Data Anonymization for Federated Learning

Diego Klabjan

Northwestern

Acknowledgements

- Fanfei Meng, ECE at Northwestern
  - Former student of mine
- Veena Mendiratta, Adjunct Professor in MLDS
  - Provides tutorials on b/f/p (p=privacy)

Anonymization

- Permanently and completely removing personal identifiers from data
  - Converting personally identifiable information into aggregated data.
- Anonymized data is data that can no longer be associated with an individual in any manner.
- Once data is stripped of personally identifying elements
  - Elements can never be re-associated with the data or the individual.

Utility (clarity, precision) Considerations

Anonymization reduces the original information in the dataset

- Increasing anonymization ➔ decreasing utility of the dataset.
- Trade-off between utility and risk of re-identification.
- Consider utility by attribute:
  - one extreme: a specific attribute is of key interest and no anonymization technique should be applied (data accuracy)
  - other extreme: an attribute is of no use in a given context, and may be dropped without impacting the utility of the data
- Additional risk if the recipient knows details of the anonymization?
  - may help the analyst to better understand the results
  - may increase the risk of re-identification

Categorization of Variables for SDC

- Direct Identifiers.
  Variables that identify a unit, for example, SSN.
- Indirect Identifiers or Key Variables.
  Set of variables that, when considered together, may be used to identify a unit. For example, using gender, age, region, and occupation it may be possible to identify individuals.

Note: Unit can be individual, household or establishment.

Data Anonymization

TECHNIQUES

Techniques for anonymization

- Generalization
  - replace the original value by a semantically consistent but less specific value
- Data Swapping
- Randomization
- Suppression
  - data not released at all
  - cell-level or (more commonly) tuple-level

Techniques for anonymization

Generalization

Replace the original value by a semantically consistent but less specific value

#	*Zip*	*Age*	*Nationality*	*Condition*
1	130**	< 40	American	Heart Disease
2	130**	< 40	American	Heart Disease
3	130**	< 40	Indian	Viral Infection
4	130**	< 40	Chinese	Cancer

Techniques for anonymization

Suppression

- Data not released at all
- Cell-level or (more commonly) tuple-level

#	*Zip*	*Age*	*Nationality*	*Condition*
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Heart Disease
3	130**	< 40	*	Viral Infection
4	130**	< 40	*	Cancer

Data Anonymization

K-ANONYMITY

Samarati, P. and Sweeney, L. "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression." (1998).

K-anonymity

Change the data such that,

- for each tuple in the resulting table,
- there are at least *(k-1)* other tuples with the same value for the quasi-identifier – K-anonymized table

#	Zip	Age	Nationality	Condition
1	130**	< 40	*	Heart Disease
2	130**	< 40	*	Heart Disease
3	130**	< 40	*	Viral Infection
4	130**	< 40	*	Cancer

4-anonymized

How do you publicly release a database without compromising individual privacy?

K-Anonymity:
- attributes are suppressed or generalized until each row is identical with at least k-1 other rows.

K-Anonymity prevents definite database linkages. At worst, the data released narrows down an individual entry to a group of k individuals.

K-anonymity assumes that each record is for a different person.

- If the same person has multiple records (doctor visits), *k*- anonymity will need to be higher than the repeat records.
  - Else, the records may be linkable, as well as re-identifiable.

K-Anonymity Drawbacks

- K-anonymity alone does not provide full privacy.
- Suppose:
  - attacker knows the non-sensitive attributes of individuals

  AND
  - Japanese have very low incidence of heart disease

	Zip	***Age***	***National***
Bob →	13053	31	American
Umeko →	13068	21	Japanese

K-Anonymity Attack

Original Data

	Non-Sensitive Data			*Sensitive Data*
#	***ZIP***	***Age***	***Nationality***	***Condition***
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

4-anonymized Table

Umeko matches here

Bob matches here

	Non-Sensitive Data			Sensitive Data
#	**ZIP**	**Age**	**Nationality**	**Condition**
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	*1485**	> = *40*	*	*Cancer*
6	*1485**	> = *40*	*	*Heart Disease*
7	*1485**	> = *40*	*	*Viral Infection*
8	*1485**	> = *40*	*	*Viral Infection*
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

