness (5 marks)

Give an informal argument for why the sender cannot determine the b at the end of the protocol.

Extending OT (5 marks)

Suppose you have access to a device that can perform 1-out-of-2 oblivious transfer. How would you use this device to create a device that can perform 1-out-of-3 oblivious transfer? You may have to use the device more than once.

Problem 6: Disparate impact (10 marks)

Data-trained predictive models are often used as black boxes that output a prediction that is used to aid decision making. Thus, it is hard to determine whether sensitive attributes such as race or gender are unduly influencing model predictions. Adler et al 2016 [1] developed a method to audit black-box models. Their tool uses techniques developed to detect and repair disparate impact in classification models to study the sensitivity of modles with respect to its features. For instance, the tool could be used to quantify the amount race contributes to sentencing decisions in court for felonies.

In this problem, you will use their tool available at `https://github.com/algofairness/BlackBoxAuditing` to study to the extent to which particular features in some real world datasets influence the outcome of classification models.

To install their tool, clone the repository from `https://github.com/algofairness/BlackBoxAuditing` and follow the instructions in the README.

1. Run the tool on the German credit dataset by changing the first line in main.py to `import experiments.german as experiment` and set `response_header = "class"`, and then running `python main.py`.

   Which attributes do you think have the most influence on the credit decision. Why? There may be multiple reasonable answers depending on your argument.

2. Run the tool with 1/3, 2/3, and 1/4 training. on the same fertility dataset used the previous question. The dataset can be downloaded from `https://archive.ics.uci.edu/ml/datasets/Fertility`.

   Display the accuracy graph produced. Based on the results, do you think any of the variables in the dataset strongly influence on the results? Why or why not?

Problem 7: Fairness in Classification (25 marks)

This problem makes use of some techniques described in the paper by Dwork et al 2011 [3].

The table below records some facts about four individuals who are about to take the LSAT test. A "0" means false, while "1" means true. The administrator of the test would like to predict *outcomes*, which is the fraction of test-takes that these individuals will do better than on the test. The possible outcomes is a number between 0 and 1. The space of individuals V is thus the finite set $\{(0,1,1,0,0,0,0),(1,1,1,1,0,0,0),(0,0,0,0,1,1,0),(1,0,0,0,1,0,1)\}$.

Student	Male?	Low Income?	First-Gen?	Alcoholic?	Honors?	Elite College?	Student Leader?
Alice	0	1	1	0	0	0	0
Bob	1	1	1	1	0	0	0
Carole	0	0	0	0	1	1	0
David	1	0	0	0	1	0	1

He needs to choose between two similarity functions to use. $\vec{x}$ represents information about an individual in the table. Assuming $\vec{x} = x_1, \cdots, x_n$ and $\vec{y} = y_1, \cdots, y_n$ and $\vec{x}, \vec{y} \in \{0,1\}^n$.

The similarity functions are:

(a) $s_1(\vec{x}, \vec{y}) = \sum_{i=1}^n |x_i - y_i|$

(b) $s_2(\vec{x}, \vec{y}) = \left|\log\left(\frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n y_i}\right)\right|$

M maps individuals from the space of individuals in the given table (i.e. there are four elements in the space of individuals) to the following probability distributions over outcomes. M attempts to capture the ground truth laid out in this study [4]. These distributions are the laplace distribution with varying parameters.

For example, Carole's entry in the table is (0,0,0,0,1,1,0). Since $|(0,0,0,0,1,1,0)| = 2$, thus M maps her to probability distribution P_2. Thus, Carole, a relatively privileged student, is likely to do better than $P_2((0,0,0,0,1,1,0)) = 86\%$ of all LSAT takers.

$$\begin{aligned}
|\textstyle\sum_{i=1}^n x_i| = 1 &\mapsto P_1(\vec{y}) := \frac{1}{2\times 0.4}\exp\left(-\frac{|\sum_{i=1}^n y_i - 4|}{0.4}\right)\\
|\textstyle\sum_{i=1}^n x_i| = 2 &\mapsto P_2(\vec{y}) := \frac{1}{2\times 0.4}\exp\left(-\frac{|\sum_{i=1}^n y_i - 1.85|}{0.4}\right)\\
|\textstyle\sum_{i=1}^n x_i| = 3 &\mapsto P_3(\vec{y}) := \frac{1}{2\times 0.4}\exp\left(-\frac{|\sum_{i=1}^n y_i - 2.95|}{0.4}\right)\\
|\textstyle\sum_{i=1}^n x_i| = 4 &\mapsto P_4(\vec{y}) := \frac{1}{2\times 0.4}\exp\left(-\frac{|\sum_{i=1}^n y_i - 3.5|}{0.4}\right)\\
|\textstyle\sum_{i=1}^n x_i| = 5 &\mapsto P_5(\vec{y}) := \frac{1}{2\times 0.4}\exp\left(-\frac{|\sum_{i=1}^n y_i - 4.5|}{0.4}\right)\\
|\textstyle\sum_{i=1}^n x_i| = 6 &\mapsto P_6(\vec{y}) := \frac{1}{2\times 0.4}\exp\left(-\frac{|\sum_{i=1}^n y_i - 5.2|}{0.4}\right)\\
|\textstyle\sum_{i=1}^n x_i| = 7 &\mapsto P_7(\vec{y}) := \frac{1}{2\times 0.4}\exp\left(-\frac{|\sum_{i=1}^n y_i - 6.3|}{0.4}\right)
\end{aligned}$$

1. Prove that the similarity functions s_1 and s_2 are metrics on the space V [3].

2. Compute the similarity metric using functions (a) and (b) between the following individuals:

   (I) Alice and Bob

   (II) Alice and David

3. Does individual fairness (described in the lecture on October 26) hold when using either of similarity metric (a) or similarity metric (b)? It may help to write a script for the computations.

[3] https://en.wikipedia.org/wiki/Metric_space

4. In the previous question, you found that one of the similarity metrics does not allow individual fairness to hold. Suggest one modification to the way ground truth about individuals is captured so that individual fairness is preserved for that similarity metric. Justify your reasoning and why it approximately captures the ground truth according to Kidder 2001 [4], which found that students from advantaged backgrounds tend to do better on standardized tests than those from disadvantaged backgrounds.

   Hint: Look at the way the headers of the columns are phrased. Another possibility is to modify the distribution functions.

5. Using your modification above, show that individual fairness is now preserved in the given database of individuals while upholding the assumption that students from advantaged backgrounds tend to do better on standardized tests than those from disadvantaged backgrounds

Submission

Submit these files:

1. Merge all the written parts into a single pdf file ⟨your_andrew_id⟩_HW5.pdf.
2. Rename the program file (.c/.cpp/.java/.py) you used for Problem 1 as ⟨your_andrew_id⟩_laplace.⟨extension⟩.

Zip these files into ⟨your_andrew_id⟩_HW5.zip and submit the zip file on BlackBoard.

References

[1] P. Adler, C. Falk, S. A. Friedler, G. Rybeck, C. Scheidegger, B. Smith, and S. Venkatasubramanian. Auditing black-box models by obscuring features. *arXiv preprint arXiv:1602.07043*, 2016.

[2] A. Datta, S. Sen, and Y. Zick. Algorithmic transparency via quantitative input influence. In *Proceedings of 37th IEEE Symposium on Security and Privacy*, 2016.

[3] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM, 2012.

[4] W. C. Kidder. Does the lsat mirror or magnify racial and ethnic differences in educational attainment?: A study of equally achieving "elite" college students. *California Law Review*, pages 1055–1124, 2001.