Concept

data mining function:

1. generalization
2. pattern discovery
3. classification
4. Cluster Analysis
5. Outliers Analysis
6. Time and Ordering: Sequential Pattern, Trend and Evolution Analysis
7. Structure and Network Analysis Graph mining

Data

characteristic of structured data:

1. Dimensionality
    1. Curse of dimensionality
2. Sparsity
    1. Only presence counts
3. Resolution
    1. Patterns depend on the scale
4. Distribution
    1. Centrality and dispersion

- Data sets are made up of data objects
- Data objects are described by *attributes*

attribute:

- dimensions, features, variables

type:

1. Nominal: `auburn, black, blond, brown, grey, red, white`
2. Binary: `0/1`
3. ordinal: `small, medium, large`
4. Interval: Measured on a scale of equal sized units `temperature`
5. Ratio: inherent zero-point `temperaturer in kelven, count`

type 2:

1. Discrete Attribute: a finite or countably infinite set of values `zip code, profession`
2. Continuous Attribute: Has real numbers as attribute values `height, weight`

statistical measurement:

mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ or $\mu = \frac{1}{N}\sum x$

weighted mean: $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$

median (approx): $L_1 + (\frac{n/2 - \sum freq_l}{freq_{median}})width$

> $\sum freq_l$: sum before the median interval
>
> $width$: interval width: $L_2 - L_1$
>
> L_1: low interval limit

mode: Value that occurs most frequently in the data

data matrix:

- A data matrix of n data points with l dimensions generate a matrix with shape $n \cdot l$
- Dissimilarity (distance) matrix: triangular matrix

$$\begin{pmatrix} 0 & & & \\ d(2,1) & 0 & & \\ \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & \dots & 0 \end{pmatrix}$$

standardizing:

- z-score: $z = \frac{x-\mu}{\sigma}$, or using mean absolute deviation

❑ An alternative way: Calculate the mean absolute deviation

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|)$$

where

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \dots + x_{nf})$$

❑ standardized measure (*z-score*): $z_{if} = \frac{x_{if} - m_f}{s_f}$

❑ Using mean absolute deviation is more robust than using standard deviation

distance:

1. Minkowski distance (L-p norm):

$$d(i,j) = \sqrt[p]{|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{il} - x_{jl}|^p}$$

properties:

❑ d(i, j) > 0 if i ≠ j, and d(i, i) = 0 (Positivity)

❑ d(i, j) = d(j, i) (Symmetry)

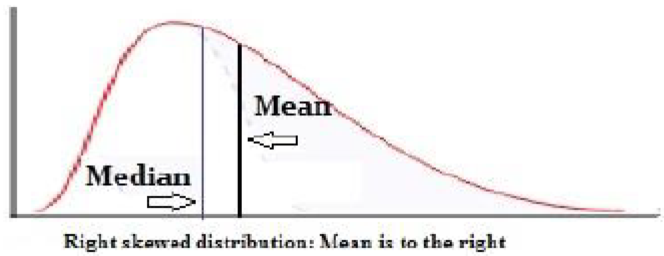
❑ d(i, j) ≤ d(i, k) + d(k, j) (Triangle Inequality)

Model

type:

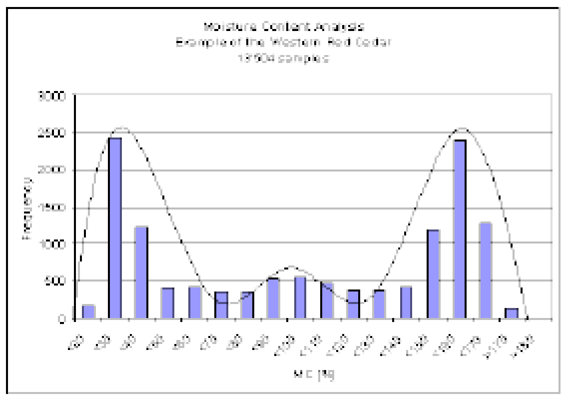
unimodal:

