Chapter II: Data

Knowledge Discovery in Databases

Luciano Melodia M.A.
Evolutionary Data Management, Friedrich-Alexander University Erlangen-Nürnberg
Summer semester 2021

Chapter II: Getting to know your data

This is our agenda for this lecture:

- **Data objects and attribute types.**
- Basic statistical descriptions of data.
- Data visualization.
- Measuring data similarity and dissimilarity.
- Summary.

Types of data sets

Records:

Relational records.
Data matrix, e.g. numerical matrix, crosstabs.
Document data: text documents, **term-frequency vectors**.
Transaction data.

Graph and network:

World wide web.
Social of information networks.
Molecular structures.

	team	couch	play	ball	score	game
Document1	3	0	5	0	2	6
Document2	0	7	0	2	1	0
Document3	0	1	0	0	1	2

TID	**Items**
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diapers, Milk
4	Beer, Bread, Diapers, Milk
5	Coke, Diapers, Milk

Types of data sets

Ordered data:

- Video data: sequences of images.
- Temporal data: time series.
- Sequential data: transaction sequences.
- Genetic sequence data.

Spatial, image and multimedia:

- Spatial data: maps.
- Image data.
- Video data.

Important characteristics of structured data

Dimensionality:
Curse of dimensionality (sparse high-dimensional data spaces).

Sparsity:
Only presence counts.

Resolution:
Patterns depend on the scale.

Distribution:
Centrality and dispersion.

Data objects

Data sets are made up of data objects.
A data object represents an entity.

Examples:

- Sales database: customers, store items, sales.
- Medical database: patients, treatments.
- University database: students, professors, courses.

They are also called:
Sampels, examples, instances, data points, objects, tuples, . . .

Data objects are described by attributes:

- Database rows $\longrightarrow$ data objects.
- Columns $\longrightarrow$ attributes.

Attributes

Attribute:
Sometimes also in other context: field, dimension, feature, variable, . . .

A data field encodes the property of an entity or feature of a data object.
E.g. `customer_ID`, `name`, `address`.

Types:

- Nominal.
- Binary.
- Ordinal.
- Numerical:
  - Interval scaled.
  - Ratio scaled.

Attribute types

Nominal:

Categories, states, or "names of things".
E.g. $\texttt{hair\_color} = \{\text{auburn, black, blond, brown, grey, red, white}\}$.
Other examples: `marital_status`, `occupation`, `ID`, `ZIP code`.

Binary:

Nominal attribute with only two states (0 and 1).
Symmetric binaries: both outcomes equally important, such as `gender`.
Asymmetric binary: outcomes not equally important.
E.g. medical test (positive vs. negative).
Convention: assign 1 to most important outcome (e.g. HIV positive).

Ordinal:

Values have a meaningful order (ranking),
but magnitude between successive values is not known.
E.g. $\texttt{size} = \{\text{small, medium, large}\}$, grades, army rankings.

Numerical attribute types

Numerical: Quantity (integer- or real-valued).

Interval scaled:

Measured on a scale of **equally sized** units.
Values have order.
E.g. temperature in C or F, calender dates.
No true zero-point.

Ratio scaled:

Inherent **zero point**.
We can speak of values as being an order of magnitude larger than the unit of measurement.
E.g. $10K$ is twice as high as $5K$.
E.g. temperature in Kelvin, length, counts, monetary quantities.

Discrete vs. continuous attributes

Discrete attribute:

Has finite or countably infinite elements.
E.g. ZIP code, profession, or the set of words in a collection of documents.
Sometimes represented as integer variables.
Note: Binary attributes are a special case of discrete attributes.

Continuous attribute:

Has real numbers as attribute values.
E.g. temperature, height, or weight.
Practically, real values can only be measured and represented using a finite number of digits.
Continuous attributes are typically represented as floating-point variables.

Chapter II: Getting to know your data

Data objects and attribute types.

Basic statistical descriptions of data.

Data visualization.

Measuring data similarity and dissimilarity.

Summary.

Basic statistical descriptions of data

Motivation:

To better understand the data: central tendency, variation and spread.

Data dispersion characteristics:

Median, max, min, quantiles, outliers, variance etc.

Numerical dimensions correspond to sorted intervals.

Data dispersion: analyzed with multiple granularities of precision.

Boxplot or quantile analysis on sorted intervals

Dispersion analysis on computed measures.

Folding measures into numerical dimensions.

Boxplot or quantile analysis on the transformed cube.

