DATA MINING 1

Data Understanding

Dino Pedreschi, Riccardo Guidotti

Revisited slides from Lecture Notes for Chapter 2 "Introduction to Data Mining", 2nd Edition by Tan, Steinbach, Karpatne, Kumar

UNIVERSITÀ DI PISA

Getting To Know Your Data

- For preparing data for data mining task it is essential to have an overall picture of your data
- Gain insight in your data
  - with respect to your project goals
  - and general to understand properties
- Find answers to the questions
  - What kind of attributes do we have?
  - How is the data quality?
  - Does a visualization helps?
  - Are attributes correlated?
  - What about outliers?
  - How are missing values handled?
  - Do we need to extract other attributes

Types of data sets

- Record
  - Data Matrix
  - Document Data
  - Transaction Data
- Graph
  - World Wide Web
  - Molecular Structures
- Ordered
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data

What is Data?

- Collection of **data objects** and their **attributes**
- An **attribute** is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, dimension, or feature
- A collection of attributes describe an **object**
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

- Such data set can be represented by an *m* by *n* matrix, where there are *m* rows, one for each object, and *n* columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Document Data

- Each document becomes a ‘term’ vector
  - Each term is a component (attribute) of the vector
  - The value of each component is the number of times the corresponding term occurs in the document.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

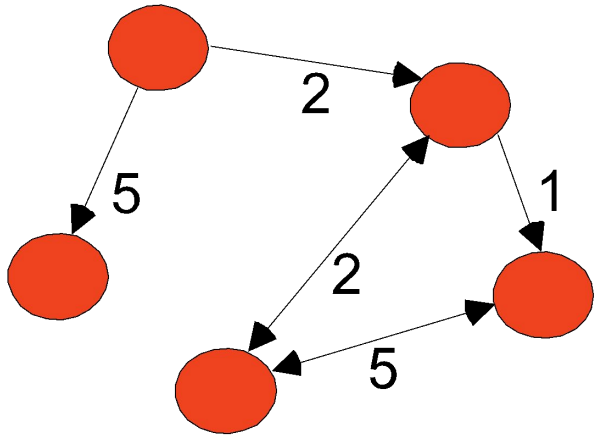
Transaction Data

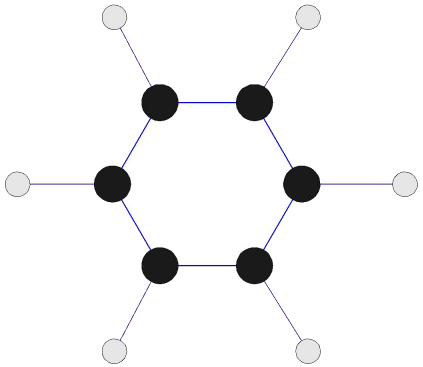
- A special type of record data, where
  - Each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

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

Graph Data

- Examples: Generic graph, a molecule, and webpages





Benzene Molecule: C6H6

Useful Links:

- Bibliography
- Other Useful Web sites
  - ACM SIGKDD
  - KDnuggets
  - The Data Mine

Knowledge Discovery and Data Mining Bibliography

(Gets updated frequently, so visit often!)

- Books
- General Data Mining

Book References in Data Mining and Knowledge Discovery

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.
Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.

General Data Mining

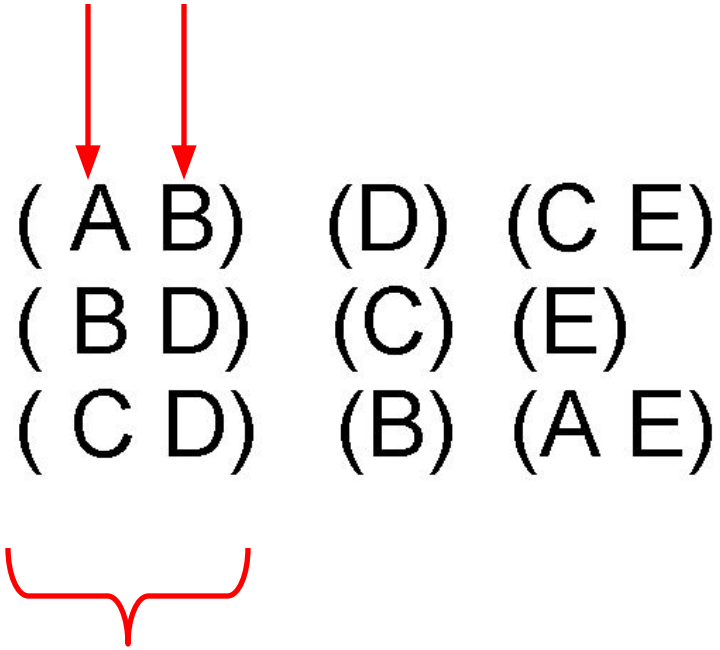
Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

