Data Cleaning

Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

- Basic methods of Data cleaning
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data

Missing Values

- Data is not always available. E.g., many tuples have no recorded value for several attributes, such as customer income in sales data

Methods to Handle Missing Data:

- ✓ Ignore the tuple.
- ✓ Fill in the missing value manually.
- ✓ Use a global constant to fill in the missing value.
- ✓ Use a measure of central tendency for the attribute (e.g., the mean or median) to
- ✓ fill in the missing value.
- ✓ Use the attribute mean or median for all samples belonging to the same class as
- ✓ the given tuple.
- ✓ Use the most probable value to fill in the missing value.

Noisy Data

Noise is a random error or variance in a measured variable. Following are the data smoothing techniques used to remove the noise.

Binning:

- Binning methods smooth a sorted data value by consulting its "neighbourhood," that is, the values around it. The sorted values are distributed into a number of "buckets," or bins.
- In smoothing by bin means, each value in a bin is replaced by the mean value of the bin.
- Smoothing by bin medians can be employed, in which each bin value is replaced by the bin median.
- In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value.

Regression: Data smoothing can also be done by regression, a technique that conforms data values to a function. Linear regression involves finding the "best" line to fit two attributes (or variables) so that one attribute can be used to predict the other. Multiple linear regression is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.

Outlier analysis: Outliers may be detected by clustering, for example, where similar values are organized into groups, or "clusters." Intuitively, values that fall outside of the set of clusters may be considered outliers

Data Cleaning as a Process

- The first step in data cleaning as a process is discrepancy detection.
- Discrepancies can be caused by several factors, including poorly designed data entry forms that have many optional fields, human error in data entry, deliberate errors (e.g., respondents not wanting to divulge information about themselves), and data decay (e.g., outdated addresses).
- Discrepancies may also arise from inconsistent data representations and inconsistent use of codes. Other sources of discrepancies include errors in instrumentation devices that record data and system errors.
- Errors can also occur when the data are used for purposes other than originally intended. There may also be inconsistencies due to data integration.