Measuring the central tendency

Mean:

N denotes the amount of samples within the data set.
The **sample mean** is given by

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i .$$

While the **population mean** is defined by

$$\mu = \sum x \cdot p(x|\theta) \cdots .$$

Measuring the central tendency (2)

Median:

The median $\tilde{x}$ minimizes the sum of absolute deviations for any x of a sample X:

$$\sum_{i=1}^{n} |\tilde{x} - x_i| \leq \sum_{i=1}^{n} |x - x_i|. \tag{1}$$

$$\tilde{x} = \begin{cases} x_{\frac{N}{2}} & \text{if } N \bmod 2 = 0, \\ \frac{x_{\frac{N-1}{2}} + x_{\frac{N+1}{2}}}{2} & \text{if } N \bmod 2 \neq 0. \end{cases} \tag{2}$$

Age	Frequency
$1-5$	200
$6-15$	450
$16-20$	300
$21-50$	1500
$51-80$	700
$81-110$	44

Measuring the central tendency (3)

Median for interval grouped data:

Let n be the total amount of data points, n_i the respective number of the ith group and l_i or u_i the lower or upper interval limit. We determine the group to which the median belongs and denote it as mth group. It is determined by

$$\sum_{k=1}^{m-1} n_k < \frac{n}{2}, \text{ but } \sum_{k=1}^{m} n_k \geq \frac{n}{2}. \tag{3}$$

If there is no information about the underlying distribution, we just assume that data is equally distributed and use linear interpolation to estimate the median:

$$\tilde{x} = l_m + \frac{\frac{n}{2} - \sum_{k=1}^{m-1} n_k}{n_m} \cdot (u_m - l_m). \tag{4}$$

Age	Frequency
$1 - 5$	200
$6 - 15$	450
$16 - 20$	300
$21 - 50$	1500
$51 - 80$	700
$81 - 110$	44

Measuring the central tendency (3)

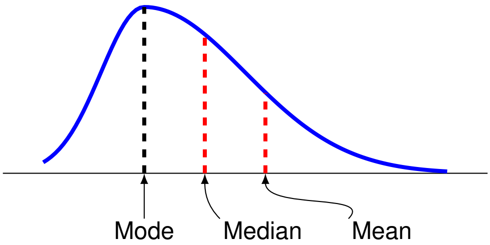
Mode:

Value that occurs most frequently within the data set.
Can be unimodal, bimodal, trimodal etc.
Empirical formula:

$$\overline{x} - \text{mode} \approx 3(\overline{x} - \tilde{x}). \tag{5}$$

Age	Frequency
1 − 5	200
6 − 15	450
16 − 20	300
21 − 50	1500
51 − 80	700
81 − 110	44

Example of mode, median and mean



$$f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{6}$$

Example of mode, median and mean

Quartiles, outliers and boxplots:

Quartiles: $\mathbf{Q}_1$ (25^{th} percentile), $\mathbf{Q}_3$ (75^{th} percentile).

Inter quartile range: $\mathrm{IQR} = Q_3 - Q_1$.

Five number summary: min, Q_1, median, Q_3, max.

Boxplot: ends of the box are the quartiles;
median is marked; add whiskers and plot outliers individually.

Outlier: usually assigned to values higher/lower than $1.5 \cdot \mathrm{IQR}$.

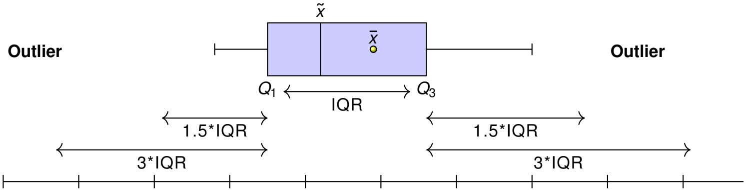
Variance σ^2 and standard deviation σ:

Empirical sample variance: $\overline{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$

Empirical population variance: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2$.

Standard deviation is the square root $\sigma = \sqrt{\sigma^2}$.

Boxplot analysis



Five number summary of a distribution:
Minimum, Q_1, median, Q_3, maximum.

Boxplot:

Data is represented with a box.

The ends of the box are at the first and third quartiles, i.e. the height of the box is IQR.

The median is marked by a line within the box.

Whiskers: two lines outside the box extended to minimum and maximum.

Outliers: points beyond a specified outlier threshold, plotted individually.

Properties of normal distribution curves

The normal distribution:

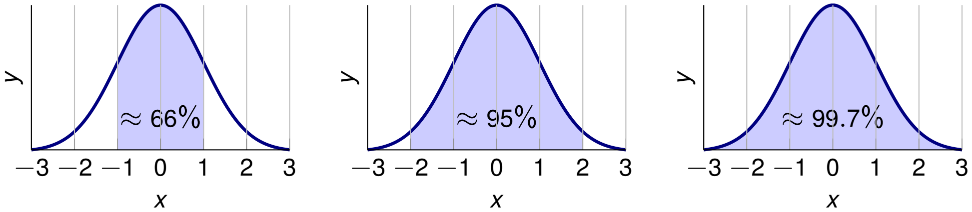
From $\mu - \sigma$ to $\mu + \sigma$: contains about 68% of the measurements.

μ: mean,

σ: standard deviation.

From $\mu - 2\sigma$ to $\mu + 2\sigma$: contains about 95% of the surface under the curve.

$\mu - 3\sigma$ to $\mu + 3\sigma$: contains about 99.7% of the surface under the curve.



Visualization of basic statistical descriptions

Boxplot: Visualization of five number summary.

Histogram: x-axis are values, y-axis represent frequencies.

Quantile plot: Each value x_i is paired with some q_i indicating that approximately $q_i \cdot 100\%$ of data are $\leq x_i$.

Quantile-quantile (q-q) plot: Graphs the quantiles of one univariate distribution against the corresponding quantiles of another.

Scatter plot: Each pair of values is a pair of coordinates and plotted as points in the plane.

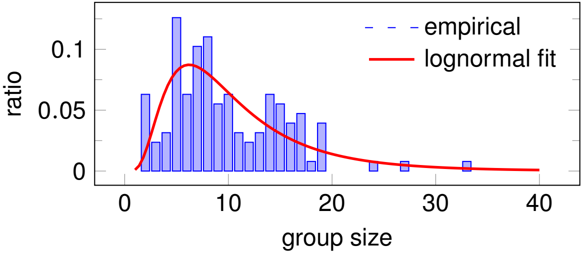
Histogram analysis

Histogram: Visualization of tabulated frequencies, shown as bars.

It shows what proportion of cases fall into each of several categories.

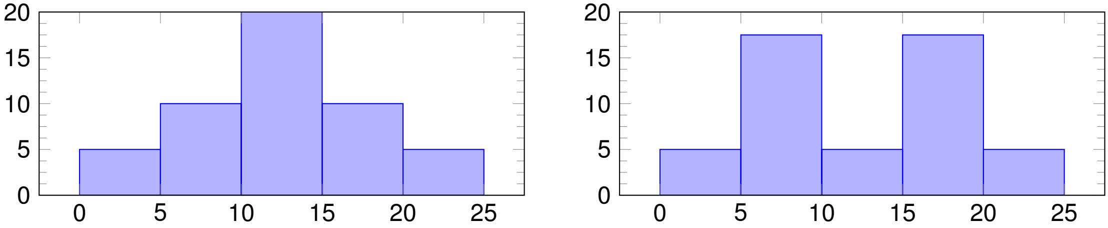
Differs from a **bar chart** in that it is the *area* of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width.

The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent.



Histograms often tell more than boxplots

The two histograms shown below may have the same boxplot representation, thus the same values for min, Q_1, median, Q_3 and for the max. But they have rather different underlying distributions.



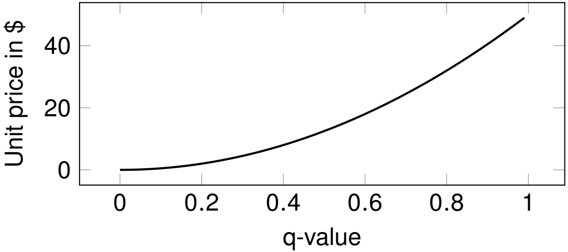
Quantile plot

Displays all of the data.

A quantile plot allows the user to assess both the overall behaviour and unusual occurrences.

Plots quantile information.

For some data point x_i, sorted in increasing order, q_i indicates that approximately $q_i \cdot 100\%$ of the data are below or equal to the value of x_i.

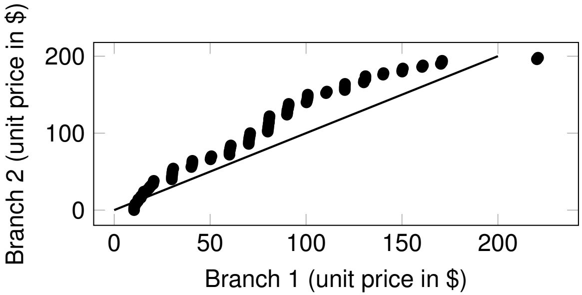


Quantile-quantile (q-q) plot

Graphs the quantiles of one univariate distribution against the corresponding quantiles of another.

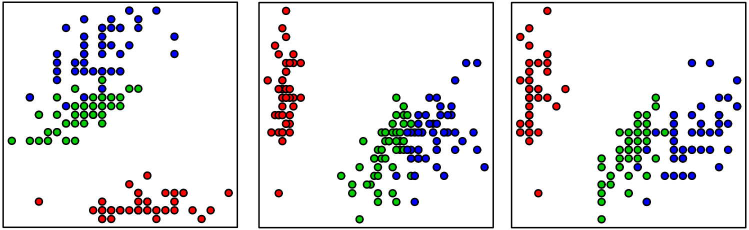
View: Do these two distributions differ?

Example shows unit price of items sold at Branch 1 vs. branch 2 for each quantile. Unit prices of items sold at branch 1 tend to be lower than those at branch 2.



Scatter plots

Provides a first look at **bivariate data** to see clusters of points, outliers or similar.
Each pair of values is treated as a pair of coordinates and plotted as points in the plane.



Data profiling

More from the database perspective.

Derive metadata such as:

- Data types and value patterns.
- Completeness and uniqueness of columns.
- Keys and foreign keys.
- Occasionally functional dependencies and association rules.
- Discovery of inclusion dependencies and conditional functional dependencies.

Statistics:

- Number of null values and distinct values in a column.
- Data types.
- Most frequent patterns of values.

Chapter II: Getting to know your data

Data objects and attribute types.

Basic statistical descriptions of data.

Data visualization.

Measuring data similarity and dissimilarity.

Summary.

Data visualization

Why visualize data?

- **Gain insight** into an information space by mapping data into graphical primitives.
- **Provide qualitative overview** of large data sets.
- **Search** for patterns, trends, structure, irregularities, relationships among data.
- **Help find interesting regions and suitable parameters** for further quantitative analysis.
- **Provide a visual proof** of computer representations derived.

Categorization of visualization methods:

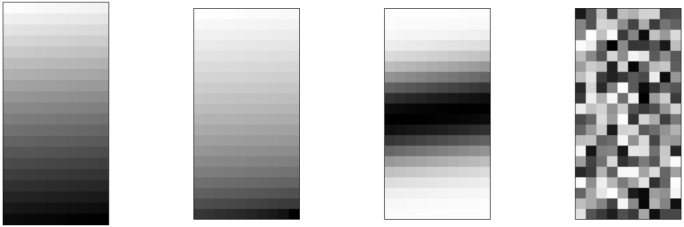
- Pixel-oriented.
- Geometric projection.
- Icon-based.
- Hierarchical.
- Visualizing complex data and relations.

Pixel oriented visualization techniques

For a data set of *m* dimensions create *m* windows on the screen, one for each dimension.

The values in dimension *m* of a record are mapped to *m* pixels at the corresponding positions in the windows.

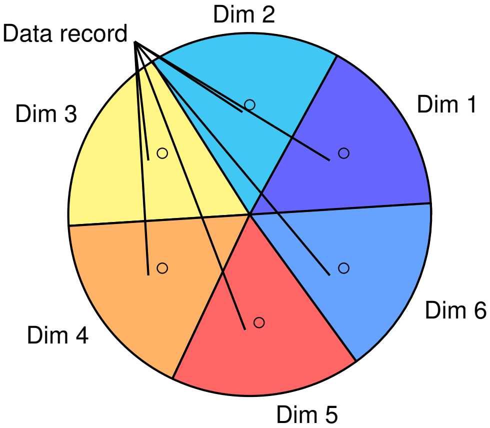
The colors of the pixels reflect the corresponding values.



a) Income. b) Credit limit. c) Transaction volume. (d) Age.

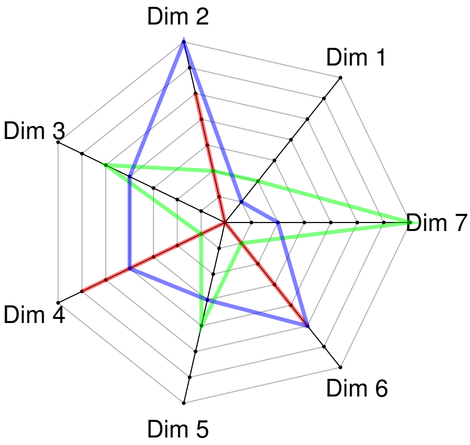
Laying out pixels in a spiderweb diagram

To save space and show the connections among multiple dimensions, space filling is often done in a spiderweb diagram.



Laying out pixels in circle segments

To save space and show the connections among multiple dimensions, space filling is often done in a circle segment.



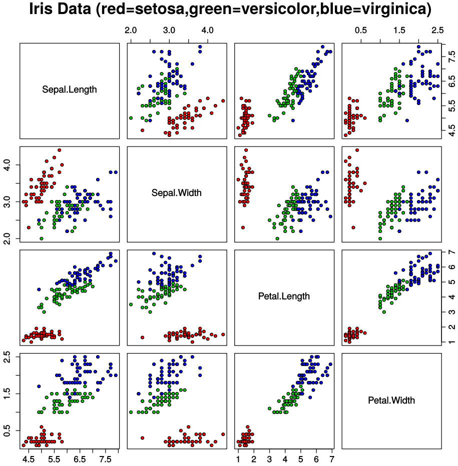
Geometric projection visualization techniques

Visualization of geometric transformations and projections of data.

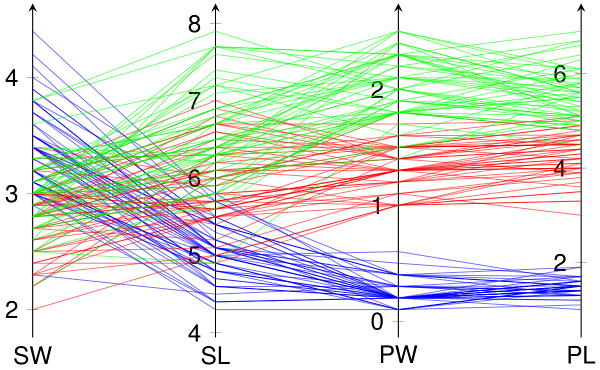
Methods:

- Scatter plot and scatter plot matrices.
- Landscapes.
- Projection pursuit technique: *Help users find meaningful projections of multidimensional data.*
- Prosection views.
- Hyperslice.
- Parallel coordinates.

Scatter plot matrices



Parallel coordinate plot



Icon based visualization

Visualization of the data values as features of icons.

Typical visualization methods:

- Chernoff faces.
- Stick figures.

General techniques:

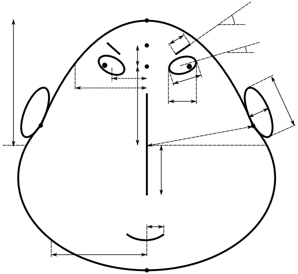
- Shape coding: *Use shape to represent certain information encoding.*
- Color icons: *Use color icons to encode more information.*
- Tile bars: *Use small icons to represent the relevant feature vectors in document retrieval.*

Chernoff faces

A way to display variables on a two-dimensional surface:

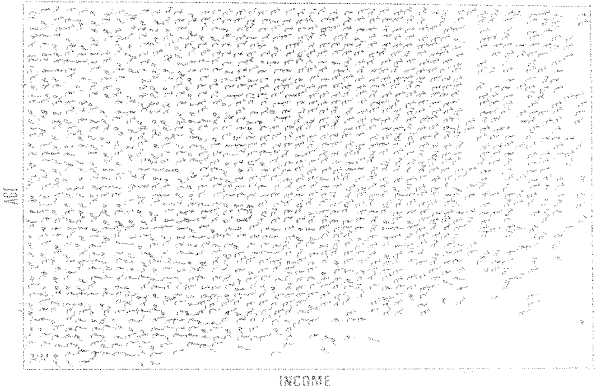
E.g. let *x* be eyebrow slant, *y* be eye size, *z* be nose length etc.

The figure shows faces produced using 10 characteristics (head eccentricity, eye size, eye spacing, eye eccentricity, pupil size, eyebrow slant, nose size, mouth shape, mouth size, and mouth opening). Each assigned one of 10 possible values, generated using Mathematica (S. Dickson).



Stick figure

A census data figure showing age, income, gender, education etc.
A 5-piece stick figure (1 body and 4 limbs w. different angle/length).



Used by permission of G. Grinstein, University of Massachusettes at Lowell.

Hierarchical visualization techniques

Visualization of the data using a hierarchical partitioning into subspaces.

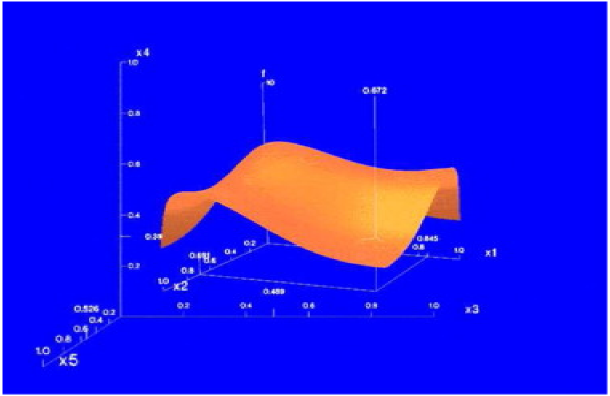
Methods:

- Worlds within worlds.
- Tree maps.
- Cone trees.
- Info cube.

Worlds within world

Assign the function and two most important parameters to innermost world.

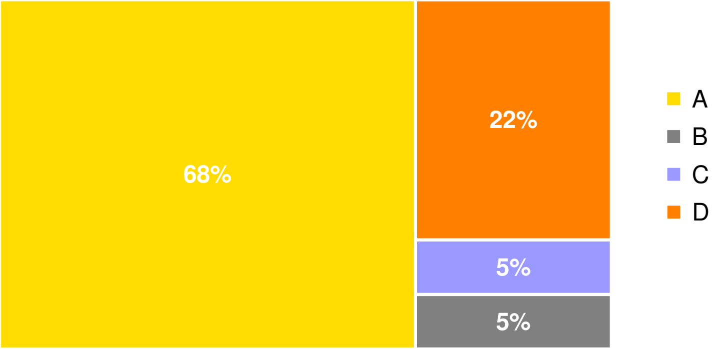
Fix all other parameters at constant values – draw other (1,2 or 3) dimensional worlds choosing these as the axes.



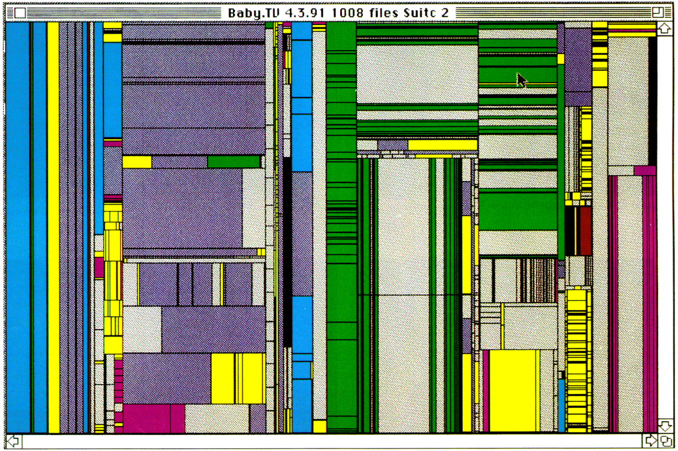
Tree maps

Screen filling method:

Uses a hierarchical partitioning of the screen into regions depending on the attribute values.
x and *y*-coordinates of the screen partitioned alternately according to the attribute values.



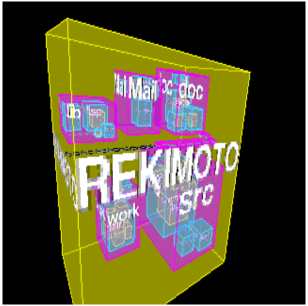
Tree map of a file system



Info cubes

3D visualization technique:

- Hierarchical information is displayed as nested semi-transparent cubes.
- The outermost cubes correspond to the top level data, while the subnodes or the lower level data are represented as smaller cubes inside the outermost cubes, and so on.



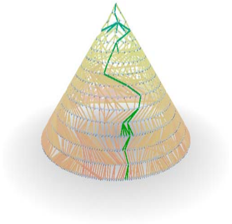
Three dimensional cone trees

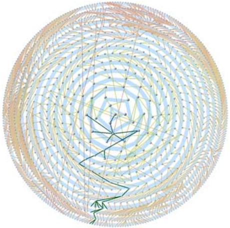
3D cone-tree visualization technique:
works well for up to approx. a thousand nodes.

Build a 2D circle tree that arranges its nodes in concentric circles centered on the root node.

Overlaps can't be avoided projecting onto 2D.

G. Robertson, J. Mackinlay, S. Card. "Cone Trees: Animated 3D Visualizations of Hierarchical Information", ACM SIGCHI'91.





Acknowledgement: http://nadeausoftware.com/articles/visualization.

Visualizing complex data and relations

Visualizing non-numerical data: text and social networks.

Tag cloud: visualizing user-generated tags.

The importance of tag is represented by font size/color.
Besides text data, there are also methods to visualize relationships, such as visualizing social networks.

Chapter 2: Getting to know your data

Data objects and attribute types.

Basic statistical descriptions of data.

Data visualization.

Measuring data similarity and dissimilarity.

Summary.

Similarity and dissimilarity

Similarity.

Numerical measure of how alike two data objects are.
Value is higher when objects are more alike.
Often chosen within the range of $[0, 1]$.

Dissimilarity.

E.g. distance.
Numerical measure of how different two data objects are.
Lower when objects are more alike.
Minimum dissimilarity is often 0.
Upper limit varies.

Proximity.

Refers to similarity or dissimilarity.

Data matrices and dissimilarity matrices

Data matrix:

n data points with dimension m. Two mode matrix.

$$\begin{Bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{Bmatrix}.$$

Dissimilarity matrix:

n data points, but registers only the distance. A triangular one mode matrix.

$$\begin{Bmatrix} 0 & 0 & 0 & \cdots & 0 \\ d(x_1, x_2) & 0 & 0 & \cdots & 0 \\ d(x_1, x_3) & d(x_2, x_3) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d(x_1, x_m) & d(x_2, x_m) & d(x_3, x_m) & \cdots & 0 \end{Bmatrix}.$$

Proximity measures for nominal attributes

Can take two or more states.

- E.g. red, yellow, blue or green.
- Generalization of a binary attribute.

Values can be the same (distance of 0) or different (distance of 1).

More options for sets of nominal attributes (variables).

(1) Method is the simple matching coefficient:

$$\mathrm{SMC} = \frac{\#\text{of matching attributes}}{\#\text{number of attributes}} = \frac{\sum_{i=1}^{n} a_{ii}}{\sum_{i,j=1}^{n} a_{ij}}. \tag{7}$$

(2) Method is to use a large number of binary attributes:

- Creating a new binary attribute for each of the nominal states.

Proximity measure for binary attributes (I)

A contingency table for binary data.

Counting matches.

$$
y \left\{ \begin{array}{c|ccc}
 & 0 & 1 & \sum \\
\hline
0 & q & r & q+r \\
1 & s & t & s+t \\
\sum & q+s & r+t & q+r+s+t
\end{array} \right.
$$

$$\underbrace{\hphantom{\sum \quad q+s \quad r+t \quad q+r+s+t}}_{x}$$

Distance measure for symmetrical binary variables:

$$d(x, y) = \frac{r+s}{q+r+s+t}. \tag{8}$$

Distance measure for asymmetrical binary variables:

$$d(x, y) = \frac{r+s}{q+r+s}. \tag{9}$$

Proximity measure for binary attributes (II)

A contingency table for binary data.