Ordered Data

- Sequences of transactions

Items/Events



An element of the sequence

Retail data

Ordered Data

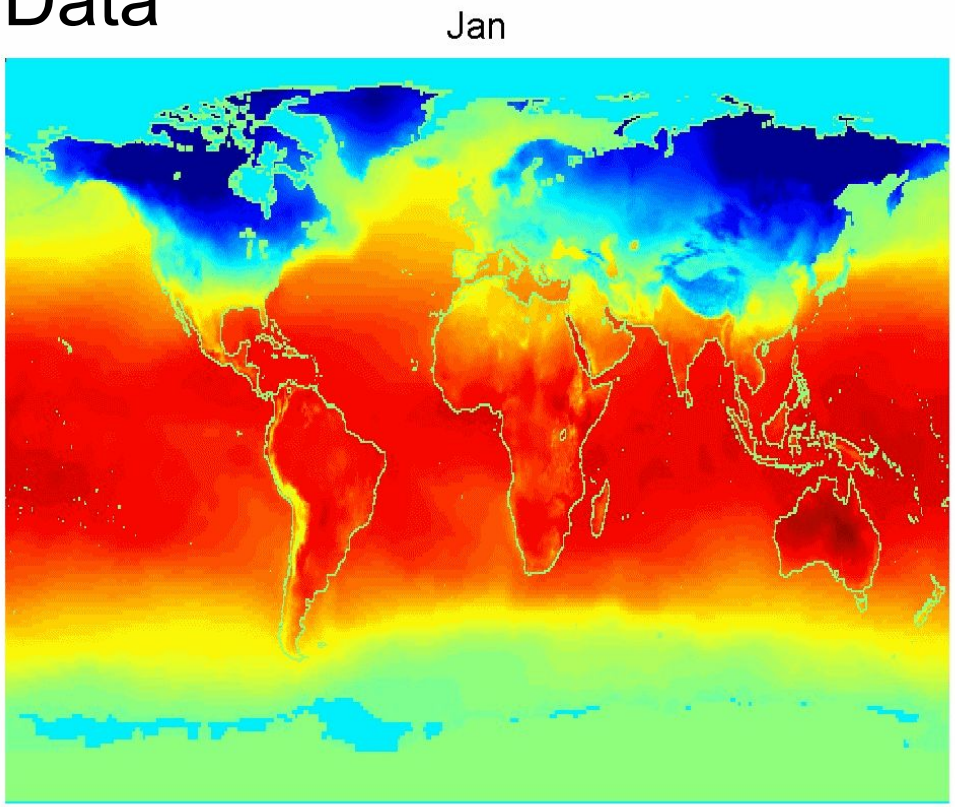
- Genomic sequence data

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Ordered Data

- Spatio-Temporal Data

Average Monthly Temperature of land and ocean



Types of Attributes

- Five are different types of attributes
  - Nominal/Categorical: attribute values in a finite domain, categories
    - **Examples**: ID numbers, eye color, zip codes
  - Binary: Nominal attribute with only 2 states (0 and 1)
    - **Symmetric binary**: both outcomes equally important (e.g., gender)
    - **Asymmetric binary:** outcomes not equally important. (e.g., medical test positive vs. negative) The convention is to assign 1 to most important outcome (e.g., having cancer)
  - Ordinal: finite domain with a meaningful ordering on the domain
    - **Examples**: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}

Types of Attributes

- Numeric: quantity (integer or real-valued)
  - **Interval-Scaled**
    - Measured on a scale of equal-sized units
    - Values have order
  - **Examples**: calendar dates, temperatures in Celsius
- Ratio-Scaled: We can speak of values as being an order of magnitude larger than the unit of measurement
  - **Examples**: length, counts, elapsed time (e.g., time to run a race)
  - A baseball game lasting 3 hours is **50% longer** than a game lasting **2 hours.**

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - **Examples**: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes
- Continuous Attribute
  - **Has real numbers** as attribute values
  - **Examples**: 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.