- Empirical formula: $mean - mode = 3 \times (mean - median)$



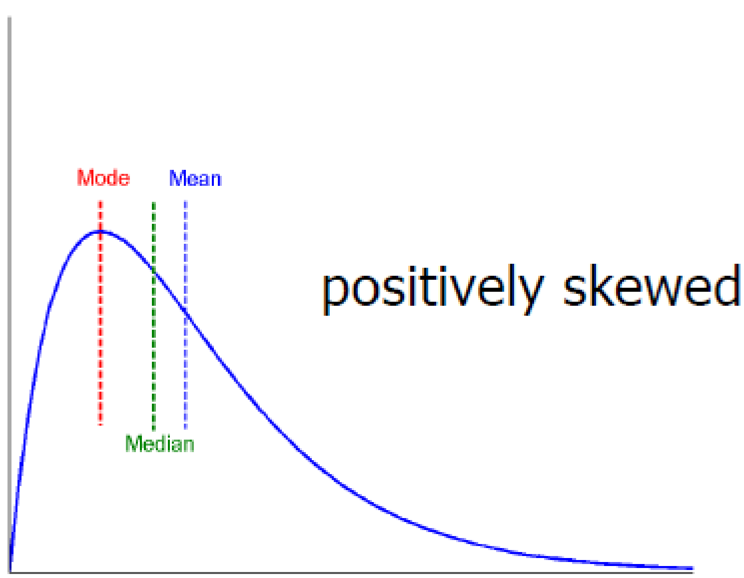
multi model:

- include bimodal and trimodal, etc. depend on peak number

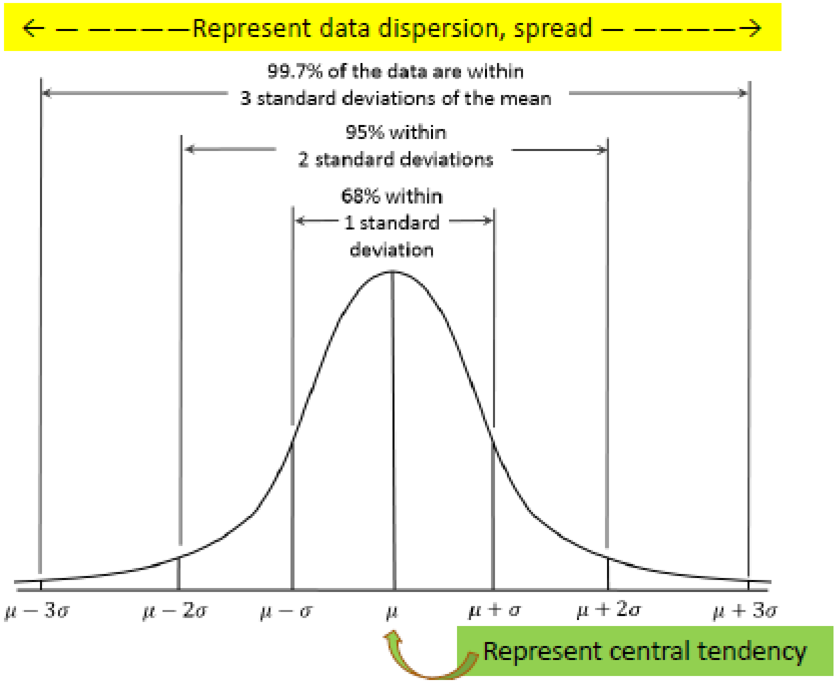


distribution:

1. symmetric
2. skewed: include positive skewed and negative skewed, their mean/median have opposite direction



normal distribution curve



measurement:

Variance ($s^2 \text{ or } \sigma^2$) and standard deviation ($s \text{ or } \sigma$) use to measure data distribution

