

CS685: DATA MINING

DATA PREPROCESSING AND DATA CLEANING

Arnab Bhattacharya
`arnabb@cse.iitk.ac.in`

Computer Science and Engineering,
Indian Institute of Technology, Kanpur
<http://web.cse.iitk.ac.in/~cs685/>

1st semester, 2021-22
Mon 1030-1200 (online)

- Data should have the following qualities
 - Accuracy
 - Completeness
 - Consistency
 - Timeliness
 - Reliability
 - Interpretability
 - Availability

Types of Data

- Data values can be classified as **discrete** or **continuous**

Types of Data

- Data values can be classified as **discrete** or **continuous**
- *Discrete*
 - Finite or countably infinite set of values
 - Countably infinite sets have a one-to-one correspondence with the set of natural numbers

Types of Data

- Data values can be classified as **discrete** or **continuous**
- *Discrete*
 - Finite or countably infinite set of values
 - Countably infinite sets have a one-to-one correspondence with the set of natural numbers
- *Continuous*
 - Real numbers
 - Precision of measurement and machine-representation limit possibilities
 - Not continuous in the actual sense

Types of Data

- Data values can be classified as **discrete** or **continuous**
- *Discrete*
 - Finite or countably infinite set of values
 - Countably infinite sets have a one-to-one correspondence with the set of natural numbers
- *Continuous*
 - Real numbers
 - Precision of measurement and machine-representation limit possibilities
 - Not continuous in the actual sense
- Data can also be classified in other ways

Categorical Data

- Categorical data is *qualitative*

Categorical Data

- **Categorical** data is *qualitative*
- **Nominal**
 - Categories
 - Example: color
 - Operations: equal, not equal
- **Binary**
 - Special case of nominal
 - Example: gender, diabetic
 - Symmetric: Two cases are equally important
 - Asymmetric: One case is more important

Categorical Data

- **Categorical** data is *qualitative*
- **Nominal**
 - Categories
 - Example: color
 - Operations: equal, not equal
- **Binary**
 - Special case of nominal
 - Example: gender, diabetic
 - Symmetric: Two cases are equally important
 - Asymmetric: One case is more important
- **Ordinal** or **Rank** or **Ordered scalar**
 - Can order
 - Example: small, medium, large
 - Operations: equality, lesser, greater

Categorical Data

- **Categorical** data is *qualitative*
- **Nominal**
 - Categories
 - Example: color
 - Operations: equal, not equal
- **Binary**
 - Special case of nominal
 - Example: gender, diabetic
 - Symmetric: Two cases are equally important
 - Asymmetric: One case is more important
- **Ordinal** or **Rank** or **Ordered scalar**
 - Can order
 - Example: small, medium, large
 - Operations: equality, lesser, greater
 - Difference has no meaning

Numeric Data

- **Numeric** data is *quantitative*

Numeric Data

- **Numeric** data is *quantitative*
- **Ratio-scaled**
 - Has a *zero point*: absolute values are ratios of each other
 - Example: temperature in Kelvin, age, mass, length
 - Operations: difference, ratio

Numeric Data

- **Numeric** data is *quantitative*
- **Ratio-scaled**
 - Has a *zero point*: absolute values are ratios of each other
 - Example: temperature in Kelvin, age, mass, length
 - Operations: difference, ratio
- **Interval-scaled**
 - Measured on equal sized units
 - Example: temperature in Celsius, date
 - No zero point: absolute value has no meaning
 - Operations: difference

Data Errors and Parameters

- Errors in data due to
 - Measurement error
 - Data collection error
 - **Noise**: probabilistic
 - **Artifact**: deterministic distortions

Data Errors and Parameters

- Errors in data due to
 - Measurement error
 - Data collection error
 - **Noise**: probabilistic
 - **Artifact**: deterministic distortions
- Parameters to measure the quality of measurements
 - **Precision**: closeness of repeated measurements
 - **Bias**: systematic variation of measurements
 - **Accuracy**: closeness of measurements to true value

