Minería de Datos

U3 Preprocesamiento de datos

Héctor Maravillo

Section 1

Dealing with Missing Values

Missing Values

Many existing, industrial and research data sets contain **Missing Values (MVs)** in their attribute values.

Intuitively a MV is just a value for attribute that was **not introduced** or **was lost** in the recording process.

There are various reasons for their existence, such as

- Manual data entry procedures.
- Equipment errors.
- Incorrect measurements.

The presence of such imperfections usually requires a **preprocessing stage** in which the data is prepared and clean, in order to be useful to and sufficiently clear for the knowledge extraction process.

Discard MVs

The simplest way of dealing with MVs is **to discard** the examples that contain them.

However, this method is practical only when the data contains a relatively **small number** of examples with MVs.

Problems associated with MVS

Inappropriate handling of the MVs in the analysis may introduce **bias** and can result in **misleading conclusions** being drawn from a research study, and can also limit the **generalizability** of the research findings.

Three types of problems are usually associated with MVs:

- Loss of **efficiency**.
- Complications in handling and analyzing the data.
- **Bias** resulting from differences between missing and complete data.

:

Treatment of MVs

Usually the treatment of MVs in DM can be handled in three different ways:

- The first approach is **to discard** the **examples** with MVs in their attributes. Therefore **deleting attributes** with elevated levels of MVs is included in this category too.
- Another approach is the use of **maximum likelihood procedures**, where the parameters of a **model** for the complete portion of the data are estimated, and later used for **imputation** by means of **sampling**.
- Finally, the **imputation** of MVs is a class of procedures that aims to fill in the MVs with **estimated** ones. In most cases, a data set's attributes **are not independent** from each other. Thus, through the identification of relationships among attributes, MVs can be determined

Assumptions and Missing Data Mechanism

Let X be a $n \times m$ rectangular matrix of data, where n is the number of instances and m the number of attributes.

Usually, it is denoted as x_i the i-th row of x.

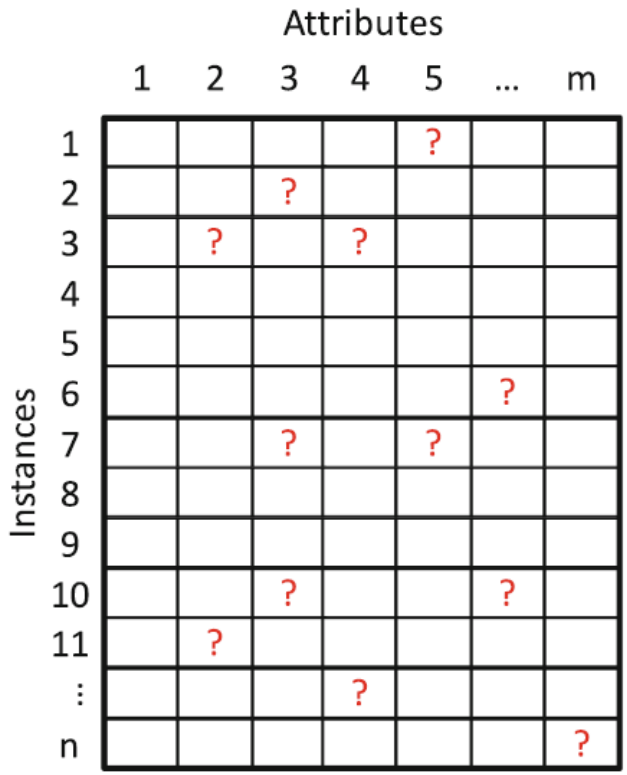

Figure: Data set with MVs denoted with a "?"

Independent and Identically Distributed assumption

A common assumption is that the instances are all **independent and identically distributed (i.i.d.)** draws of some **multivariate probability distribution**.

If we consider the i.i.d. assumption, the **probability function** of the complete data can be written as follows:

$$P(X|\theta) = \Pi_{i=1}^{n} f(x_i|\theta)$$

where f is the probability function for a single case and θ represents the parameters of the model that yield such a particular instance of data.

The main problem is that the particular parameters' value θ for given data are very rarely known.

Independent and Identically Distributed assumption

For this reason authors usually overcome this problem by considering distributions that are commonly found in nature and their properties are well known as well.

The three **distributions** that standout among these are:

- The **multivariate normal distribution** in the case of only real valued parameters.
- The **multinomial model** for **cross-classified categorical data** when the data consists of nominal features.
- **Mixed models** for combined normal and categorical features in the data.

Missing at Random (MAR)

Let X_{obs} be the observed part of X and we denote the missing part as X_{mis} so that

$$X = (X_{obs}, X_{mis})$$

Informally talking, when the probability that an observation is missing may depend on X_{obs} but not on X_{mis} we can state that the missing data is **missing at random (MAR)**.

In the case of MAR missing data mechanism, given a particular value or values for a set of features belonging to X_{obs}, the distribution of the rest of features is the same between the observed cases and the missing cases.

Missing at Random (MAR)

Let suppose that we dispose an $n \times m$ matrix called B of variables whose values are 1 or 0 when X elements are observed and missing respectively.

The distribution of B should be related to X and to some unknown parameter ζ, so we dispose a probability model for B described by $P(B|X, \zeta)$.

Having a **MAR assumption** means that this distribution does not depend on X_{mis}:

$$P(B|X_{obs}, X_{mis}, \zeta) = P(B|X_{obs}, \zeta)$$

Missing Completely at Random (MCAR)

The **MAR assumption** does not suggest that the missing data values constitute just another possible sample from the probability distribution.

This condition is known as **missing completely at random (MCAR)**.

MCAR is a special case of MAR in which the distribution of an example having a MV for an attribute does not depend on either the observed or the unobserved data, that is:

$$P(B|X_{obs}, X_{mis}, \zeta) = P(B|\zeta)$$

Missing Completely at Random (MCAR)

Although there will generally be some **loss of information**, comparable results can be obtained with missing data by carrying out the same analysis that would have been performed with no MVs.

In practice this means that, under MCAR, the analysis of **only those units with complete data** gives valid inferences.

Not Missing at Random (NMAR)

A third case arises when MAR does not apply as the MV depends on both the rest of observed values and the proper value itself. That is

$$P(B|X_{obs}, X_{mins}, \zeta)$$

is the actual probability estimation.

This model is usually called **not missing at random (NMAR)** or **missing not at random (MNAR)** in the literature.

This model of missingness is a challenge for the user as the only way to obtain an **unbiased estimate** is to also model missing data.

Missing data randomness

In summary, **missing data randomness** can be divided into three classes:

- **Missing Completely at Random (MCAR)**. This is the highest level of randomness. It occurs when the probability of an instance (case) having a missing value for an attribute **does not depend** on either the **known values** or the **missing data**. In this level of randomness,**any** missing data treatment method can be applied **without risk** of introducing **bias** on the data.
- **Missing At Random (MAR)**. When the probability of an instance having a missing value for an attribute **may depend** on the **known values**, but not on the value of the missing data itself.
- **Not Missing At Random (NMAR)**. When the probability of an instance having a missing value for an attribute **could depend** on the **value of that attribute**.

Missing data randomness

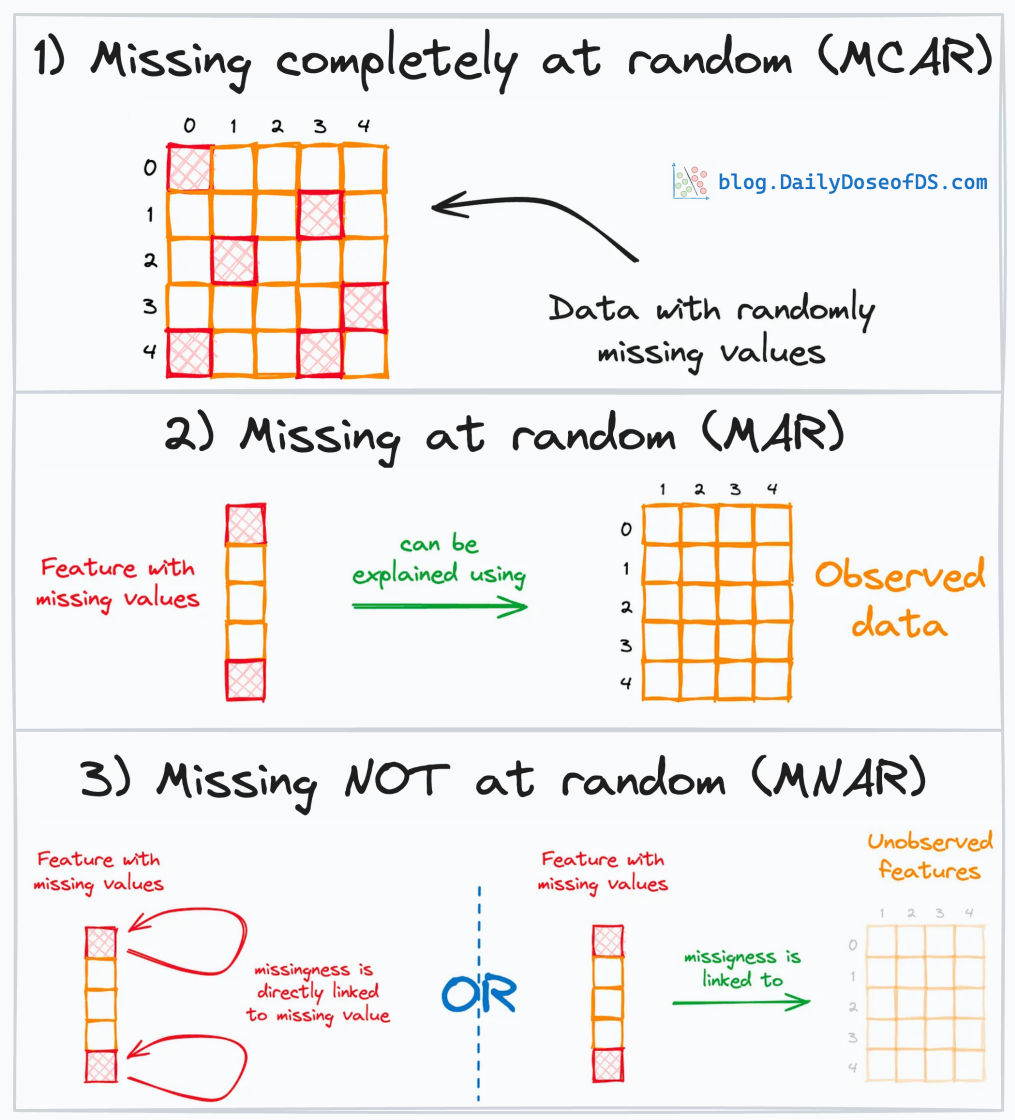

Figure: Missing data randomness

Ignore Missing (IM)

A very common approach in the specialized literature is to apply **case deletion** or **ignore missing (IM)**.

Using this method, all instances with **at least one MV** are **discarded** from the data set.

Although IM often results in a substantial decrease in the **sample size** available for the analysis, it does have important advantages.

Ignore Missing (IM)

- Under the assumption that data is **MCAR**, it leads to **unbiased parameter estimates**.
  Unfortunately, even when the data are MCAR there is a loss in **power**[1] using this approach, especially if we have to rule out a large number of subjects.
- When the data is **not MCAR**, it **biases** the results.
  For example when low income individuals are less likely to report their income level, the resulting mean is biased in favor of higher incomes.

[1] The **power** of a statistical test is defined as the probability of correctly rejecting the **null hypothesis** H_0 when it is actually false:

$$P(\text{reject } H_0 \mid H_0 \text{ is false})$$

Substitution of MVs

The **substitution** of the MVs for the global:

- **Most common** attribute value for nominal attributes.
- **Average value** for numerical attributes

is widely used, especially when many instances in the data set contain MVs and to apply a approach of **to do not impute (DNI)**, would result in a very reduced and unrepresentative pre-processed data set.

Hot Deck

In the **Hot Deck method**, a missing attribute value is filled in with a value from an **estimated distribution** for the missing value from the current data.

Hot deck is typically implemented in two stages.

- In the first stage, the data are **partitioned** into clusters.
- In the second stage, each instance with missing data is **associated** with one cluster.

The complete cases in a **cluster** are used **to fill** in the missing values.

This can be done by calculating the mean or mode of the attribute within a cluster.

Cold deck imputation is similar to hot deck, but the data source must be other than the current data source

Imputation with K-Nearest Neighbor (KNNI)

Using this instance-based algorithm, every time an MV is found in a current instance, **KNNI** computes the K-**Nearest Neighbor (KNN)** and a value from them is imputed.

- For nominal values, the **most common value (mode)** among all neighbors is taken.
- For numerical values the **average value** among all neighbors is used.

Imputation with K-Nearest Neighbor (KNNI)

The KKNI selects instances with expression profiles similar to the instance of interest **to impute** missing values.

If we consider instance A that has one missing value in attribute 1, this method would find K other instances, which have a value present in attribute 1, with values most similar to A in attributes 2-m (where m is the total number of attributes).

Imputation with K-Nearest Neighbor (KNNI)

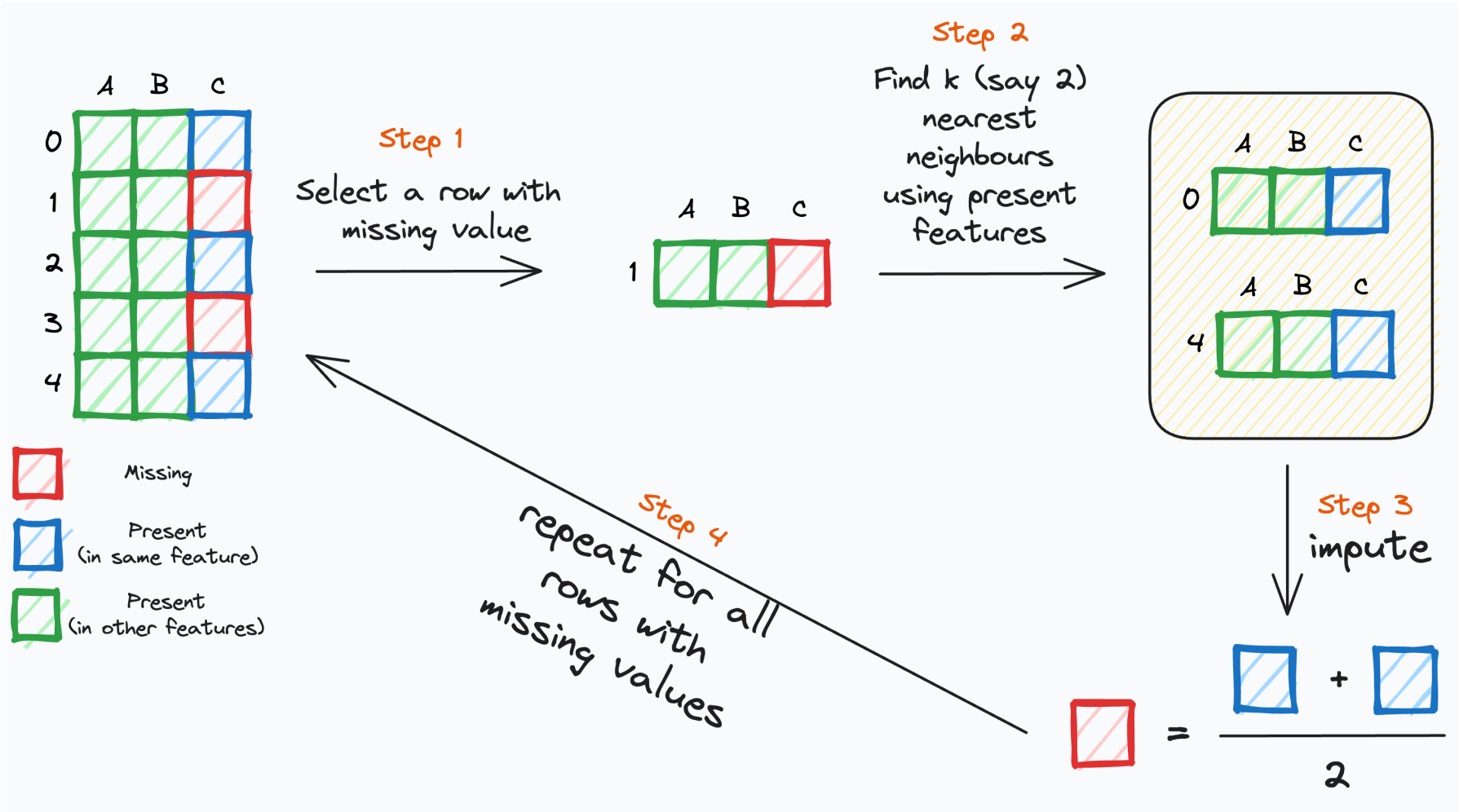

Figure: Visual illustration to kNN Imputer (Chawla, 2023)

Imputation with K-Nearest Neighbor (KNNI)

A **weighted average** of values in attribute 1 from the K closest instances is then used as an estimate for the missing value in instance A.

In the weighted average, the contribution of each instance is weighted by similarity of its expression to that of instance A.

Proximity measure

To define the **neighborhoods**, a proximity measure between instances is needed.

A measure based in Euclidean distance is the most commonly used in the literature. This formulation ignores feature coordinates with a missing value in either sample and scales up the weight of the remaining coordinates.

