Data Analysis and Visualisation

A data object typically represents an entity, described by attributes. An attribute is a data field representing a characteristic feature of an object.

**Qualitative Attributes**

Nominal attributes are symbols or names of things. They are also called categorical attributes.

Binary attributes are attributes with only two categories or states: 0 and 1.

A binary attribute is symmetric if both the values 0 and 1 are equally likely and asymmetric otherwise.

An ordinal attribute is one where the attribute’s values have some meaningful ordering among them, but the magnitude between successive values is not known.

**Quantitative Attributes**

A numeric attribute is quantitative. It can be

1. Interval scaled
   1. Measured on a scale of equal sized units
2. Ratio scaled
   1. Defined zero-point

A discrete attribute is one that has a finite or countably infinite set of values, whereas a continuous attributes represents attributes that are not discrete.

Central tendency is measured usually using the mean, median and mode. Mean can be taken as a weighted mean as well, with each data point associated with a weight.

Extreme values may cause the mean to be pulled up or down, causing inconsistencies in the central tendency, so a trimmed mean is used by chopping off values at the low and high extremes.

For skewed symmetric data, the median is a better measure.

The mode is another measure of central tendency. Datasets with one mode are called unimodal, two are called bimodal and three are called trimodal and the rest after 2 are called multimodal.

The midrange can also be used as a measure which is the average values of the highest and lowest value.

When the mode < median, data is positively skewed, and when mode > median, data is negatively skewed.

**Dispersion**

Range = max value - min value

Quantiles are points taken at regular intervals of a data distribution, dividing the data into equal sized consecutive sets. When it is divided into 4 parts, these are called quartiles. When there are a 100 parts, these are called percentiles.

IQR = Q3 - Q1which is the interquartile range.

The five number summary is <Minimum, Q1, Median, Q3, Maximum>. A box plot is used to model this 5 number summary.

Variance and Standard deviation are the most commonly used measures of dispersion. Mathematically, by Chebyshev’s inequality, it can be shown that at least (1 - 1/k2)\*100% observations are no more than k standard deviations from the mean.

**Graphical Descriptions**

1. Quantile plot
   1. Displays all data for a given attribute
   2. Plots quantile information
   3. Calculate fi = (i - 0.5)/N and graph xi vs fi
2. Quantile-Quantile plot
   1. Plot the quantiles of one univariate distribution against another
3. Histograms
   1. Summarize distributions for attribute X
   2. Bar chart if categorical, histogram if numeric
   3. Range of a bucket is known as width
4. Scatter Plot
   1. Trends between attributes
   2. Increasing means positive correlation, decreasing means negative correlations
   3. For n attributes, a scatter-plot matrix is formed

**Data Visualisation**

1. Pixel oriented
   1. Create m windows on the screen, one for each dimension
   2. The m dimension values of a record are mapped to m pixels in the corresponding positions
2. Geometric projection
   1. Scatter plots and scatter plot matrices
   2. Parallel coordinates can be used when dimensionality increases, by having n equally spaced axes, one for each dimension, and a record is a polygonal line intersecting each axis at a point to the dimension value.
3. Icon based
   1. Chernoff faces : cartoon human face used by facial characteristics denoting the values of the dimensions
   2. Stick figures : maps multidimensional data to five piece stick figures.
4. Hierarchical
   1. nVision or Worlds-within-Worlds : 3D plots with inner plots for each subsequent dimension
   2. Treemaps : nested rectangles, each dimension mapped to a unique color
5. Complex
   1. Tag clouds : statistics of user-generated tags

**Measuring Data Similarity and Dissimilarity**

1. A data matrix stores the n data objects in the form of a n x p matrix for p attributes, each row corresponding to an object. Also called two mode matrix.
2. A dissimilarity matrix stores a collection of proximites available for each pair of n objects. Also called one mode matrix.

**Proximity measures**

1. Nominal attributes
   1. d(a,b) = p - m / p where m is the number of matching attributes and p is the total.
   2. Similarity is just 1 - d(a,b)
2. Binary attributes
   1. q is the number of attributes = 1 for both objects
   2. r is the number of attributes which are 1 for object i but 0 for j
   3. s is the number of attributes which are 0 for object i but 1 for j
   4. t is the number of attributes = 0 for both objects
   5. p is the total number of attributes
   6. For symmetric binary dissimilarity where each state is equally valuable,

