Data Preprocessing

Andrew McCarren

Data Preprocessing

- Why preprocess the data?
- Steps in Preprocessing
  - Data Collection/Integration
  - Data cleaning and handling Missing Data
  - Data Augmentation
  - Feature engineering
-

Data Cleaning

- Data cleaning tasks
  - Fill in missing values
  - Identify outliers
  - noisy data
  - Transformations (Feature Engineering)

Structured vs. Unstructured Data

- Structured data
  - Loadable into “spreadsheets”
  - Arranged into rows and columns
  - Each cell filled or could be filled
  - Data mining friendly

- Unstructured daa
  - Microsoft Word, HTML, PDF documents, PPTs
  - Usually converted into XML → semi structured
  - Not structured into cells
  - Variable record length, notes, free form survey-answers
  - Text is relatively sparse, inconsistent and not uniform
  - Also images, video, music etc.

Missing Data

- Missing data – values, attributes, entire records, entire sections

- Missing values and defaults are indistinguishable
- Truncation/censoring – not aware, mechanisms not known
- Problem: Misleading results, bias.

Detecting Missing Data

- Overtly missing data
  - Match data specifications against data – are all the attributes present?
  - Scan individual records – are there gaps?
  - Rough checks : number of files, file sizes, number of records, number of duplicates
  - Compare estimates (averages, frequencies, medians) with “expected” values and bounds; check at various levels of granularity since aggregates can be misleading.

Missing data detection (cont.)

- Hidden damage to data
  - Values are truncated or censored – check for spikes and dips in distributions and histograms
  - Missing values and defaults are indistinguishable – too many missing values? metadata or domain expertise can help
  - Errors of omission e.g. all calls from a particular area are missing – check if data are missing randomly or are localized in some way

Missing Data Mechanisms

Missing Completely at Random (MCAR)

Missing value (y) neither depends on x nor y

if the events that lead to any particular data-item being missing are independent both of observable variables and of unobservable parameters of interest, and occur entirely at random.

Missing at Random (MAR)

Missing value (y) depends on x, but not y

Example: Respondents in service occupations less likely to report Income

Missing not at Random (NMAR)

The probability of a missing value depends on the variable that is missing Example:

Respondents with high income less likely to report income

Imputing Values to Missing Data

- In federated data, between 30%–70% of the data points will have at least one missing attribute – data wastage if we ignore all records with a missing value
- Remaining data is seriously biased
- Lack of confidence in results
- Understanding pattern of missing data unearths data integrity issues

Exploring missing data mechanisms

- Can‟t be 100% sure about probability of missing (since we don‟t actually know the missing values)

- Could test for MCAR
- Many missing data methods assume MCAR or MAR

But our data is often NMAR

- Some methods specifically for NMAR

Questions

- The salaries of numerous professions have been collected.
- The table outlining the number of missing items in each category is given below.
- How would you go about testing if there is a relationship between missing values and the Profession. Is this MCAR or MAR?

Number	Accountant	IT Professional	Lecturer
Missing	60	40	20
Not Missing	200	150	100

Missing Value Imputation – 1

- Standalone imputation
  - Mean, median, other point estimates
  - Assume: Distribution of the missing values is the same as the non-missing values.
  - Does not take into account inter-relationships
  - Introduces bias
  - Convenient, easy to implement

Missing Value Imputation – 2

- Better imputation – use attribute relationships
- Assume : all prior attributes are populated
  - That is, *monotonicity* in missing values.

X1	X2	X3	X4	X5
1.0	20	3.5	4	.
1.1	18	4.0	2	.
1.9	22	2.2	.	.
0.9	15	.	.	.

- Two techniques
  - Regression (parametric),
  - Propensity score (nonparametric)

Missing Value Imputation -3

- Regression method
  - Use linear regression, sweep left-to-right

    $X3=a+b*X2+c*X1;$

    $X4=d+e*X3+f*X2+g*X1,$ and so on
  - X3 in the second equation is estimated from the first equation if it is missing

Missing Value Imputation – 3

- Propensity Scores (nonparametric)
  - Let $Y_j=1$ if X_j is missing, 0 otherwise
  - Estimate $P(Y_j=1)$ based on X_1 through $X_{(j-1)}$ using logistic regression
  - Create propensity score $P(Y_j=1)$ groups. Then align with non-missing X's who have a similar propensity scores
  - Within each group, estimate missing X_js from known X_j's using approximate Bayesian bootstrap.
  - Repeat until all attributes are populated.

Missing Value Imputation

- Arbitrary missing pattern
  - Markov Chain Monte Carlo (MCMC)
  - Assume data is multivariate Normal, with parameter Θ
  - (1) Simulate missing X, given Θ estimated from observed X ; (2) Re-compute Θ using filled in X
  - Repeat until stable.
  - Expensive

- Note that imputed values are useful in aggregates but can't be trusted individually

What is a propensity score (nice Papers)

- https://www.youtube.com/watch?v=ACVyPp1Fy6Y
- https://www.futuremedicine.com/doi/full/10.2217/cer-2017-0071

Does and don’ts

- Don’t impute output variables
- Individual points that have been imputed should only be relied up as aggregations.
- Don’t impute NMAR variables.

