NYU

DS-GA 3001.009 Responsible Data Science Lab 7

Center for Data Science
Haoyue Ping| Tandon School of Engineering

NYU

Outline

Introduction

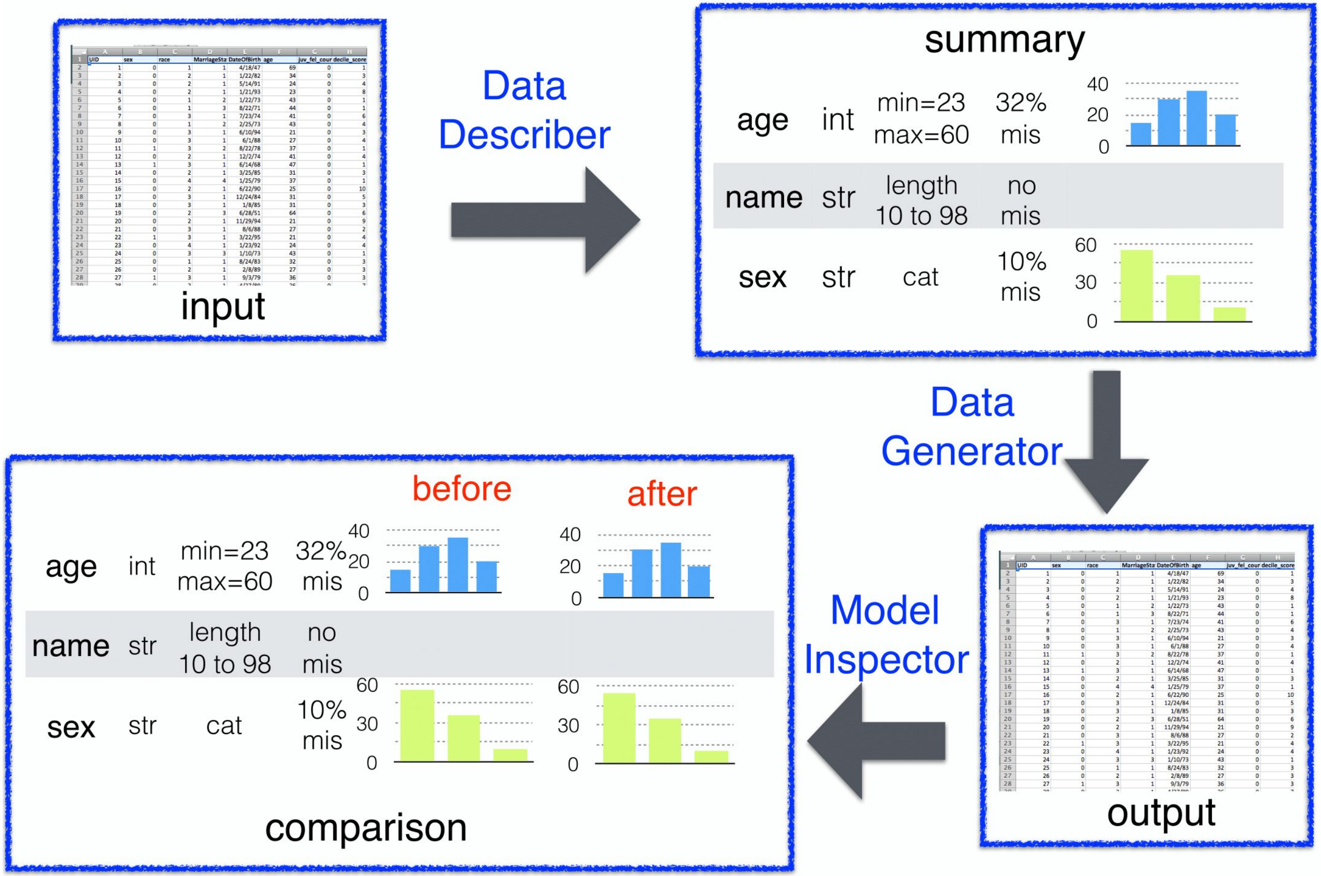
Download DataSynthesizer and setup the running environment

