Missing Data Imputation

Alexander Hanf, Arber Qoku

11.12.2019

Agenda

Statistical background

Imputation Methods

Tools and libraries for dealing with missing data

Live Demo

Statistical background

Missing data

- Very common in practical problems
- Data can be missing for many reasons
  - Survey participant is unreachable or refuses to answer
  - Question does not apply to patient (sex-specific)
  - Subset of sensors not working for a period of time
- Complicates statistical analysis (bias, statistical power, ...)
- Many machine learning models can not handle missing data.

Types of missing data

- Missing completely at random (MCAR): the pattern of missing values is independent of all other covariates (both observed and unobserved)
- Missing at random (MAR): the pattern of missing values depends only on observed covariates
- Missing not at random (MNAR): the pattern of missing values also depends on unobserved covariates

Idea of Imputation

- Use statistical techniques to fill in the missing values
- Make the most of the data we have
- Caution is required! We can bias our analysis or create nonsensical data

Imputation Methods

Complete case analysis

- Row-wise or column-wise deletion

+ Very simple
+ Valid inferences when MCAR

– Biased results on MAR or MNAR
– Too few datapoints (if any) left!
  If only 5% missing independently on a dataset with 20 columns, we end up with $0.95^{20} \approx 35\%$ of the original data

Big problem in predictive modelling!

Imputation of constant values

- Substitute mean, median, mode, constant value

+ Simple
+ Retains all observed values
+ Can preserve statistical quantities

– Underestimates standard errors
– May impute unrealistic values

Hot-Deck Imputation

- General idea: reuse values from other observations
- Assumption: identically distributed data points
- "Donor" observations can be chosen in multiple ways
    - Naively: imputing value from randomly chosen observation
    - Distance-based methods
    - Predictive mean matching (Regression & Distance)

$$X = \underbrace{\begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ - & x_{32} & x_{33} \end{pmatrix}}_{\text{original design matrix}} \rightarrow \underbrace{\begin{pmatrix} x_{12} & x_{13} \\ x_{22} & x_{23} \end{pmatrix}}_{\text{design matrix}}, \underbrace{\begin{pmatrix} x_{11} \\ x_{21} \end{pmatrix}}_{\text{response}} \rightarrow \underbrace{\begin{pmatrix} \hat{x}_{11} & x_{12} & x_{13} \\ \hat{x}_{21} & x_{22} & x_{23} \\ \hat{x}_{31} & x_{32} & x_{33} \end{pmatrix}}_{\text{compute distances}}$$

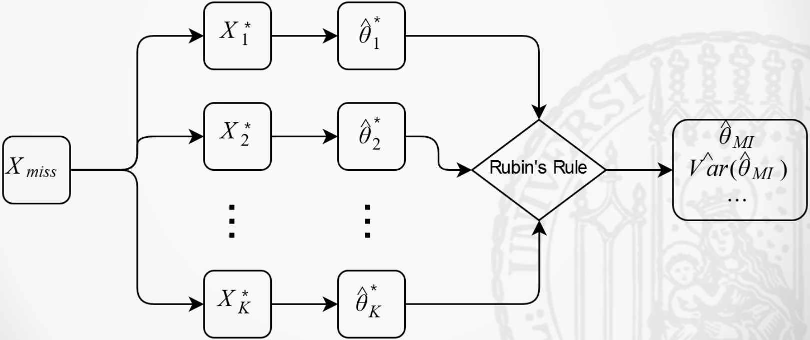
$+$ Easy to implement and computationally efficient

$-$ Donor values might get reused often

$-$ Underestimates standard errors

Multiple Imputation

- Use given data to quantify uncertainty of imputations



Specification Overhead

- Imputation model: linear regression, random forest, ...
- Predictor vs response variables: leave-one-out, interactions, auxiliary, ...
- Order of variable imputation: random, least/most missing first, ...
- Initial imputations, number of iterations (convergence condition)
- Number of multiply imputed datasets (cycles)

Specification Overhead

- Imputation model: linear regression, random forest, ...
- Predictor vs response variables: leave-one-out, interactions, auxiliary, ...
- Order of variable imputation: random, least/most missing first, ...
- Initial imputations, number of iterations (convergence condition)
- Number of multiply imputed datasets (cycles)

$\Rightarrow$ Multiple Imputation by Chained Equations (MICE)

