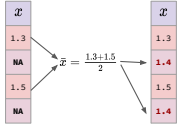
Applied Machine Learning

Imputation:
Introduction and Simple Methods

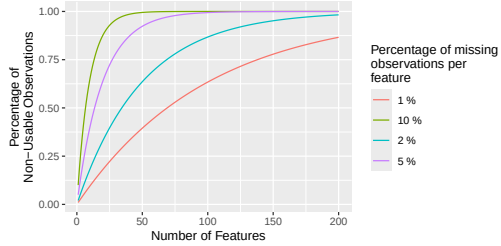


Learning goals

- Understanding missing data mechanisms
- Simple imputation strategies

MOTIVATING EXAMPLE

- Assume each feature in your dataset has 2% missing values.
- The missing values are randomly distributed over the observations.
- How many rows can be used if all observations that contain at least a missing value is dropped?



With 100 features and 2% missing values only 13% of our data can be used.

VISUALIZING MISSING VALUES

WHY IS DATA MISSING?

▸ Morvan and Varoquaux 2024

- faulty measurements
- unanswered questionnaire items
- unreported data

MISSINGNESS MECHANISMS

- MCAR (Missing Completely At Random)
  - Data is missing independently of missing or observed values
- MAR (Missing At Random)
  - Missingness depends only on observed values
- MNAR (Missing Not At Random)
  - Missingness depends on observed and missing values

MISSINGNESS MECHANISMS

- MCAR (Missing Completely At Random)
  - Data is missing independently of missing or observed values
- MAR (Missing At Random)
  - Missingness depends only on observed values
- MNAR (Missing Not At Random)
  - Missingness depends on observed and missing values

Which one is most realistic to assume in the real world?

MISSINGNESS MECHANISMS - EXAMPLES

- MCAR (Missing Completely At Random)
  - Participants skip a page of a questionaire because they flip two pages at once
- MAR (Missing At Random)
  - Certain medical examinations are conducted more often for certain patient groups (age groups, different gender)
  - Certain demographic groups are less likely to fill in surveys about depression (but demographic groups is often fully observable, e.g., gender)
- MNAR (Missing Not At Random)
  - Missingness depends on observed and missing values

POSSIBLE WAYS TO DEAL WITH MISSING VALUES

- Remove observations that contain missing values.
  But: Could lead to a very small dataset.
- Remove features that contain mostly missing values.
  But: Can lose (important) information.
- Use models that can handle missing values, e.g., (most) tree-based methods
  But: Restriction in model choice.
- **Imputation**
  $\rightarrow$ Replace missing values with *plausible* values.

REASONS TO IMPUTE MISSING VALUES

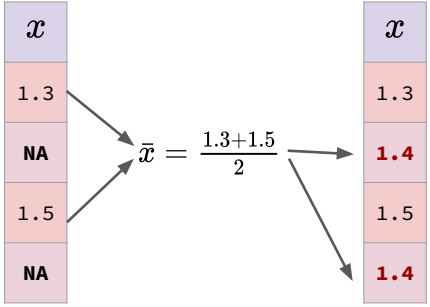
- Prediction
  - Goal: allow for maximal predictive performance
- Inference
  - Goal: estimate parameters such as mean and variance of the variable distribution
  - Not part of this lecture

$\rightarrow$

- Imputation depends on the goal
- Imputation needs to be benchmarked wrt the downstream task

SIMPLE IMPUTATION METHODS

A very simple imputation strategy is to replace missing values with univariate statistics, e.g. mean or median, of the feature:



SIMPLE IMPUTATION METHODS

The statistic used to impute the missing values has to match the type of the feature:

- Numeric features: mean, median, quantiles, mode, ...
- Categorical features: mode, ...

Alternatively missing values can be encoded with new values

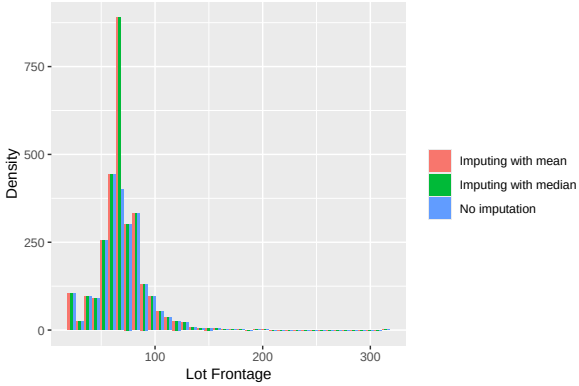
- Numeric features: `2*max`, ...
- Categorical features: `__MISS__`, ...

IMPUTATION NOTES

- To ensure that the information regarding which values were imputed is not lost, we can add a binary indicator variable.
  - Provides additional information in MNAR
- Domain knowledge is highly important: Missing Credit can mean that the individual has 0 debt.
- Encoding numeric values with out-of-range values has been shown to work well in practice for complex ML models.
  - This is especially useful for tree-based methods, as it allows separating observations with missing values in a feature.
  - But using out-of-range imputation when estimating global effects (e.g. in linear models) can skew the results

DISADVANTAGE OF CONSTANT IMPUTATION

By imputing a feature with one value we shift the distribution of that feature towards a single value.



IMPUTATION BY SAMPLING

A way out of this problem is to sample values to replace each missing observation from

- the empirical distribution or histogram, for a numeric feature.
- the relative frequencies of levels, for a categorical feature.

This ensures that the distribution of the features does not change much.

BENCHMARK OF SIMPLE IMPUTATION

To illustrate the effect of imputation on the performance we evaluate a linear model on the Ames housing dataset. Evaluation is done with a 10-fold cross-validation: