Handling Missing Values

Tuesday 14 May 2024 12:51 AM

https://stefvanbuuren.name/fimd/ch-introduction.html

→ Why are values missing? For each column

Types of missing values
- MCAR (Missing completely at Random)
- MAR (Missing at Random):
- MNAR (Missing not at Random):

MCAR (Missing completely at Random)

If the probability of being missing is the same for all cases, then the data are said to be missing completely at random (MCAR). This effectively implies that causes of the missing data are unrelated to the data. We may consequently ignore many of the complexities that arise because data are missing apart from the obvious loss of information. An example of MCAR is a weighing scale that ran out of batteries. Some of the data will be missing simply because of bad luck. Another example is when we take a random sample of a population, where each member has the same chance of being included in the sample. The (unobserved) data of members in the population that were not included in the sample are MCAR. While convenient, MCAR is often unrealistic for the data at hand.

MAR (Missing at Random):

If the probability of being missing is the same only within groups defined by the observed data, then the data are missing at random (MAR). MAR is a much broader class than MCAR. For example, when placed on a soft surface, a weighing scale may produce more missing values than when placed on a hard surface. Such data are thus not MCAR. If, however, we know surface type and if we can assume MCAR *within* the type of surface, then the data are MAR. Another example of MAR is when we take a sample from a population, where the probability to be included depends on some known property. MAR is more general and more realistic than MCAR. Modern missing data methods generally start from the MAR assumption.

Depends on observed data
reason for missing can be explained by other columns (observed data)

MNAR (Missing not at Random):

If neither MCAR nor MAR holds, then we speak of missing not at random (MNAR). In the literature one can also find the term NMAR (not missing at random) for the same concept. MNAR means that the probability of being missing varies for reasons that are unknown to us. For example, the weighing scale mechanism may wear out over time, producing more missing data as time progresses, but we may fail to note this. If the heavier objects are measured later in time, then we obtain a distribution of the measurements that will be distorted. MNAR includes the possibility that the scale produces more missing values for the heavier objects (as above), a situation that might be difficult to recognize and handle. An example of MNAR in public opinion research occurs if those with weaker opinions respond less often. MNAR is the most complex case. Strategies to handle MNAR are to find more data about the causes for the missingness, or to perform what-if analyses to see how sensitive the results are under various scenarios.

Rubin's distinction is important for understanding why some methods will work, and others not. His theory lays down the conditions under which a missing data method can provide valid statistical inferences. Most simple fixes only work under the restrictive and often unrealistic MCAR assumption. If MCAR is implausible, such methods can provide biased estimates.

missingo library

→ Imputing missing values with incorrect assumptions can introduce BIAS into the dataset, skewing model's predictions.

→ Imputing with incorrect values might also lead to incorrect interpretations.

Handling missing values
- Removing
- Imputation
- using Algos that can handle missing values

Removing
- By dropping cols
  - % of missing data
  - Importance of the feature
- By dropping rows
  - Small % of missing data ~ ≤ 5%
  - MCAR
  - Preserves the distribution
  - Expecting no missing values in production [not common in industry projects] (educated imputation is preffered)

Imputation
- univariate
- Multivariate

Univariate imputation

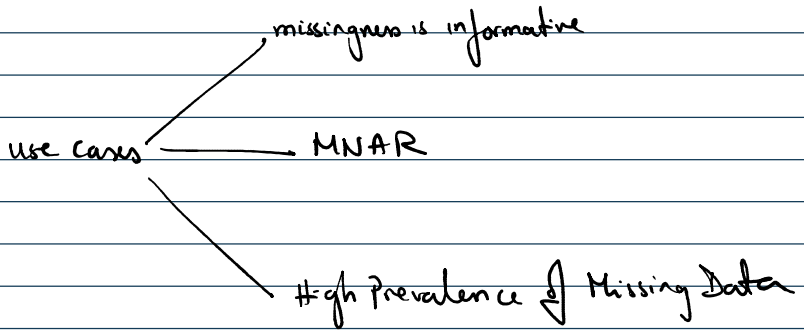
1) Missing Indicator [sklearn]

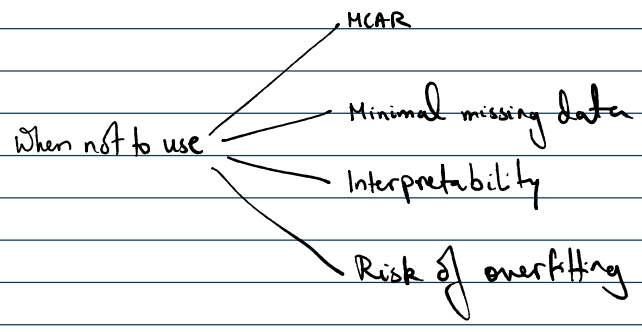
→ Binary feature that indicates whether data was missing for a certain observation in another column.

→ for each feature in the dataset with missing values, create a new feature that will have a binary value.

→ 1 - missing 0 - not missing

→ use this new feature / use both

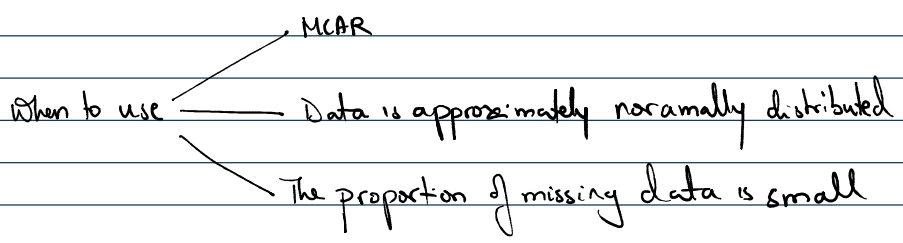


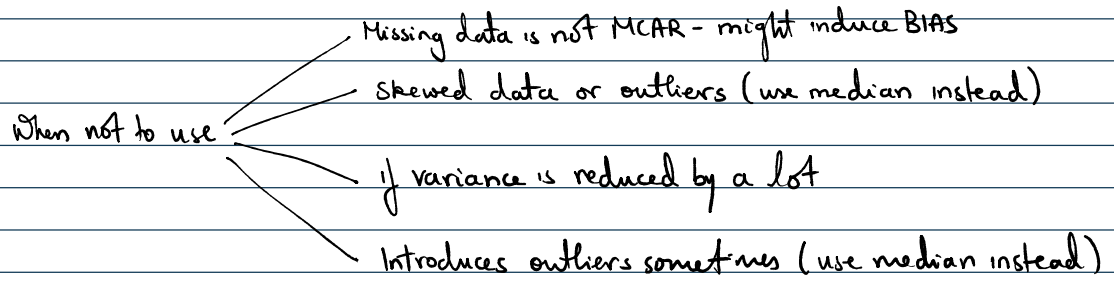


i.e. something is better than nothing.

2) Simple Imputer [sklearn]

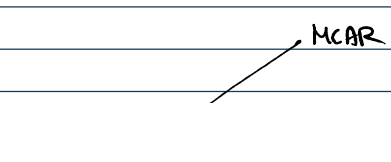
→ Mean & Median [numerical features]

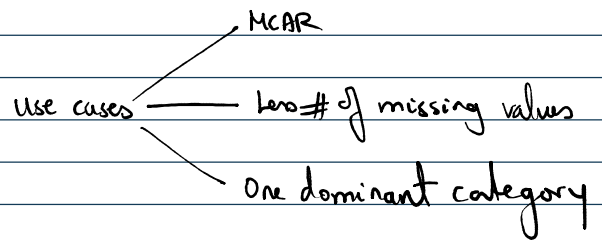




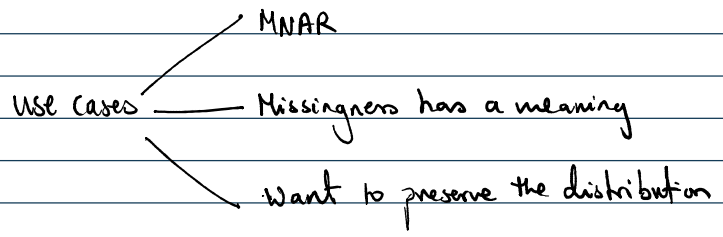
→ Check impact on correlation matrix
→ kde, box plots b4 & after.

→ Most Frequent [categorical]





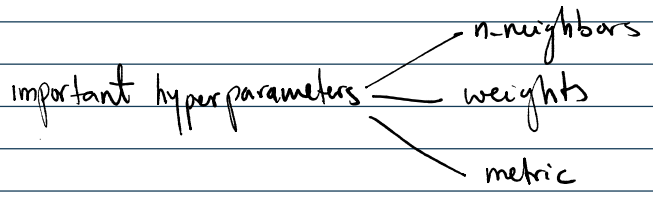
→ Constant [numerical & categorical]

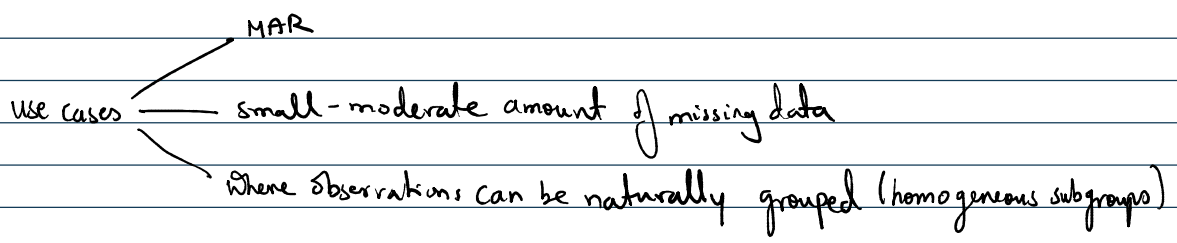


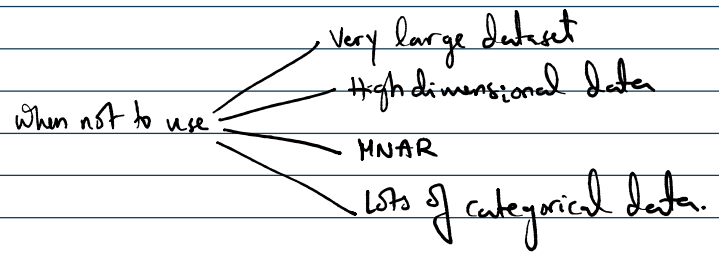
Multivariate Imputation

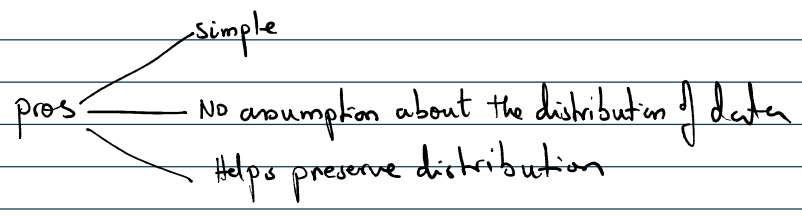
1) KNN Imputer [sklearn]

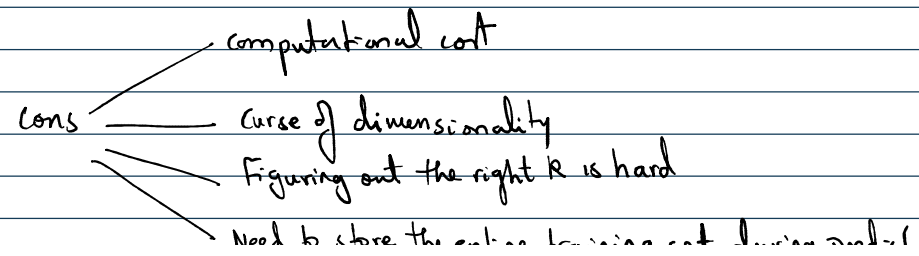
→ nan euclidean distance [sklearn does]











- Figuring out the right k is hard
- Need to store the entire training set during prediction

2) Iterative Imputer [sklearn]

- MICE Algorithm → Multiple Imputation Chained equations
- Flexible

Use cases
- missing values across multiple feature
- MAR
- datasets where relationship b/w features is complex & non-linear

When not to use
- MCAR/MNAR
- more than 50% of data is missing
- categorical Data

pros
- Flexible estimator choice

Cons
- Computation
- overfitting Risk
- sensitive to initialization [initial strategy parameter]
- convergence issue [But no loss function]

Read Docs