DataSynthesizer usage

- **Random mode**
- **Independent attribute mode**
- **Correlated attribute mode**

Some useful statistical measures

Introduction



DataSynthesizer installation

GitHub repo

https://github.com/DataResponsibly/DataSynthesizer

- **Download it**
- **Add ./DataSynthesizer/ into sys.path**

Random mode

- **Generate type-consistent data**
- **Learn the domains of attributes**
    - Data type
    - Categorical vs non-categorical
        - Threshold = 20 by default
        - True for rating, gender
        - False for score, name
    - Numerical vs non-numerical
        - Integer, Float, Datetime are numerical
        - Datetimes → timestamps if non-categorical
    - Active domain
        - if is_categorical:
            - Attribute values in dataset
        - else if is_numerical:
            - Range(min, max)

Data Type	Example
Integer	ID, age
Float	Score, rating
String	Name, gender
Datetime	Birthday, event time

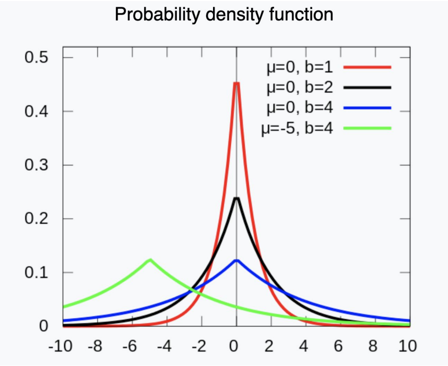
Independent attribute mode

Assume the attributes (or columns) are independent.

- **Run random mode first to get the attribute domains**

- **Model attribute distributions**
  - Bar charts for categorical attributes
  - Histograms for numerical attributes

- **Inject Laplace noise into the bar charts / histograms.**
  - Sensitivity = 2/n
  - d = #attributes, then privacy budget is ε/d for each attribute.
  - Inject Lap(2d/nε)



Correlated attribute mode

Parameters

- epsilon: the privacy budget
- k: #parents in Bayesian network (BN)

Run GreedyBayes to construct a BN

- Connect attributes with high mutual information
- Randomize the attribute connections
- Cost epsilon/2, half of the privacy budge

Populate conditional probability tables (CPTs)

- Inject Laplace noise into CPTs
- Cost epsilon/2, half of the privacy budge

Randomize BN structure

Algorithm 1 GreedyBayes(D, A, k)

Require: Dataset D, set of attributes A, maximum number of parents k

```
1:  Initialize N = ∅ and V = ∅.
2:  Randomly select an attribute X_1 from A.
3:  Add (X_1, ∅) to N; add X_1 to V.
4:  for i = 2, .., |A| do
5:      Initialize Ω = ∅
6:      p = min(k, |V|)
7:      for each X ∈ A\V and each Π ∈ (V choose p) do
8:          Add (X, Π) to Ω
9:      end for
10:     Compute mutual information based on D for all pairs in Ω.
11:     Select (X_i, Π_i) from Ω with maximal mutual information.
12:     Add (X_i, Π_i) to N.
13: end for
14: return N
```

Select the (child, parents) among all combinations in Ω **with a probability proportional to** $\exp(I(X,\Pi)/2\Delta)$

Where $I()$ is mutual information.

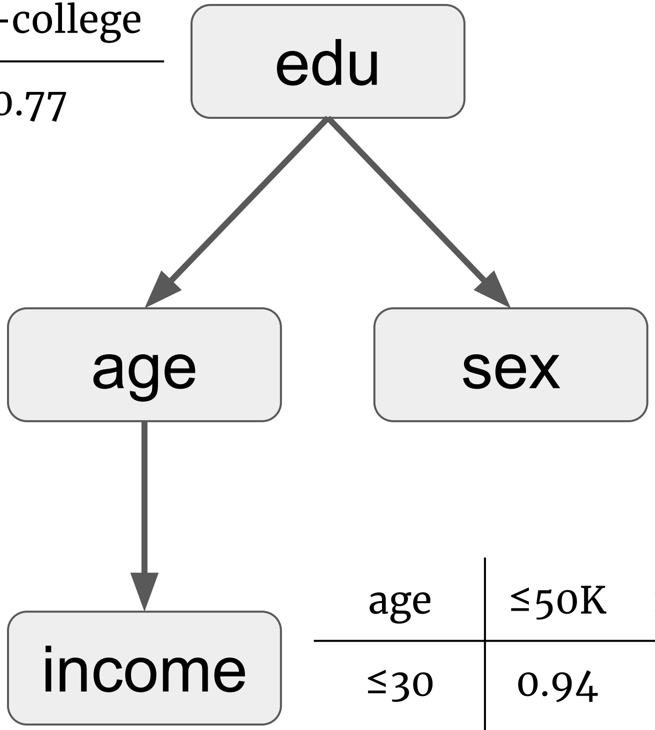
$$\Delta = (d-1)S(I)/\varepsilon$$

$$S(I(X,\Pi)) = \begin{cases} \frac{1}{n}\log(n) + \frac{n-1}{n}\log\left(\frac{n}{n-1}\right), & \textit{if } X \textit{ or } \Pi \textit{ is binary;} \\ \frac{2}{n}\log\left(\frac{n+1}{2}\right) + \frac{n-1}{n}\log\left(\frac{n+1}{n-1}\right), & \textit{otherwise,} \end{cases}$$

n is the number of tuples in D.

Randomize BN structure

college	non-college
0.23	0.77



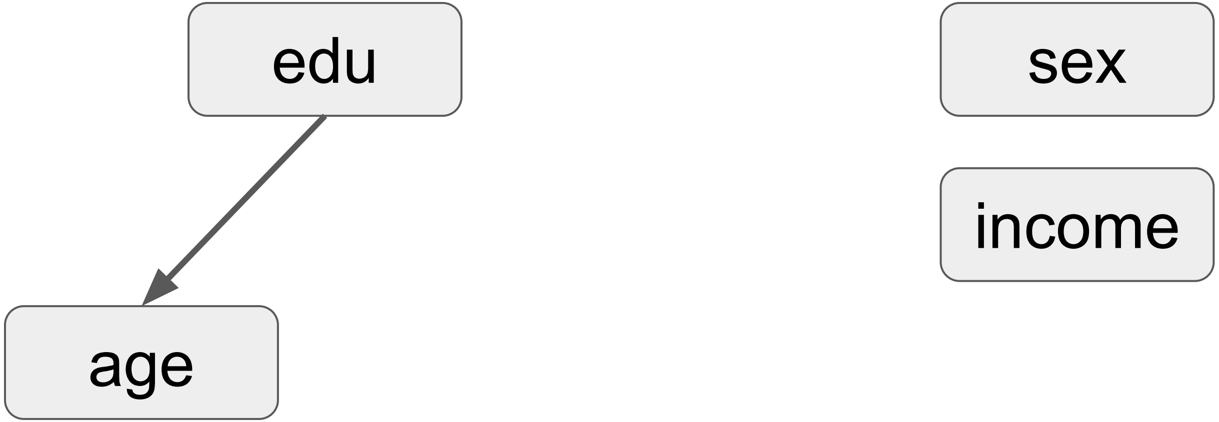
edu	≤30	30~50	>50
college	0.24	0.56	0.20
non-college	0.35	0.45	0.20