Counting matches.

$$
y\left\{
\begin{array}{c|ccc}
 & 0 & 1 & \sum \\
\hline
0 & t & s & t+s \\
1 & r & q & r+q \\
\sum & t+s & s+q & q+r+s+t
\end{array}
\right.
$$

$$\underbrace{\hphantom{xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx}}_{x}$$

Jaccard coefficient for asymmetrical binary variables:

$$\text{Jaccard}(x, y) = \frac{q}{q + r + s}. \tag{10}$$

Jaccard coefficient corresponds to "coherence":

$$d(x, y) = \frac{\sup(x, y)}{\sup(x) + \sup(y) - \sup(x, y)} = \frac{q}{(s + q) + (r + q) - q}. \tag{11}$$

Dissimilarity between binary variables

Example:

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Bob	M	Y	N	P	N	N	N
Alice	F	Y	N	P	N	P	N
Charlie	M	Y	P	N	N	N	N

Gender is a symmetrical attribute.

The remaining attributes are asymmetrical binary.

Let the values Y and P be equal to 1 and the value of N be 0, then

$$\mathrm{Jaccard}(\mathrm{Bob}, \mathrm{Alice}) = \frac{0+1}{2+0+1} \approx 0.33, \tag{12}$$

$$\mathrm{Jaccard}(\mathrm{Bob}, \mathrm{Charlie}) = \frac{1+1}{1+1+1} \approx 0.67, \tag{13}$$

$$\mathrm{Jaccard}(\mathrm{Charlie}, \mathrm{Alice}) = \frac{1+2}{1+1+2} = 0.75. \tag{14}$$

Standardizing numerical data

z-Score:

$$z = \frac{x - \mu}{\sigma}. \tag{15}$$

x is the score to be standardized; μ ist the population mean; σ is the standard deviation.
The distance between the raw score and the population mean in units of the standard deviation.
Negative when the raw score is below the mean, positive else.
An alternative way is to compute the average absolute deviation:

$$\mathrm{MAD}\big(X = \{x_1, x_2, \ldots, x_n\}\big) = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|, \tag{16}$$

$$\text{where } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \text{ thus } z_i = \frac{x_i - \bar{x}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}}. \tag{17}$$

Example: data matrix and dissimilarity matrix

Data matrix:

Point	Attribute 1	Attribute 2
x_1	1	2
x_2	3	5
x_3	2	0
x_4	4	5

Dissimilarity matrix (with Euclidean distance):

	x_1	x_2	x_3	x_4
x_1	0			
x_2	$3,61$	0		
x_3	$2,24$	$5,1$	0	
x_4	$4,24$	1	$5,39$	0

Distance on numerical data: Minkowski distance

Minkowski distance: a popular distance measure, given by:

$$d(x, y) = \sqrt[n]{\sum_{i=1}^{n} |x_i - y_i|^n}, \tag{18}$$

where $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$
are two n-dimensional data objects and n if the order.

In fact, this distance induces a norm over real vector space, called L_n-norm.

Properties:

$d(x, y) \geq 0$, positive definiteness.
$d(x, y) = d(y, x)$, symmetry.
$d(x, y) \leq d(x, z) + d(z, y)$, triangle inequality.

A distance satisfying this properties is called **metric**.

Special cases of Minkowski distance

$n = 1$: **Manhatten** (city block, L_1-norm) distance:

E.g. the Hamming distance: the number of bits that differ in two binary vectors, given by

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|. \tag{19}$$

$n = 2$: **Euclidean** (L_2-norm) distance:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \tag{20}$$

$n \to \infty$: **supremum** ($L_{\max}$-norm, L_∞-norm) distance:

This is the maximum difference between any component (attribute) of the vectors.

$$d(x, y) = \lim_{n\to\infty} \left(\sum_{i=1}^{n} |x_i - y_i|^n \right)^{\frac{1}{n}} = \max_i |x_i - y_i|. \tag{21}$$

Example: Minkowski, Euclidean and Supremum distance

Point	Attribute 1	Attribute 2
x_1	1	2
x_2	3	5
x_3	2	0
x_4	4	5

Manhatten (L_1):

	x_1	x_2	x_3	x_4
x_1	0			
x_2	5	0		
x_3	3	6	0	
x_4	6	1	7	0

Euclidean (L_2):

	x_1	x_2	x_3	x_4
x_1	0			
x_2	$3,61$	0		
x_3	$2,24$	$5,1$	0	
x_4	$4,24$	1	$5,39$	0

Supremum (L_∞):

	x_1	x_2	x_3	x_4
x_1	0			
x_2	3	0		
x_3	2	5	0	
x_4	3	1	5	0

Ordinal variables

- An ordinal variable can be discrete or continuous.
- Order is important e.g. rank.
- Can be treated like interval-scaled:
  - Replace x_j by their rank $r_j \in \{1, \ldots, N\}$.
  - Map the range of each variable onto $[0, 1]$ by replacing the i-th position of an object by

$$z_i = \frac{r_i - 1}{N - 1}. \tag{22}$$

  - Compute the dissimilarity using methods for interval-scaled variables.