Data Errors and Parameters

- Errors in data due to
 - Measurement error
 - Data collection error
 - **Noise**: probabilistic
 - **Artifact**: deterministic distortions
- Parameters to measure the quality of measurements
 - **Precision**: closeness of repeated measurements
 - **Bias**: systematic variation of measurements
 - **Accuracy**: closeness of measurements to true value
- Data problems
 - Missing values
 - Noise
 - Outliers
 - Inconsistent values
 - Duplicate objects

Data Errors and Parameters

- Errors in data due to
 - Measurement error
 - Data collection error
 - **Noise**: probabilistic
 - **Artifact**: deterministic distortions
- Parameters to measure the quality of measurements
 - **Precision**: closeness of repeated measurements
 - **Bias**: systematic variation of measurements
 - **Accuracy**: closeness of measurements to true value
- Data problems
 - Missing values
 - Noise
 - Outliers
 - Inconsistent values
 - Duplicate objects
- *Domain knowledge* about data and attributes helps data mining

Data Preprocessing

- **Data preprocessing** is the process of preparing the data to be fit for data mining algorithms and methods
- Known as **ETL** (Extract, Transform, Load)
- It may involve one or more of the following steps
 - Data cleaning
 - Data reduction/summarization
 - Data integration
 - Data transformation

Data Cleaning

- Process of handling errors in data
- Different ways
- Filling in missing values
- Handling noise
- Removing outliers
 - One of the main methods in handling noise
- Resolving inconsistent data
 - Out of range
 - Once identified as inconsistent data, handled as missing value
- De-duplicating duplicated objects

Missing Values

- Ignore the data object

Missing Values

- Ignore the data object
- Ignore only the missing attribute during analysis

Missing Values

- Ignore the data object
- Ignore only the missing attribute during analysis
- Estimate the missing value

Missing Values

- Ignore the data object
- Ignore only the missing attribute during analysis
- Estimate the missing value
- Use a measure of overall *central tendency*
 - Mean or median
- Use a measure of central tendency from only the *neighborhood*

Missing Values

- Ignore the data object
- Ignore only the missing attribute during analysis
- Estimate the missing value
- Use a measure of overall *central tendency*
 - Mean or median
- Use a measure of central tendency from only the *neighborhood*
- Interpolation
 - Useful for temporal and spatial data

Missing Values

- Ignore the data object
- Ignore only the missing attribute during analysis
- Estimate the missing value
- Use a measure of overall *central tendency*
 - Mean or median
- Use a measure of central tendency from only the *neighborhood*
- Interpolation
 - Useful for temporal and spatial data
- Use the most probable value
 - Mode

Noise

- **Noise** is a *random* perturbation in the data
- It is generally assumed that magnitude of noise is smaller than magnitude of attribute of interest
 - **Signal-to-noise** ratio should not be too low
- **White noise**
 - Gaussian distribution with zero mean

Noise

- **Noise** is a *random* perturbation in the data
- It is generally assumed that magnitude of noise is smaller than magnitude of attribute of interest
 - **Signal-to-noise** ratio should not be too low
- **White noise**
 - Gaussian distribution with zero mean
- As opposed to noise, bias can be corrected since it is deterministic

Noise

- **Noise** is a *random* perturbation in the data
- It is generally assumed that magnitude of noise is smaller than magnitude of attribute of interest
 - **Signal-to-noise** ratio should not be too low
- **White noise**
 - Gaussian distribution with zero mean
- As opposed to noise, bias can be corrected since it is deterministic
- Histogram binning
 - Bin values are replaced by mean or median
 - Equi-width histograms are more common than equi-depth

