Data Preprocessing

There are several techniques used to preprocess data

1. Data cleaning
2. Data integration
3. Data reduction
4. Data transformation

The quality of data is defined by

1. Accuracy
2. Completeness
3. Consistency
4. Timeliness
5. Believability
6. Interpretability

**Data Cleaning**

Missing values are rectified by

1. Ignore the tuple
2. Fill the missing value manually
3. Use a global constant to fill the value
4. Use a measure of central tendency to fill a value
5. Use attribute mean or median for all samples of the same class of the tuple
6. Use the most probable value

Noisy data can be removed by techniques such as

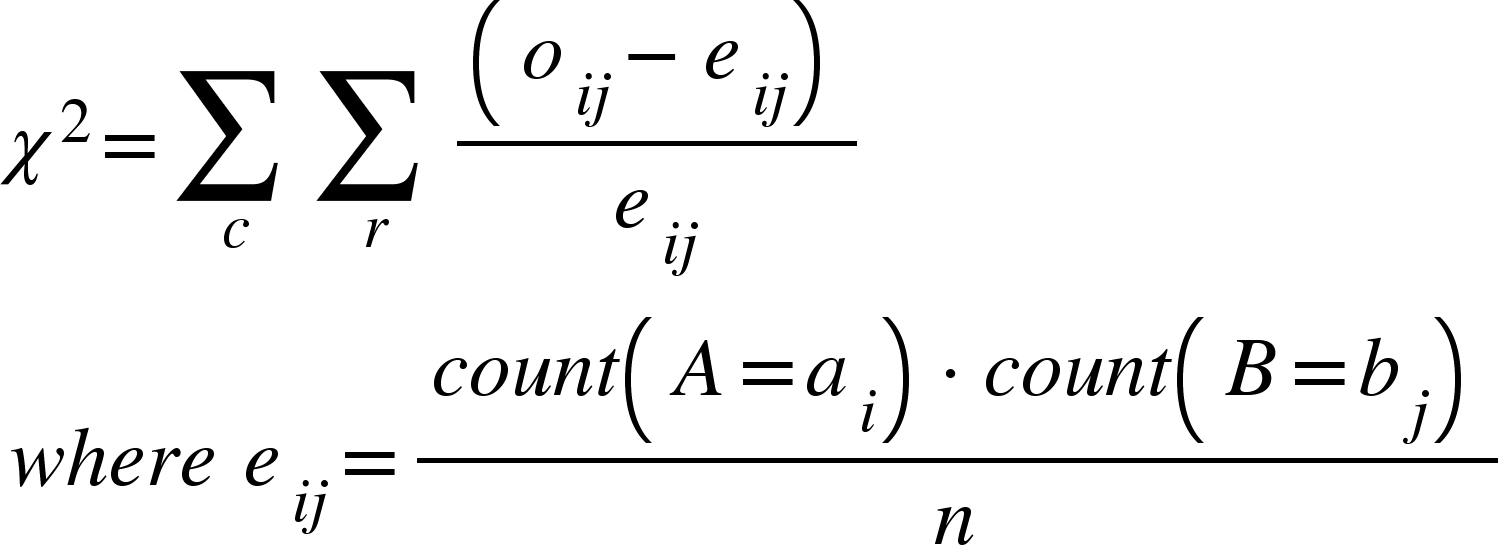
1. Smoothing
   1. Consult a neighborhood of values
   2. Bin means
   3. Bin medians
   4. Bin boundaries
2. Regression
3. Outlier analysis

Process of data cleaning is based on discrepancy detection

1. Discrepancies can occur due to poorly designed data entry, deliberate errors,etc
2. Use metadata as ground truth to identify the discrepancies
3. Data should be observed with regards to
   1. Unique rule : Each value must be unique for that attribute
   2. Consecutive rule : There can be no missing values
   3. Null rule : Specifies the use of blanks or nulls
4. Data scrubbing and auditing tools can be used as well
5. Data migration can be used for simple transforms, and ETL can be used for the same
6. But this is error prone and time consuming and might cause nested discrepancies
7. One approach is to build a series of transformations by composing or debugging individual transformations.
8. Another is to develop declarative languages for defining data cleaning specifications