R Packages for Missing Data

- VIM (Uses regression technique)

- The Amelia Package. (Various options
- mvnmle package (to create a complete variance/covariance matrix. Based on ML of imputed missing values)

- The SeqKnn and rrcovNA (Uses k-nearest neighbours (Assumes data is missing at Random)

Question?

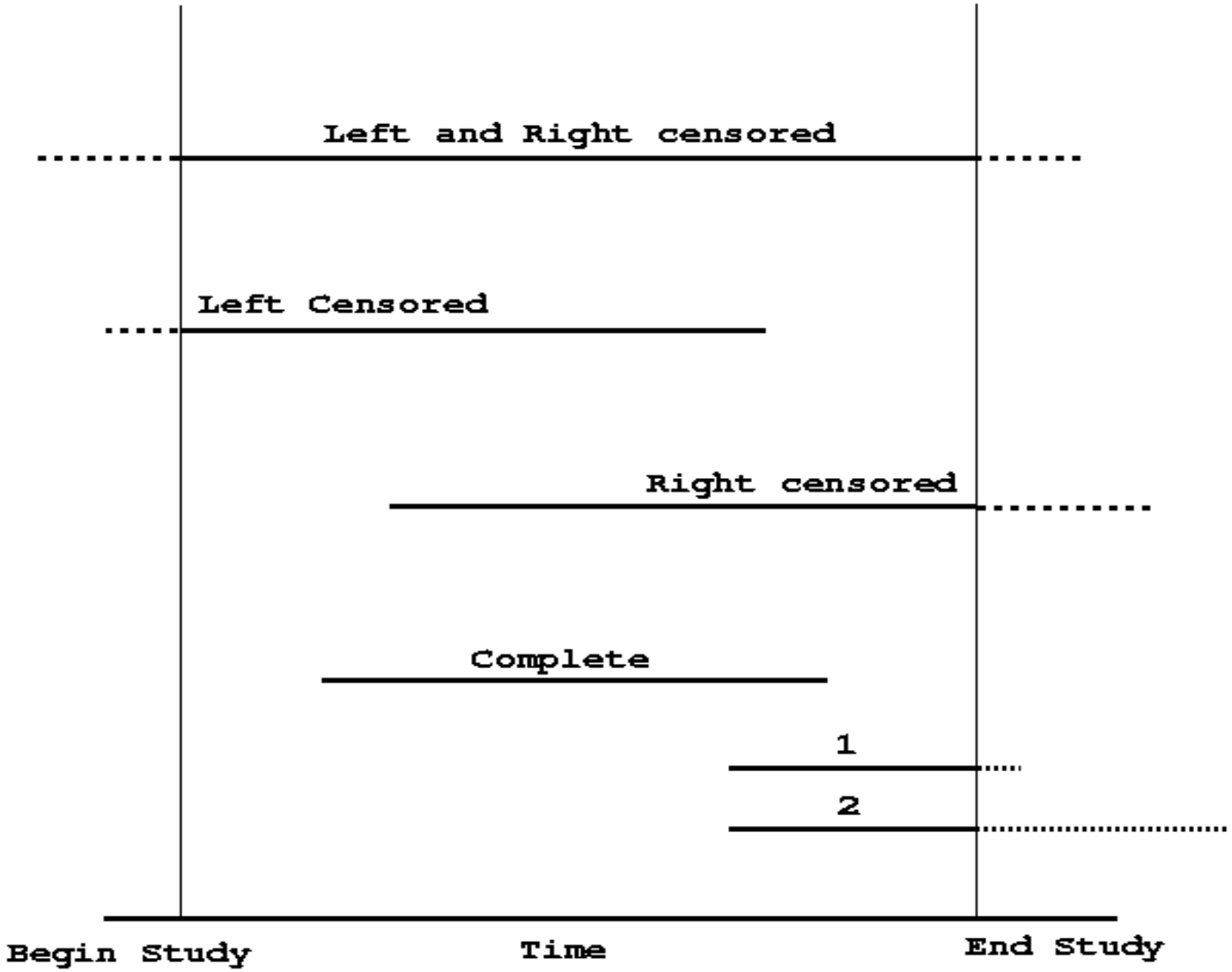
- Which Missing data mechanism should not be applied with missing values generation packages?
  - 1. MCAR
  - 2. MAR
  - 3. NMAR

Censoring and Truncation

- Well studied in Biostatistics, relevant to time dependent data e.g. duration

- *Censored* – Measurement is bounded but not precise e.g. Call duration > 20 are recorded as 20

- *Truncated* – Data point dropped if it exceeds or falls below a certain bound e.g. customers with less than 2 minutes of calling per month



Censored time intervals

Censoring/Truncation (cont.)

- If censoring/truncation mechanism not known, analysis can be inaccurate and biased.
- But if you know the mechanism, you can mitigate the bias from the analysis.
- Metadata should record the existence as well as the nature of censoring/truncation

Question?

- If the censoring mechanism is known then we can mitigate the bias from our analysis?
  - 1. TRUE
  - 2. FALSE