edu	female	male
college	0.30	0.70
non-college	0.34	0.66

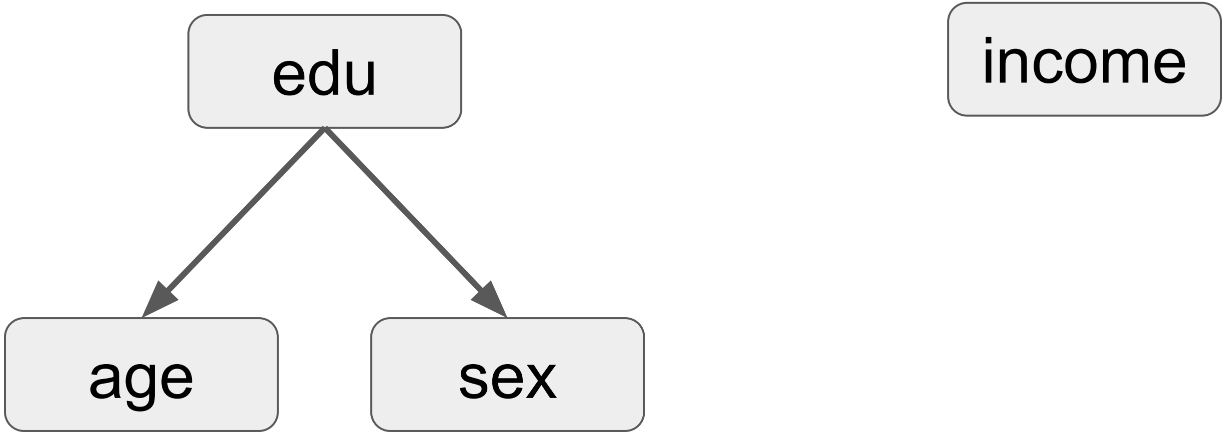
age	≤50K	>50K
≤30	0.94	0.07
30~50	0.67	0.33
>50	0.68	0.32