Properties of Attribute Values

The type of an attribute depends on which of the following properties/operations it possesses:

- Distinctness: = ≠
- Order: < >
- Differences are meaningful : + -
- Ratios are meaningful * /

- Nominal attribute: distinctness
- Ordinal attribute: distinctness & order
- Interval attribute: distinctness, order & meaningful differences
- Ratio attribute: all 4 properties/operations

	Attribute Type	Description	Examples	Operations
Categorical Qualitative	Nominal	Nominal attribute values only distinguish. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, $\chi 2$ test
	Ordinal	Ordinal attribute values also order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric Quantitative	Interval	For interval attributes, differences between values are meaningful. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
	Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, current	geometric mean, harmonic mean, percent variation

Data Quality

- Poor data quality negatively affects many data processing efforts

"The most important point is that poor data quality is an unfolding disaster.

- Poor data quality costs the typical company at least ten percent (10%) of revenue; twenty percent (20%) is probably a better estimate."

  Thomas C. Redman, DM Review, August 2004

- Data mining example: a classification model for detecting people who are loan risks is built using poor data
  - Some credit-worthy candidates are denied loans
  - More loans are given to individuals that default

Data Quality …

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

- Examples of data quality problems:
  - Wrong data
  - Duplicate data
  - Noise and outliers
  - Missing values

Data Quality issues …

- **Syntactic accuracy:** Entry is not in the domain.
  - **Examples**: fmale in gender, text in numerical attributes, ... Can be checked quite easy.
- **Semantic accuracy:** Entry is in the domain but not correct
  - **Example**: John Smith is female
  - Needs more information to be checked (e.g. “business rules”).
- **Completeness**: is violated if an entry is not correct although it belongs to the domain of the attribute.
  - **Example**: Complete records are missing, the data is biased (A bank has rejected customers with low income.)
- **Unbalanced data**: Dataset might be extremely biased to one type of records.
  - **Example**: Defective goods are a very small fraction of all.
- **Timeliness:** Is the available data up to date?

Duplicate Data

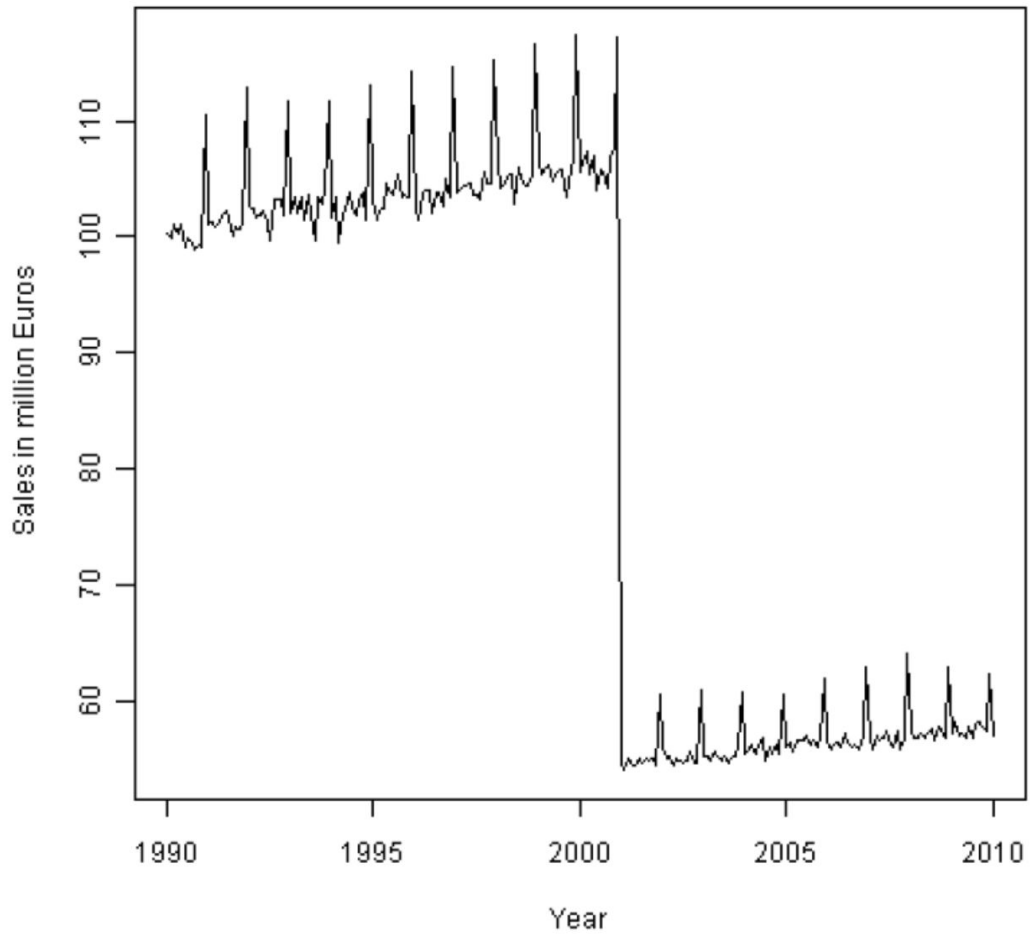
- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeneous sources
- Examples:
  - Same person with multiple email addresses
- Data cleaning
  - Process of dealing with duplicate data issues