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

> n: sample size, N: population size

Graph

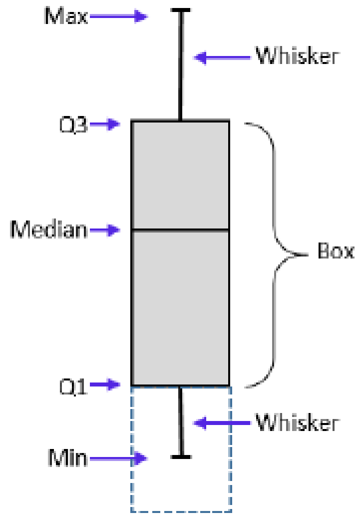
1. Boxplot: graphic display of five number summary
2. Histogram: x axis are values, y axis are frequencies
3. Quantile plot: each value x i is paired with f indicating that approximately 100 f% of data are $\leq$ x i
4. Quantile-quantile (q-q) plot : graphs the quantiles of one univariate distribution against the corresponding quantiles of another
5. Scatter plot: each pair of values is a pair of coordinates and plotted as points in the plane

box plot:

Quartiles: `Q1` (25 th percentile), `Q3` (75 th percentile)

IQR: `Q3` - `Q1`

Five number summary: `min`, `Q1` , `Q3` , `max`



Histogram:

Graph display of tabulated frequencies, shown as bars

- Differences between histograms and bar charts: Histograms are used to show distributions of variables while bar charts are used to compare variables
- Histograms Often Tell More than Boxplots: different histogram may have the same boxplot representation

Correlation

cosine Similarity:

$$cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\| d_1 \| \times \| d_2 \|}$$

chi-square test:

- The larger the X2 value, the more likely the variables are related
- Null hypothesis: The two distributions are independent
- Correlation does not imply causality

$$\chi^2 = \sum_i^n \frac{(\overset{\text{observed}}{O_i} - E_i)^2}{\underset{\text{expected}}{E_i}}$$

variance:

variance for single variable: $E((X-\mu)^2)$

covariance for two variable:
$E((X_1-\mu_1)(X_2-\mu_2)) = E(X_1X_2) - \mu_1\mu_2 = E[X_1X_2] - E[X_1]E[X_2]$

- the sign of covariance indicate the relation direction

- if `X1` and `X2` are independent, $\sigma_{12} = 0$, but reverse is not true

$$\hat{\sigma}_{11} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i1} - \widehat{\mu_1})(x_{i1} - \widehat{\mu_1})$$

correlation:

if $\rho_{12} > 0$, positive correlation, $\rho_{12} = 0$, uncorrelated, $\rho_{12} < 0$, negative correlated

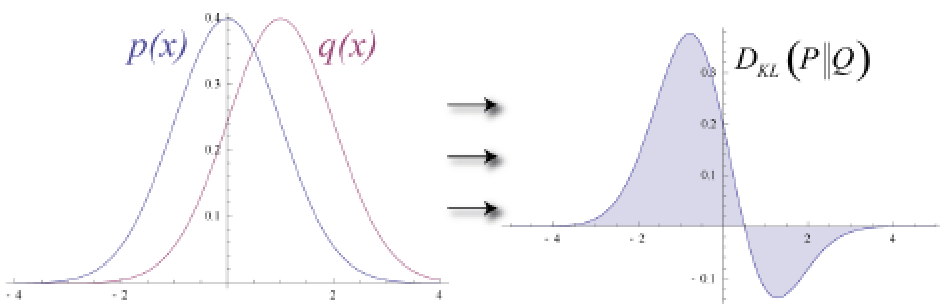
$$\rho_{12} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}$$

Kullback Leibler (KL) divergence:

Measure the difference between two probability distributions over the same variable x

$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x) ln \frac{p(x)}{q(x)}$$

$$D_{KL}(p(x)||q(x)) = \int_{-\infty}^{\infty} p(x) ln \frac{p(x)}{q(x)}$$



Original Gaussian PDF's KL Area to be Integrated

- when $p \neq 0$ but $q = 0$, the D_{KL} is given as ∞, because one predict possible and one predict impossible

Data cleaning

missing data:

1. Incomplete: `Salary = ""`
2. Noisy: `Salary = 10” (an error)`
3. Inconsistent: `Age=“42”, Birthday = “03/07/2022`
4. Intentional: `Jan. 1 as everyone’s birthday?`

data Integration:

Combining data from multiple sources into a coherent store