4-anonymized Table

	Non-Sensitive Data			*Sensitive Data*
#	**ZIP**	**Age**	**Nationality**	**Condition**
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	*1485**	*> = 40*	*	*Cancer*
6	*1485**	*> = 40*	*	*Heart Disease*
7	*1485**	*> = 40*	*	*Viral Infection*
	*1485**	*> = 40*	*	*Viral Infection*
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Umeko matches here

Bob has Cancer

Bob matches here

4-anonymized Table

Umeko has Viral Infection

Umeko matches here

Bob has Cancer

Bob matches here

	Non-Sensitive Data			*Sensitive Data*
#	**ZIP**	**Age**	**Nationality**	**Condition**
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	*1485**	> = *40*	*	*Cancer*
6	*1485**	> = *40*	*	*Heart Disease*
7	*1485**	> = *40*	*	*Viral Infection*
8	*1485**	> = *40*	*	*Viral Infection*
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Algorithm [Kenig et al, 2012]

- Quality of k-anonymous database
  - Maximize $\frac{1}{n}\sum_i \frac{\bar{n}(i)-1}{n(i)-1}$
  - $n(i) =$ the number of distinct values for feature i in raw database
  - $\bar{n}(i) =$ the number of distinct values for feature i in anonymized database
- Cluster samples so that the cardinality of each cluster is within $[k, 2k-1]$
- For each public feature and each cluster set the value of this feature for the samples in the cluster to any existing value
  - All samples in a cluster get the same feature value
- Easy and practical
  - Approximately solves the problem

K-Anonymity Drawbacks

Basic Reasons for leak

- Sensitive attributes lack diversity in values
  - Homogeneity attack
- Attacker has additional background knowledge
  - Background knowledge attack

Hence another solution has been proposed
(in-addition to k-anonymity)
– l-diversity

An equivalence class is said to have l-diversity if there are at least l "well-represented" values for the sensitive attribute.

How to select a value for k?

- It is context dependent.
- Clearly k=1 (no anonymity) and k=n (no utility) are generally useless for a dataset of size n
- In the anonymity vs utility tradeoff, an appropriate privacy level can be bounded from either direction:
  - given a certain analysis that should remain possible with the anonymized data set, what is the maximum *k* we can tolerate?
  - given the privacy guarantees we'd like to make, what is the minimum *k* we require?
- For basic demographic data a level around $5 \leq k \leq 20$ is appropriate.
- A low *k* may be implicitly required if the dataset is small.

Federated Learning with Anonymity

Easy Going

- Each client applies k-anonymity on its local data
  - k does not have to be the same among clients
- Unclear how to apply k-anonymity on embeddings
  - Their values change during algorithm execution

FL and DP

Scheme & Privacy \ Datasets		MNIST Acc.	MNIST Clu.	FASHION Acc.	FASHION Clu.	KDD Acc.	KDD Clu.
Centralized	N/A	0.993	60k	0.895	60k	0.925	300k
FATE	Homo.	0.941	60k	0.795	60k	0.923	300k
FedEmb	Ran.	0.961	15k	0.828	15k	0.923	15k
	KM.	0.950	15k	0.812	15k	0.923	15k
	Ran.	0.954	1.2k	0.825	1.2k	0.924	1.2k
	KM.	0.951	1.2k	0.812	1.2k	0.924	1.2k
	Ran.	0.947	0.6k	0.820	0.6k	0.924	0.6k
	KM.	0.950	0.6k	0.813	0.6k	0.923	0.6k
FedEmb (G.)	Ran. + 0.21%	0.955	15k	0.798	15k	0.923	15k
	Ran. + 0.83%	0.951	15k	0.799	15k	0.923	15k
	Ran. + 3.33%	0.952	15k	0.784	15k	0.922	15k
	Ran. + 0.21%	0.952	1.2k	0.813	1.2k	0.924	1.2k
	Ran. + 0.83%	0.954	1.2k	0.793	1.2k	0.924	1.2k
	Ran. + 3.33%	0.955	1.2k	0.804	1.2k	0.923	1.2k
FedEmb (P.)	Ran. + (0, 0.25)	0.952	15k	0.804	15k	0.923	15k
	Ran. + (0, 0.50)	0.888	15k	0.803	15k	0.924	15k
	Ran. + (0, 1.00)	0.952	15k	0.815	15k	0.922	15k

- P = differential privacy
- Listed (μ, β)
- Not big decrease due to DP