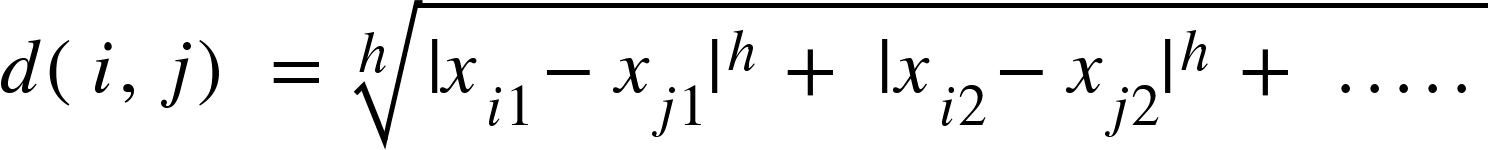
d(i, j) = (r + s) / (q + r + s + t)

* 1. For asymmetric, we just ignore t

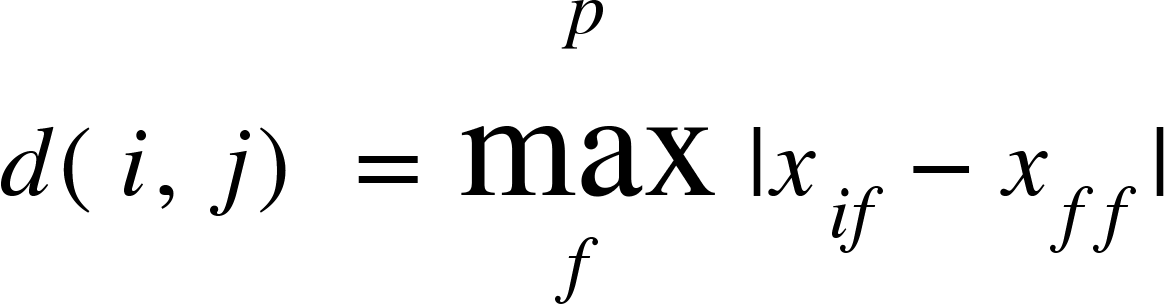
d(i, j) = (r + s) / (q + r + s)

* 1. The coefficient sim(i, j) is the Jaccard coefficient

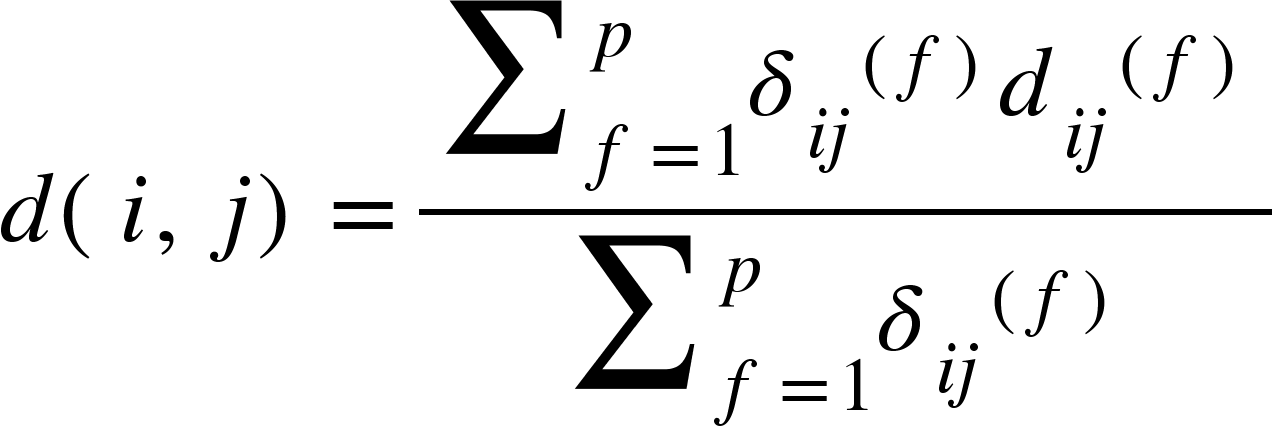
1. Numeric attributes
   1. Euclidean Distance
   2. Manhattan Distance
   3. Minkowski Distance



* 1. Supremum Distance



1. Ordinal values
   1. Replace each x with its rank in the order
   2. Normalise
   3. Calculate dissimilarity using any distance measure
2. Mixed attributes
   1. Clustering is one approach
   2. Other approach



where del(i, j) is 0 if either

1. xif or xjf is missing
2. Xif = xjf = 0

Else use the same distances as shown above, but normalise numeric attributes.

**Cosine Similarity**

This is usually used for term frequency in a document. These vectors are very long and sparse.

sim(**x, y**) = **(x.y)/||x||||y||**

When attributes are binary valued

sim(x, y) = (**x.y**)/(**x.x + y.y - x.y’**)

And this is called the Tanimoto coefficient.