Step 0: add root

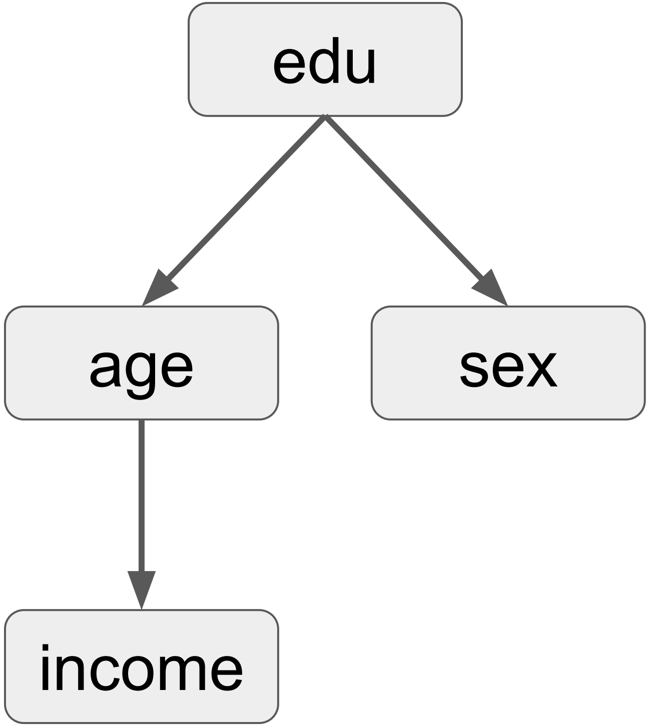
Step 1: add the 1st child



Step 2: add the 2nd child

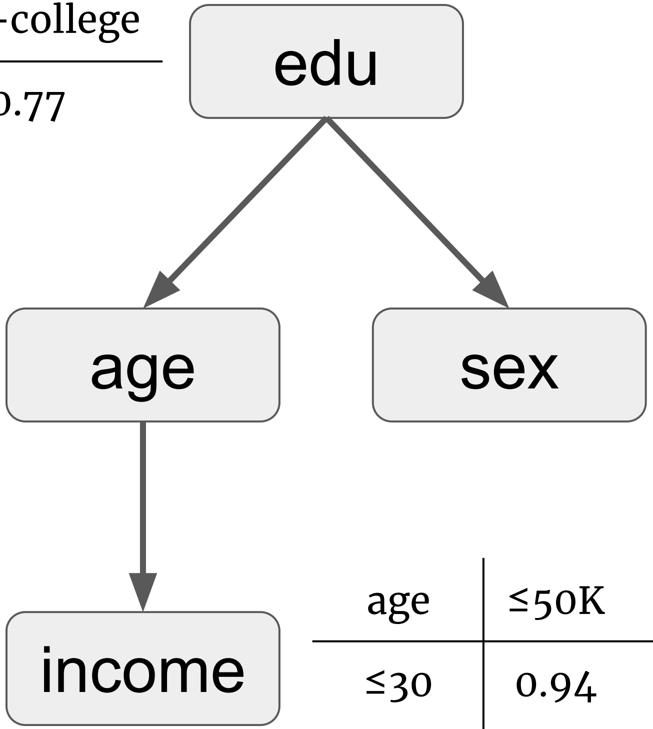


Step 3: add the 3rd child



CPTs with Lap(4(d-k)/(n·ε))* noise

college	non-college
0.23	0.77



edu	≤30	30~50	>50
college	0.24	0.56	0.20
non-college	0.35	0.45	0.20

edu	female	male
college	0.30	0.70
non-college	0.34	0.66

age	≤50K	>50K
≤30	0.94	0.07
30~50	0.67	0.33
>50	0.68	0.32

*Lap(4(d-k)/(n·ε))
- d=#attributes in BN
- k=#parents in BN
- n=sensitive dataset size

Statistical measures

Mutual information

- **How much information can be obtained from one random variable about another random variable?**

Two-sample Kolmogorov–Smirnov test

- **How different are two continuous distributions?**

KL-divergence

- **How different are two categorical distribution?**

Mutual information*

- **The "amount of information" obtained from one random variable about another random variable.**

$$\mathrm{I}(X;Y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

- **MI(X, Y) = 0 if random variables X and Y are independent**

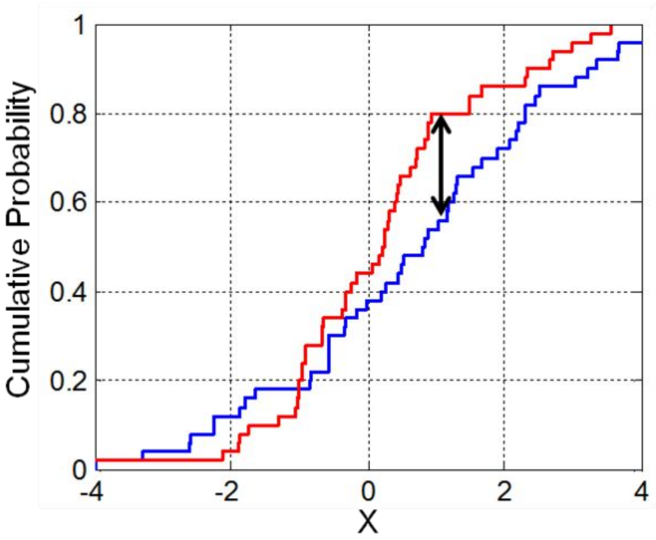
$$\log\left(\frac{p(x,y)}{p(x)\,p(y)}\right) = \log 1 = 0$$

*From https://en.wikipedia.org/wiki/Mutual_information

Two-sample Kolmogorov–Smirnov test*

- **Test whether two underlying one-dimensional probability distributions differ.**

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$$



*From https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test#Two-sample_Kolmogorov%E2%80%93Smirnov_test

KL-divergence*

- **How different are two categorical distribution P and Q?**

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

- **$D_{KL}(P||Q) = 0$ if P and Q are identical.**

- **The KL-divergence is defined only if for all x, Q(x)=0 implies P(x)=0**

*From https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence

Thank you!