Attributes of mixed type

A database may contain all attribute types:

Nominal, symmetric binary, asymmetric binary, numerical, ordinal.

One can use a weighted formula to combine their effects:

$$d(x, y) = \frac{\sum_{i=1}^{n} w_i d(x_i, y_i)}{\sum_{i=1}^{n} w_i}. \tag{23}$$

If x_i, y_i are binary or nominal, then

$$d(x_i, y_i) = 0 \text{ if } x_i = y_i \text{ or } d(x_i, y_i) = 1 \text{ otherwise.} \tag{24}$$

If x_i, y_i are numeric, we use the normalized distance.

If x_i, y_i are ordinal, we compute their ranks $r_i^{(x)}, r_i^{(y)}$, compute z_i and treat z_i as interval-scaled.

Cosine similarity

A document can be represented by thousands of attributes, each recording the frequency of a particular word (such as keywords) or phrase in the document.

Document	Team	Coach	Hockey	Baseball	Soccer	Penalty	Score	Win	Loss
Document1	5	0	3	0	2	0	0	2	0
Document2	3	0	2	0	1	1	0	1	0
Document3	0	7	0	2	1	0	0	3	0
Document4	0	1	0	0	1	2	2	0	3

Other vector objects: gene features in microarrays and more.

Applications: information retrieval, biologic taxonomy, gene-feature mapping and many more.

Cosine measure: Let x and y be two vectors (e.g. term-frequency vectors), then

$$\text{sim}(x, y) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}. \tag{25}$$

Example: cosine similarity (I)

$$\text{sim}(x, y) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}. \tag{26}$$

Compute the similarity between x and y for $x = (5, 0, 3, 0, 2, 0, 0, 2, 0)$, $y = (3, 0, 2, 0, 1, 1, 0, 1, 0)$.

$$\sum_{i=1}^{n} x_i \cdot y_i = 5 \cdot 3 + 0 \cdot 0 + 3 \cdot 2 + 0 \cdot 0 + 2 \cdot 1 + 0 \cdot 1 + 0 \cdot 0 + 2 \cdot 1 + 0 \cdot 0 = 25, \tag{27}$$

$$\sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{5^2 + 0^2 + 3^2 + 0^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2} \approx 6,48, \tag{28}$$

$$\sqrt{\sum_{i=1}^{n} y_i^2} = \sqrt{3^2 + 0^2 + 2^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2} = 4. \tag{29}$$

(30)

Example: cosine similarity (II)

$$\text{sim}(x, y) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}. \tag{31}$$

Compute the similarity between x and y for $x = (5, 0, 3, 0, 2, 0, 0, 2, 0)$, $y = (3, 0, 2, 0, 1, 1, 0, 1, 0)$.

$$\text{sim}(x, y) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}} \approx \frac{25}{6,48 \cdot 4} \approx 0.96. \tag{32}$$

Chapter 2: Getting to know your data

Data objects and attribute types.

Basic statistical descriptions of data.

Data visualization.

Measuring data similarity and dissimilarity.

Summary.

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

Summary

Data attribute types:
Nominal, binary, ordinal, interval-scaled or ratio-scaled.

Many types of data sets:
E.g. numerical, text, graph, web, image.

Gain insight into the data by:

- Basic statistical data description: *Central tendency, dispersion and graphical display.*
- Data visualization: *Map data onto graphical primitives.*
- Measure data similarity.

Above steps are the beginning of data preprocessing.

Many methods have been developed but still an active area of research.

References I

W. Cleveland: Visualizing Data, Hobart Press, 1993.

T. Dasu and T. Johnson: Exploratory Data Mining and Data Cleaning, John Wiley, 2003.

U. Fayyad, G. Grinstein, and A. Wierse: Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001.

L. Kaufman and P. J. Rousseeuw: Finding Groups in Data: an Introduction to Cluster Analysis, John Wiley & Sons, 1990.

H. V. Jagadish et al.: Special Issue on Data Reduction Techniques, Bulletin of the Tech. Committee on Data Eng., 20(4), 1997.

D. Keim: Information visualization and visual data mining, IEEE Trans. on Visualization and Computer Graphics, 8(1), 2002.

F. Naumann: Data Profiling Revisited, ACM SIGMOD Record, 32(4), 2013, 40-49.

D. Pyle: Data Preparation for Data Mining, Morgan Kaufmann, 1999.

References II

S. Santini and R. Jain: Similarity measures, IEEE Trans. on Pattern Analysis and Machine Intelligence, 21(9), 1999.

E. R. Tufte: The Visual Display of Quantitative Information, 2nd ed., Graphics Press, 2001.

C. Yu et al.: Visual data mining of multimedia data for social and behavioral studies, Information Visualization, 8(1), 2009.

Thank you for your attention.

Any questions about the second chapter?

Ask them now, or again, drop me a line:

luciano.melodia@fau.de.