Statistics & Visualization

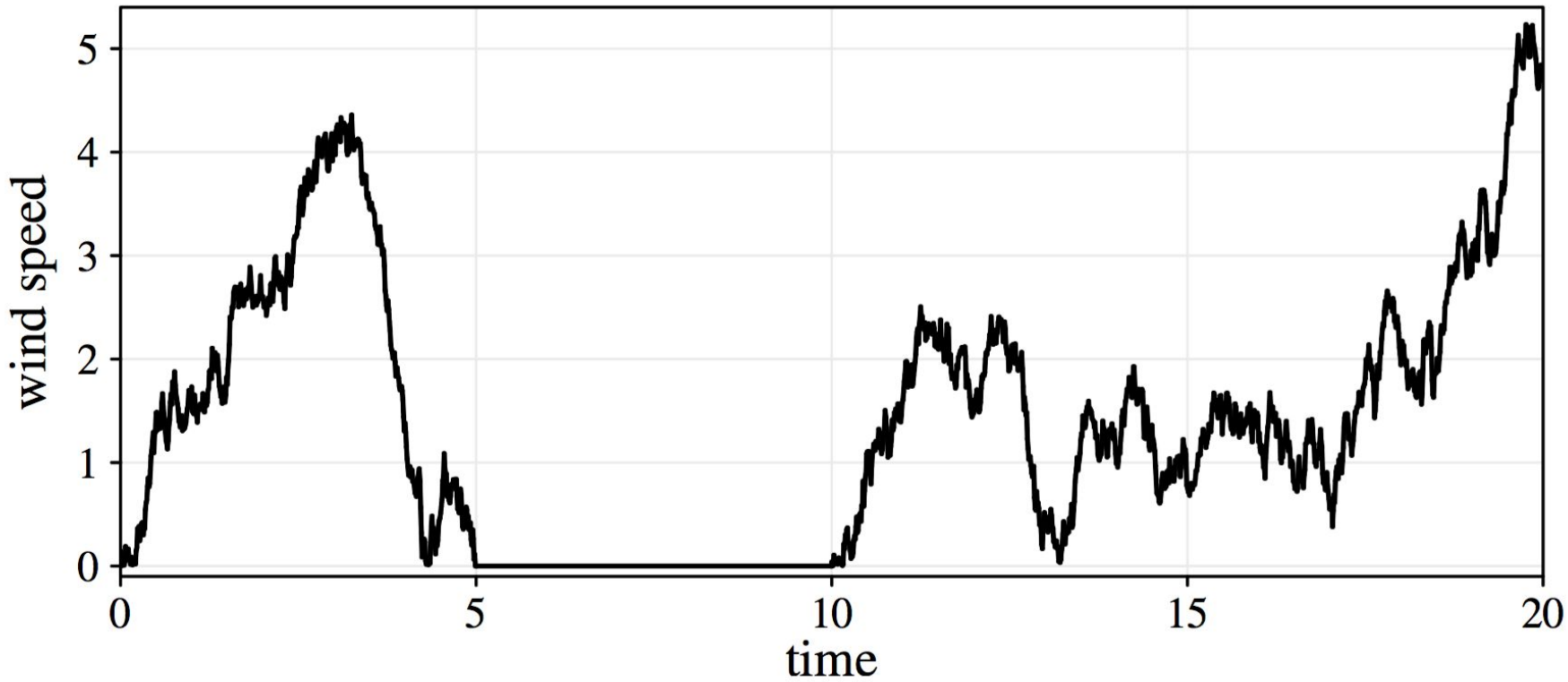
In order to know our data and discovery quality issues we need:

- Use descriptive statistics for getting a global picture and summarize properties of data
- Compare statistics with the expected behaviour
- Exploit visualization techniques that can help in detecting
  - general patterns and trends
  - outliers and unusual patterns

