PROBLEM DATA: MISSING DATA AND CAUSES

WHAT ARE MISSING DATA?

- In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation.

- This situation mostly occur as a result of manual data entry procedures, equipment errors and incorrect measurements.

PROBLEMS OF MISSING DATA

-loss of efficiency

- complications in handling and analyzing the data

- bias resulting from differences between missing and complete data.

TYPES OF MISSING DATA

- MISSING COMPLETELY AT RANDOM (MCAR)

Values in a data set can miss completely at random (MCAR) if the events that lead to any particular data-item missing are independent both of observable variables and of unobservable parameters of interest, and occur entirely at random.

TYPES OF MISSING DATA (Cont'd)

- MISSING AT RANDOM (MAR)

Missing at random (MAR) is an alternative, and occurs when the missing-ness is related to a particular variable, but it is not related to the value of the variable that has missing data.

An example of this is accidentally omitting an answer on a questionnaire.

TYPES OF MISSING DATA (Cont'd)

- MISSING NOT AT RANDOM (MNAR)

This is data that is missing for a specific reason (i.e. the value of the variable that is missing is related to the reason it is missing).

An example of this is if certain question on a questionnaire tend to be skipped deliberately by participants with certain characteristics.

HOW TO DEAL WITH MISSING DATA

Missing data reduce the representativeness of the sample and can therefore distort inferences about the population.

- There is no need to use a special method for dealing missing values if method that is used for data analysis has its own policy for handling missing values.

HOW TO DEAL WITH MISSING DATA

- Decision rules extraction methods may consider attributes with missing data as irrelevant

- Association rules extraction methods may ignore rows with missing values (conservative approach)

- handle missing values as they are supporting the rule (optimistic approach)

HOW TO DEAL WITH MISSING DATA

- REDUCING THE DATA SET

The simplest solution for the missing data imputation problem is the reduction of the data set and elimination of all missing values. This can be done by elimination of samples (rows) with missing values or elimination of attributes (columns) with missing values

HOW TO DEAL WITH MISSING DATA

TREATING MISSING ATTRIBUTE VALUES AS SPECIAL VALUES.

This method deals with the unknown attribute values using a totally different approach. Rather than trying to find some known attribute value as its value, we treat missing value itself as a new value for the attributes that contain missing values and treat it in the same way as other values.

HOW TO DEAL WITH MISSING DATA

REPLACE MISSING VALUE WITH MEAN

This method replaces each missing value with mean of the attribute.

The mean is calculated based on all known values of the attribute. This method is usable only for numeric attributes and is usually combined with replacing missing values with most common attribute value for symbolic attributes.

HOW TO DEAL WITH MISSING DATA

REPLACE MISSING VALUE WITH MEAN FOR THE GIVEN CLASS

This method is similar to the previous one. The difference is in that the mean is not calculated from all known values of the attribute, but only attributes values belonging to given class are used.

HOW TO DEAL WITH MISSING DATA

REPLACE MISSING VALUE WITH MOST COMMON ATTRIBUTE VALUE

Since the mean is affected by the presence of outliers it seems natural to use the median instead just to assure robustness. In this case the missing data for a given attribute is replaced by the median of all known values of that attribute in the class where the instance with the missing value belongs

HOW TO DEAL WITH MISSING DATA

CLOSEST FIT

The closest fit algorithm for missing attribute values is based on replacing a missing attribute value with an existing value of the same attribute from another case that resembles as much as possible the case with missing attribute values

HOW TO DEAL WITH MISSING DATA

E.g. if 5 attributes of given case is known and 6th is being searched, the 6th attribute value is taken from the case which has other 5 attributes values most similar to the given case. The proximity measure between two cases x and y is the Manhattan distance between x and y, i.e.,

HOW TO DEAL WITH MISSING DATA

$$\text{distance}(x, y) = \sum_{i=1}^{n} \text{distance}(x_i, y_i),$$

Where

$$\text{distance}(x_i, y_i) = \begin{cases} 0 \text{ if } x_i = y_i, \\ 1 \text{ if x and y are symbolic and } x_i \neq y_i \text{ or } x_i = ? \text{ or } y_i = ? \\ \dfrac{|x_i - y_i|}{r} \text{ if } x_i \text{ and } y_i \text{ are numbers and } x_i \neq y_i \end{cases}$$

where r is the difference between the maximum and minimum of the known values.

Problem with using only one closest case may lied in replacing missing value by an outlier.

HOW TO DEAL WITH MISSING DATA

The problem is illustrated by following example

Tab. 1: Closest fit example

Property table	
floor area [m^2]	rental price [CZK]
45	10 000
52	13 000
52	11 000
54	?
55	18 000
62	13 500
62	12 000

THANK YOU ALL FOR LISTENING