For example, the distance between

$$X = (3, *, *, 6)^T \qquad \wedge \qquad y = (1, *, 4, 5)^T$$

is

$$\sqrt{\frac{4}{2}\left[(3-1)^2 + (6-5)^2\right]}$$

Imputation of missing values

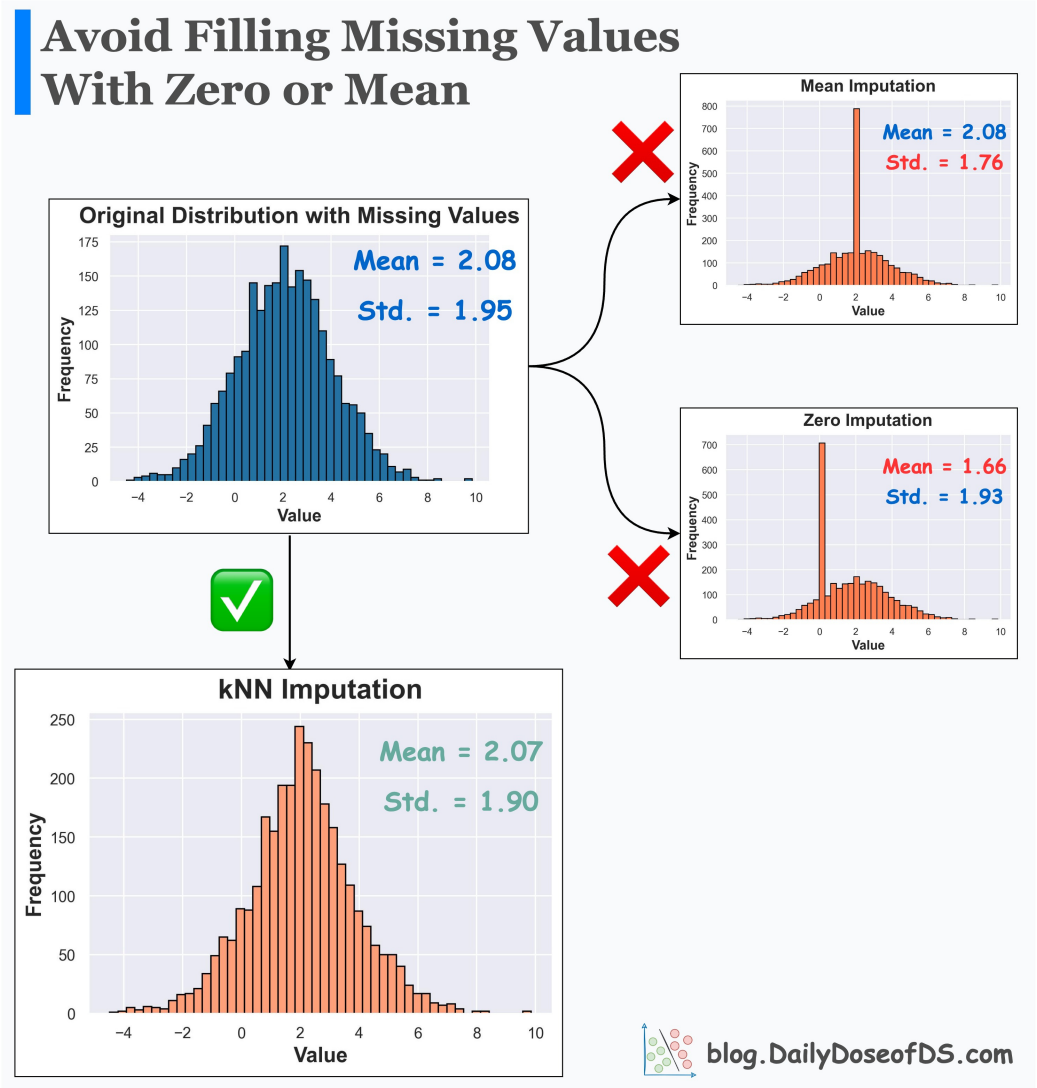

Figure: Imputation of missing values (Chawla, 2023)

References

- Batista, G. & Monard, M. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17: 519–533.
- García, S., Luengo, J. & Herrera, F. *Data Preprocessing in Data Mining*, Springer, 2015.
- Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P. Hastie, T., Tibshirani, R., Botstein, D. & Altman, B. (2001). Missing value estimation methods for DNA microarrays, *Bioinformatics*, 17(6): 520–525.