Data Privacy and Security: Homework #1

Due on November 16, 2020 at 11:59pm

Professor: Mehmet Emre Gürsoy
TA: Çağhan Köksal

Muhammet Soytürk

Problem 1

Principle #1: Proactive not reactive; preventive not remedial

Since health data is sensitive and not accessed frequently, there will be a three factor authentication mechanism for remote access attempts. These three factors will be password, email and short message service.

Principle #2: Privacy as the default setting

Before we build the online system, we will reach each patient that was admitted in the past and ask whether they are willing to be a part of the online system. If we cannot reach them, they will not be a part of the system as default. Also, when a new patient arrives, their consent will be asked. If it's an emergency, their data won't be transferred to the online system.

Principle #3: Privacy embedded into design

There will be different layers of access types. For example, a physician will only see patient's physical condition or a nurse can see patient's data only when the patient is currently admitted. The privacy of the system will be subjected to external reviews and audits regularly.

Principle #4: Full functionality – positive-sum, not zero-sum
Only the doctor, nurse or physician of the patient will be able to see patient's data. If the other doctors, nurses or physicians want to see other patients' data to get an insight about the similar patients, they will see anonymized data. I believe that this will be a positive-sum game since the hospital personnel will still be able to get insight about similar patients and the patient's information will not be shared with anyone but his/her doctor, nurse or physician.

Principle #5: End-to-end security – full lifecycle protection
The system will comply to AES encryption method which is the standard encryption method used in several countries. Also, salting will be used to obstruct dictionary attacks. There will also be a strong access control for remote users. The destruction of data will be carried out carefully to ensure that the erased data cannot be recovered.

Principle #6: Visibility and transparency – keep it open

There will be a record for each access to a patient's information. The patients will be able to see when and who accessed their data. Also, regular external reviews and audits will ensure the transparency of the system. Anonymized data samples will be released to increase the transparency of the system.

Principle #7: Respect for user privacy – keep it user-centric

At any time, the patients will have the control to delete their data to empower patients to play an active role in the management of their own data.

Problem 2

An honest but curious doctor: An honest but curious doctor could try to infer information from the anonymized data of other patients by following the allowed protocol in the system.

A malicious nurse: A malicious nurse could try to read other patient's raw data instead of anonymized version of similar patients by deviating from the specified protocol and this would be a privacy threat.

A patient's honest-but-curious mother: An honest but curious mother can try to infer some information about the doctor, nurse or physician by reading the access log of doctor, nurse or physician to her daughter's/son's information.

A man-in-the-middle (MITM) adversary: A man-in-the-middle adversary can gather all information of the patient who is trying to connect to the system remotely since the information will be sent to the device of the patient via network.

Problem 3

(a) The dataset has 11 attributes which are: age, workclass, education, marital-status, occupation, relationship, race, gender, hours-per-week, native-country and income.

(b) The clean dataset have 45222 rows.

(c) Check figure1.png under this directory.

(d) There 2 probabilities that need to be calculated for each race.

FIRST PROBABILITY (>50K):
$\Pr(income > 50K \mid race = X) = \frac{\Pr(income>50K \cap race=X)}{\Pr(race=X)}$

Queries for White race:

nominator: SELECT count(*) FROM ADULT WHERE race="White" and income=">50K";

denominator: SELECT count(*) FROM ADULT WHERE race="White";

Queries for Black race:

nominator: SELECT count(*) FROM ADULT WHERE race="Black" and income=">50K";

denominator: SELECT count(*) FROM ADULT WHERE race="Black";

Queries for Asian-Pac-Islander race:

nominator: SELECT count(*) FROM ADULT WHERE race="Asian-Pac-Islander" and income=">50K";

denominator: SELECT count(*) FROM ADULT WHERE race="Asian-Pac-Islander";

SECOND PROBABILITY (<= 50K):
$\Pr(income <= 50K \mid race = X) = \frac{\Pr(income<=50K \ \& \ race=X)}{\Pr(race=X)}$

Queries for White race:

nominator: SELECT count(*) FROM ADULT WHERE race="White" and income="<=50K";

denominator: SELECT count(*) FROM ADULT WHERE race="White";

Queries for Black race:

nominator: SELECT count(*) FROM ADULT WHERE race="Black" and income="<=50K";

denominator: SELECT count(*) FROM ADULT WHERE race="Black";

Queries for Asian-Pac-Islander race:

nominator: SELECT count(*) FROM ADULT WHERE race="Asian-Pac-Islander" and income="<=50K";

denominator: SELECT count(*) FROM ADULT WHERE race="Asian-Pac-Islander";

(e) Check q3.py under this directory (Read README.md for installation). The results are following:
Conditional probability 1 (>50K) for White race: 0.2623705112716243

Conditional probability 1 (>50K) for Black race: 0.12630085146641437

Conditional probability 1 (>50K) for Asian-Pac-Islander race: 0.2831926323867997

Conditional probability 2 (<=50K) for White race: 0.7376294887283757

Conditional probability 2 (<=50K) for Black race: 0.8736991485335857

Conditional probability 2 (<=50K) for Asian-Pac-Islander race: 0.7168073676132003

(f) Conditional probabilities for this dataset show that for each race, majority of people in that race earn less than 50K a year (74% of whites, 87% of blacks and 72% of Asian-Pac-Islanders). It is not wise to come to any other conclusion by only looking at these conditional probabilities.

Problem 4

(a)

$$TPR = \frac{TP}{P}$$

In our case TP is 850 and P is 865. So, true positive rate (TPR) is:

$$\frac{850}{865} = 0.98$$

(b)

$$TNR = \frac{TN}{N}$$

In our case TN is 85 (since there were 100 attempts by other people and 15 were accepted and 85 was rejected) and N is 135. So, true positive rate (TNR) is:

$$\frac{85}{135} = 0.63$$

(c)

$$PPV = \frac{TP}{TP + FP}$$

In our case TP is 850. FP is 15.So, precision (PPV) is:

$$\frac{850}{850 + 15} = 0.98$$

(d)

$$ACC = \frac{TP + TN}{P + N}$$

In our case TP is 850. TN is 85. P is 865 and N is 135. So, accuracy (ACC) is:

$$\frac{850 + 85}{865 + 135} = 0.935$$

(f)

$$F_1 = \frac{PPV \times TPR}{PPV + TPR}$$

In our case PPV is 0.98 and TPR is 0.98. So, F_1 score is:

$$\frac{0.98 * 0.98}{0.98 + 0.98} = 0.49$$

Problem 5

(a) Check q5.py under this directory for implementation and dict.csv for results.

(b)

Password of Alice: iloveyou2
Password of Bob: gangsta
Password of Charlie: beautiful

(c) No, it would not work as it is because they added salts to the passwords. Now, we need to create a dictionary for each salt.

(d) My strategy was to create a dictionary for each salt value. This attack obviously requires more computation and more storage since we now store three dictionaries instead of one and also we need to traverse three dictionaries to find the passwords instead of one.

(e) Check q5.py for the implementation.

Password of Dave: manutd
Password of Elaine: hello
Password of Faith: 0123456789

Problem 6

(a) GRANT INSERT ON STUDENTS TO Frank;

(b) GRANT ALL ON COURSES TO Jill;

(c) REVOKE INSERT ON STUDENTS FROM Frank

(d) GRANT UPDATE('capacity') ON COURSES TO Matthew;