- **Noise** is a *random* perturbation in the data
- It is generally assumed that magnitude of noise is smaller than magnitude of attribute of interest
 - **Signal-to-noise** ratio should not be too low
- **White noise**
 - Gaussian distribution with zero mean
- As opposed to noise, bias can be corrected since it is deterministic
- Histogram binning
 - Bin values are replaced by mean or median
 - Equi-width histograms are more common than equi-depth
- Regression
 - Fitting a function to describe the values
 - Small values of noise do not affect the overall fit
 - Noisy value replaced by most likely value predicted by the function

- **Noise** is a *random* perturbation in the data
- It is generally assumed that magnitude of noise is smaller than magnitude of attribute of interest
 - **Signal-to-noise** ratio should not be too low
- **White noise**
 - Gaussian distribution with zero mean
- As opposed to noise, bias can be corrected since it is deterministic
- Histogram binning
 - Bin values are replaced by mean or median
 - Equi-width histograms are more common than equi-depth
- Regression
 - Fitting a function to describe the values
 - Small values of noise do not affect the overall fit
 - Noisy value replaced by most likely value predicted by the function
- **Outlier** identification and removal

Data Duplication

- Same (or almost same) values

Data Duplication

- Same (or almost same) values
- Duplicate objects may appear during data insertion or data transfer
- Mostly due to data collection errors

Data Duplication

- Same (or almost same) values
- Duplicate objects may appear during data insertion or data transfer
- Mostly due to data collection errors
- Introduces errors in statistics about the data
- If most attributes are exact copies, then it is easy to remove
- Sometimes one or more attributes are slightly different
- *Domain knowledge* needs to be utilized to identify such cases
- Process is called **de-duplication**

Data Integration

- **Data integration** is the process of transforming multiple data sources into one single coherent source
- Useful when there are multiple databases about the same set of objects

Data Integration

- **Data integration** is the process of transforming multiple data sources into one single coherent source
- Useful when there are multiple databases about the same set of objects
- **Schema matching** and **entity identification**
 - Is `cust_id` equal to `cust_number`?
- Correlation analysis to reduce redundancy
- Chi-square test for categorical data
- De-duplication

Data Transformation

- Data transformation is useful when

Data Transformation

- **Data transformation** is useful when
 - Identifying trends
 - Normalizing to correctly get statistics
 - Applying particular data mining algorithms

Data Transformation

- **Data transformation** is useful when
 - Identifying trends
 - Normalizing to correctly get statistics
 - Applying particular data mining algorithms
- Smoothing of bins using histograms
- Aggregation and summarization
- Generalization
- Normalization

Normalization

- **Normalization** changes the range of values

Normalization

- **Normalization** changes the range of values
- **Min-max normalization**

$$x' = \frac{x - \min}{\max - \min}$$

- This puts range to

Normalization

- **Normalization** changes the range of values
- **Min-max normalization**

$$x' = \frac{x - \min}{\max - \min}$$

- This puts range to (0, 1)
- If new range is (min', max')

Normalization

- **Normalization** changes the range of values
- **Min-max normalization**

$$x' = \frac{x - \min}{\max - \min}$$

- This puts range to (0, 1)
- If new range is (\min' , \max')

$$x' = \left(\frac{x - \min}{\max - \min} \right) (\max' - \min') + \min'$$

Normalization

- **Normalization** changes the range of values
- **Min-max normalization**

$$x' = \frac{x - \min}{\max - \min}$$

- This puts range to (0, 1)
- If new range is (\min' , \max')

$$x' = \left(\frac{x - \min}{\max - \min} \right) (\max' - \min') + \min'$$

- **Z-score normalization**

$$x' = \frac{x - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation

- This puts range to

Normalization

- **Normalization** changes the range of values
- **Min-max normalization**

$$x' = \frac{x - \min}{\max - \min}$$

- This puts range to (0, 1)
- If new range is (min', max')

$$x' = \left(\frac{x - \min}{\max - \min} \right) (\max' - \min') + \min'$$

- **Z-score normalization**

$$x' = \frac{x - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation

- This puts range to $(-\infty, +\infty)$
- Also called **standard score** or **z-score** since it corresponds to the standard normal distribution $N(0, 1)$