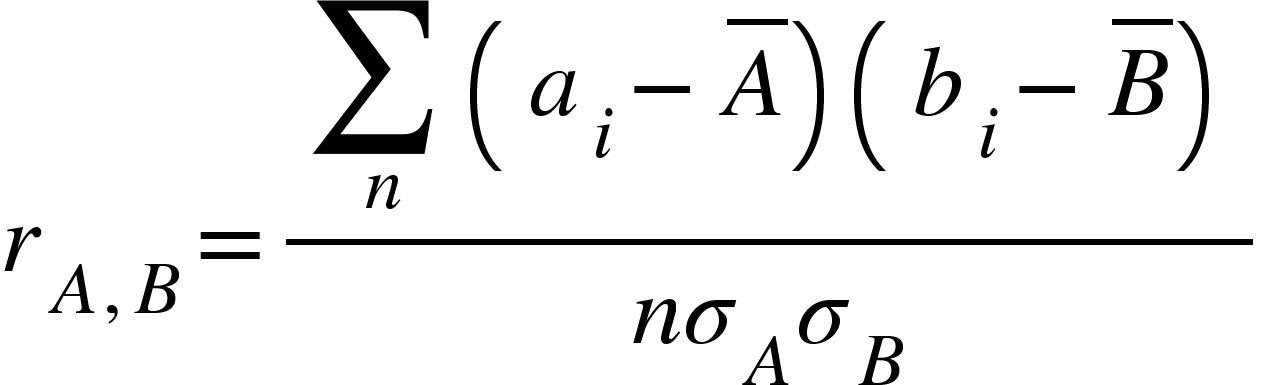
**Data Integration**

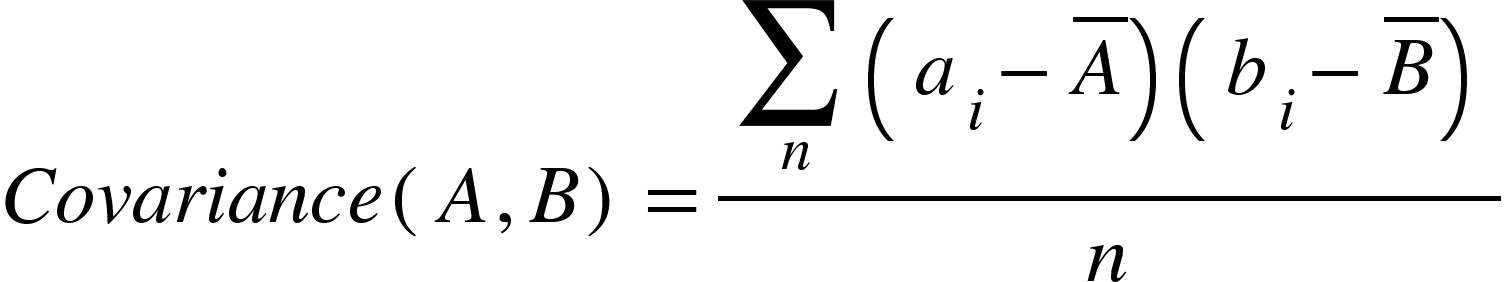
1. Entity identification problem
   1. This matches up multiple entities from different sets of data to merge them
2. Redundancy and Correlation Analysis
   1. Chi-square tests can be used to find related attributes for nominal data



This tests the hypothesis that A and B are independent.

* 1. For numeric data, Pearson’s product moment coefficient can be used as





1. Tuple Duplication
2. Data value conflict and resolution

**Data Reduction**

1. Dimensionality Reduction
2. Numerosity reduction
   1. Replace data volume by parametric models like regression or non parametric models like histograms
3. Data compression
   1. Lossy if data can’t be obtained back
   2. Lossless otherwise

**Discrete Wavelet Transforms**

This is a linear signal processing technique where a data vector X is transformed into a numerically different vector X’ of wavelet coefficients. This vector is a compressed version of the original data, which is an approximation of lossless compression.

This allows data to be transformed into a sparse form, and removes noise without smoothing out the main features of the data.

The main algorithm is

1. The length L of the input data vector is made to a power of 2 by padding the vector with 0s
2. Each transform involves applying 2 functions
   1. Data smoothing such as sum or weighted average
   2. Weighted difference to bring out features
3. Two functions applied to pairs of data points, resulting in two datasets of length L/2
4. Recursively apply these functions till length is 2
5. Selected values obtained are designated as wavelets of the data

The other data reduction techniques are

1. PCA
2. Attribute selection
   1. Stepwise forward selection
   2. Stepwise backward elimination
   3. Combination of a and b
   4. Decision tree induction
3. Regression and Log Linear models
4. Clustering
5. Sampling

**Regression and Log Linear Models : Parametric Data Reduction**

Here, we model the data as y = wx + b where w is the weight vector and b is the bias of the model.

Log linear models approximate discrete multidimensional probability distributions.

**Histograms**

They divide data into buckets of ranges where the values can exist.

1. Equal width = Width of each bucket range is uniform
2. Equal frequency = Frequency of each bucket is constant

**Sampling**

Can be of 4 types

1. Simple random with replacement
2. Simple random without replacement
3. Cluster sampling
4. Stratified sampling

Data cube aggregation is used to aggregate a set of tables to make it a representative of a higher range of data in time.

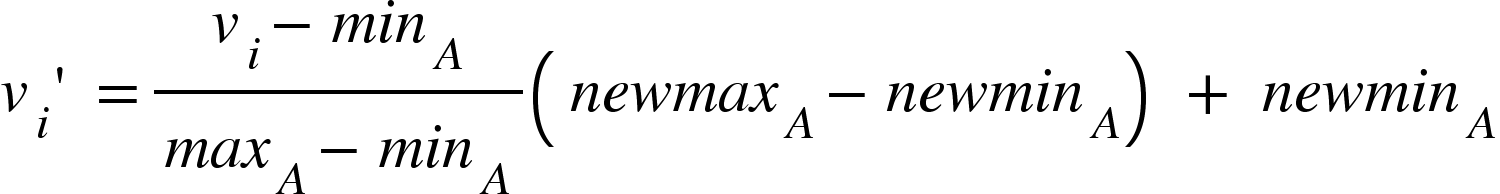
**Data Transformation**

Strategies for data transformation include

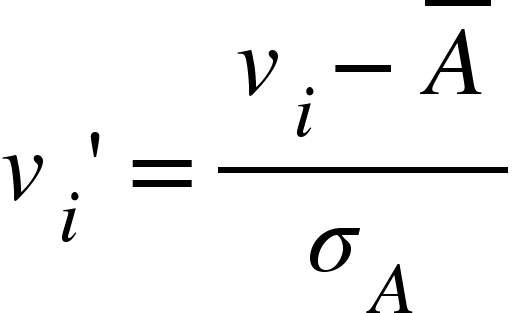
1. Smoothing
2. Attribute construction
3. Aggregation
4. Normalisation
5. Discretization
   1. Supervised using class descriptions
   2. Unsupervised
   3. Can be top-down or bottom-up
6. Concept hierarchy generation for nominal data

**Data Transformation by Normalisation**

1. Min-max normalisation
   1. Transform data to lie between new\_min and new\_max as



1. Z score normalisation



1. Normalisation by decimal scaling