Multiple Imputation by Chained Equations (MICE)

$$X = \underbrace{\begin{pmatrix} x_{11} & - & x_{13} \\ x_{21} & x_{22} & - \\ - & x_{32} & - \\ x_{41} & x_{42} & x_{43} \end{pmatrix}}_{\text{original design matrix}} \rightarrow \underbrace{\begin{pmatrix} x_{11} & \bar{x}_{.2} & x_{13} \\ x_{21} & x_{22} & \bar{x}_{.3} \\ \bar{x}_{.1} & x_{32} & \bar{x}_{.3} \\ x_{41} & x_{42} & x_{43} \end{pmatrix}}_{\text{initial imputation}} \rightarrow \underbrace{\begin{pmatrix} \bar{x}_{.2} & x_{13} \\ x_{22} & \bar{x}_{.3} \\ x_{42} & x_{43} \end{pmatrix}}_{\text{design matrix}}, \underbrace{\begin{pmatrix} x_{11} \\ x_{21} \\ x_{41} \end{pmatrix}}_{\text{response}}$$

$$\rightarrow \begin{pmatrix} x_{11} & \bar{x}_{.2} & x_{13} \\ x_{21} & x_{22} & \bar{x}_{.3} \\ \hat{x}_{31} & x_{32} & \bar{x}_{.3} \\ x_{41} & x_{42} & x_{43} \end{pmatrix} \rightarrow \begin{pmatrix} x_{21} & \bar{x}_{.3} \\ \hat{x}_{31} & \bar{x}_{.3} \\ x_{41} & x_{43} \end{pmatrix}, \begin{pmatrix} x_{22} \\ x_{32} \\ x_{42} \end{pmatrix} \rightarrow \cdots \underbrace{\begin{pmatrix} x_{11} & \hat{x}_{12} & x_{13} \\ x_{21} & x_{22} & \hat{x}_{23} \\ \hat{x}_{31} & x_{32} & \hat{x}_{33} \\ x_{41} & x_{42} & x_{43} \end{pmatrix}}_{X_i^*}$$

Deep Learning for Imputation

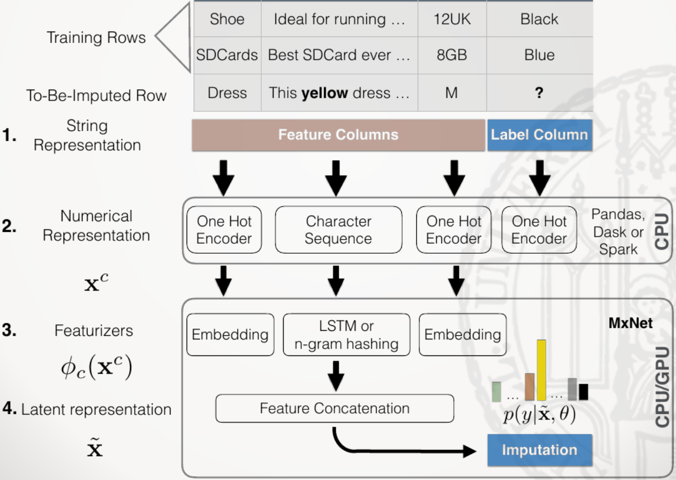
- Recent advances in deep learning improved model performance on a wide variety of NLP tasks
- This can be used to impute textual features or other categorical features
- Example: searching for ”yellow dress” on Amazon - textual description might exist, but ”colour” is missing in database

Amazon DataWig

scalable, high-precision, language-agnostic, end-to-end pipeline

Imputation of attribute `color`

Product Type	Description	Size	Color
Shoe	Ideal for running …	12UK	Black
SDCards	Best SDCard ever …	8GB	Blue
Dress	This **yellow** dress …	M	?



Tools and libraries for dealing with missing data

Useful libraries

Exploratory data analysis

- General purpose: pandas-profiling
- Specific to missing data: missingno

Imputation

- General purpose ($\sim$ `mice` in R): scikit-learn
- Hot-deck with KNNs: fancyimpute
- Random Forest ($\sim$ `missForest` in R): missingpy
- Imputation of Time Series Data (WIP): impyute

Live Demo

References

- Biessmann, Felix, et al. ”Deep Learning for Missing Value Imputation in Tables with Non-Numerical Data.” Proceedings of the 27th ACM International Conference on Information and Knowledge Management. ACM, 2018.
- Buuren, S. van, and Karin Groothuis-Oudshoorn. ”mice: Multivariate imputation by chained equations in R.” Journal of statistical software (2010): 1-68.
- Wulff, Jesper N., and Linda Ejlskov. ”Multiple Imputation by Chained Equations in Praxis: Guidelines and Review.” Electronic Journal of Business Research Methods 15.1 (2017).
- Joenssen, Dieter William Hermann. Hot-Deck-Verfahren zur Imputation fehlender Daten: Auswirkungen des Donor-Limits. Diss. 2015.