lecture 12: Socio-cultural impact

wbg231

January 2023

1 Socio-cultural impact of recommender systems

- recommender systems killed buz feed :(
- the rest of this was covered in the similarly lecture

privacy and de-anonymization

open data and anonymization

- scientific process is built on open data but human data can be sensitive
- anonymization of data is not enough
- nor is just providing statistical summaries of the data

strategies for protecting users in open data

- could hash user names to a key
- could add noise to observations (but can be undone in some cases and can bias results)
- limit users to some number of queries
- just provide statistical summaries (this is not good either)

what is a de-anonymization attack

- suppose we have "anonymized" dataset $R = (r_1 ... r_n)$
- given some partial or even potentially inaccurate information for an individual, can we identify them or get more information about them?
- yeah in most cases

why k-anonymity is not enough

- $k-anonymity$ is the idea that we only include a row in a data set if each attribute in that row has the same value in at least k other rows. so no single attribute is identifiable
- but combinations of attributes are identifiable in most cases in large dimensional collections
- people are high dimensional and idiosyncratic which is often reflected in there data
- for rows R_u, R_v define there similarly as $sim(R_u, R_v) = \frac{\sum_i Sim(R_{u,i}, R_{v,u})}{|R_u \cup R_v|}$ that is, the sum of there similarly in each attribute over the cardinality of there union
- given a partial observation q compute the similarly in each row
- determine a threshold by comparing top scores in the second most similar row
- if it is sufficiently large (ie it si much closer to one row than any other) report a match
- otherwise report no match.
- how much partial data is required for this?
- observation similarly if $|R_{u,i} - R_{v,i}| = 0$ the to rows are exactly the same
- if $|R_{u,i} - R_{v,i}| \leq 1$ they differ by at most 1 unit
- we can define a threshold naturally as confidence interval $sim(q, R_1) - sim(q, R_1) > 1.5 * \sigma_w(sum(w, R_w))$ so that is (we define our threshold as 1.5 times the variance of the similirty between q and the rows)
- with 8 ratings (if we perturb two ratings) and add 14 days of error on the data data 99% of records can be identifiable

why does this matter

- breaches are irrevocable and may have implications on peoples future privacy
- once data is out there it can be used in linkage attacks

Differential privacy

what are we putting out to the world

- a whole dataset
- a set of statistic measured on teh data
- a statistical model from the data
- an api to ask queries about the data?

Differential privacy

- the high level idea is that if an individual is excluded from the data the results of a computation should not change
- we achieve this by randomizing teh computation carefully
- DP is a property of the algorithm not the data
- for any two datasets D, D' diverging by one row ie $D' = D + \{x\}$ a randomized algorithm is ϵ differentially private if

$$P(A(D) \in S) \leq e^{\epsilon} P(A(D') \in S)$$

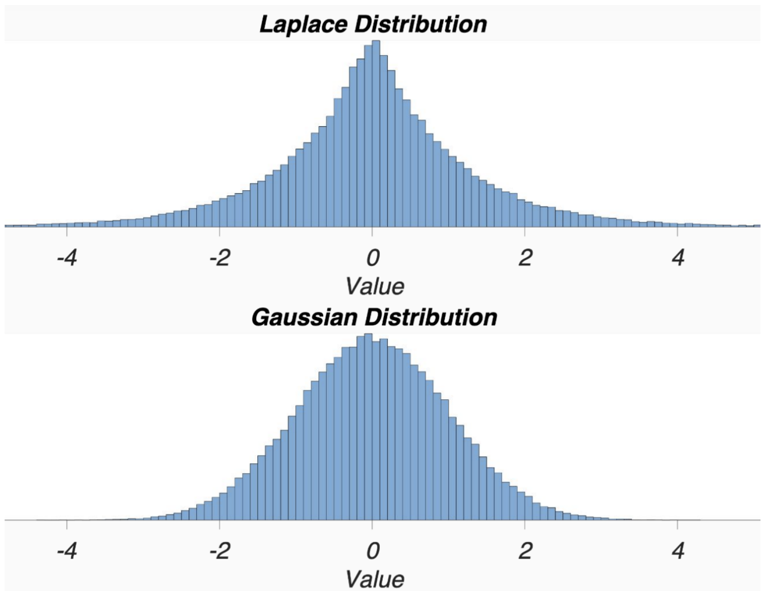
$\forall S \in range(A)$

- the idea is if we observe a certain output we should not be relibaly able to tell if it came from A(D) or A(D')
- DP says $P(A(D) \in S) \leq e^{\epsilon} P(A(D') \in S)$
- for $\epsilon \approx 0$ we will have $P(A(D) \in S) \leq P(A(D') \in S)$
- DP is symetric
- when ϵ is large the bound is louser

tuning the noise

- say we have a vector value private fucntion $f : D \rightarrow R^d$
- how different are $f(d)$ and $f(D')$ if the two data set differ only by a single row?
- call the sensitivity of f $\Delta f = max_{D,D'} \sum_i [f(D)[i] - F(D')[i]]$
- let $A(D) = f(D) + z$ where $z[i] \sim Laplace(0, \frac{\Delta f}{\epsilon})$ where if this is the case

- in this case A is ϵ differentially private

- 

- that is what a Laplacian looks like versus a guassian

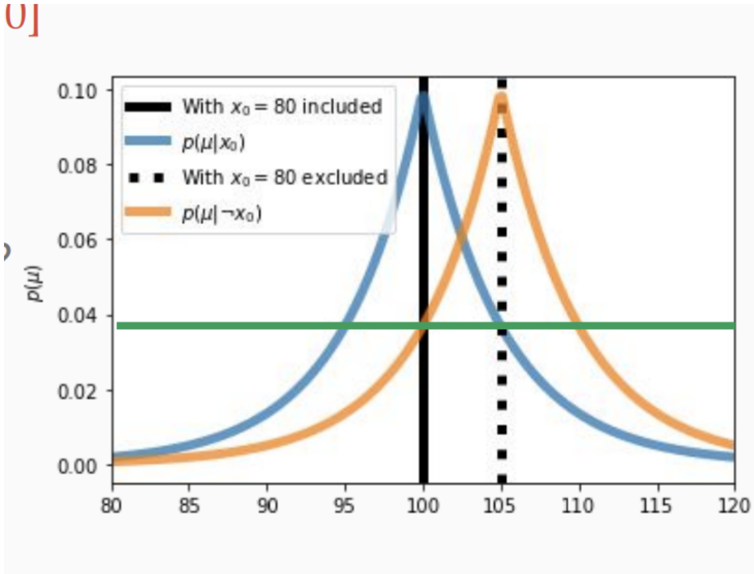
- so the key is that the tails are wider and it is zero mean and symetric about the origin

why is this a good idea

- say we have a dataset $X = [8, 9, 10, 11, 12]$ if we compute the mean of the data then remove 8 and recompute the mean of data

- the mean of the two data sets have shifted but but if we find that $A(D) = 105$ we can see that with either of our datasets there is not a low probability of this outcome

- so there is no real evedince that 8 was no included in the data set just becuase ouf our result

- 

- a lot of noise means high privacy
- note that external aggregators like min, max are very sensative (so are hard to make differentially private) so it is better to report percentiles
- sensitivity goes down as n decreaes
- privacy is easier with big data sets

what about multiple queries

- Differential privacy is at the query level so each time you query you will get a random results
- Differential privacy composition theorem: if you make a sequence of queries A_i each being ϵ_i-DP then teh results will be $\sum_i \epsilon_i$ DP
- the good news any deterministic post procesing preserves privacy
- DP is a property of an algorithm not a dataset
- privacy requires scale
-