Data Visualization

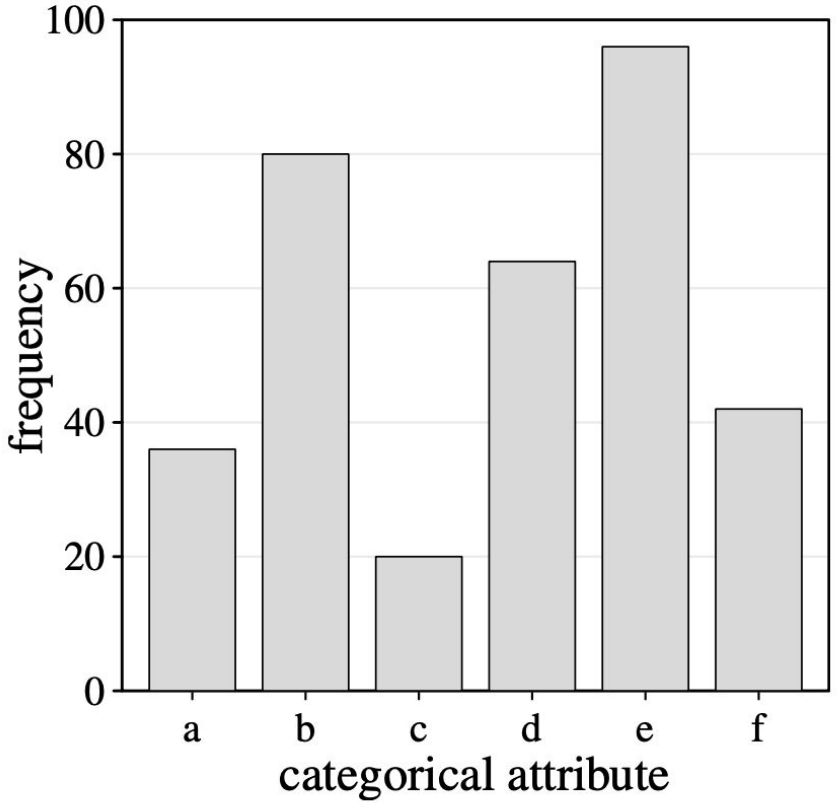


Data Visualization



The zero values might come from a broken or blocked sensor and might be consider as missing values.

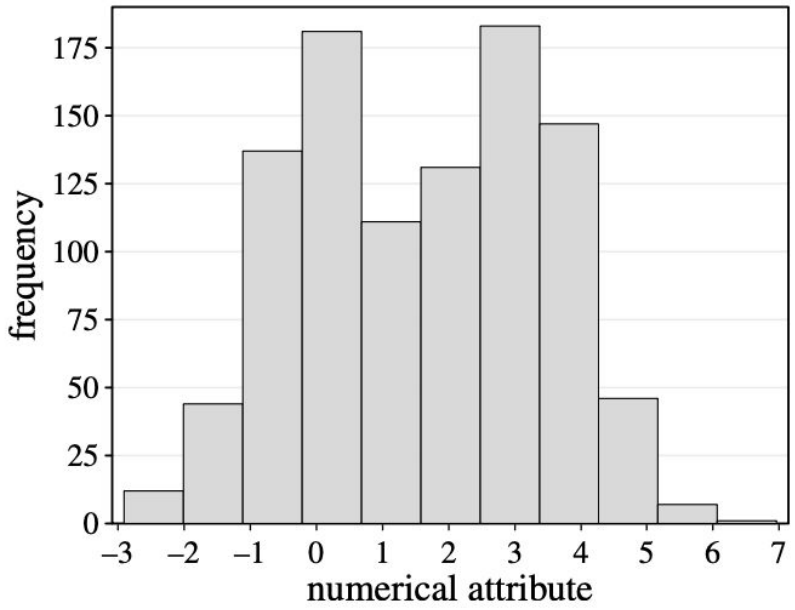
Bar Chart for Categorical Attributes



A bar chart is a simple way to depict the frequencies of the values of a categorical attribute.

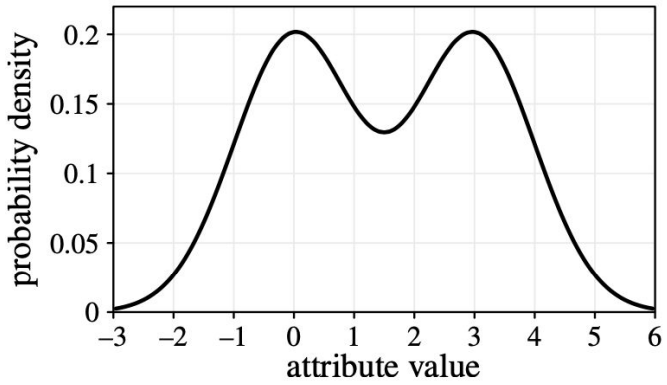
Histograms for Numerical Attributes

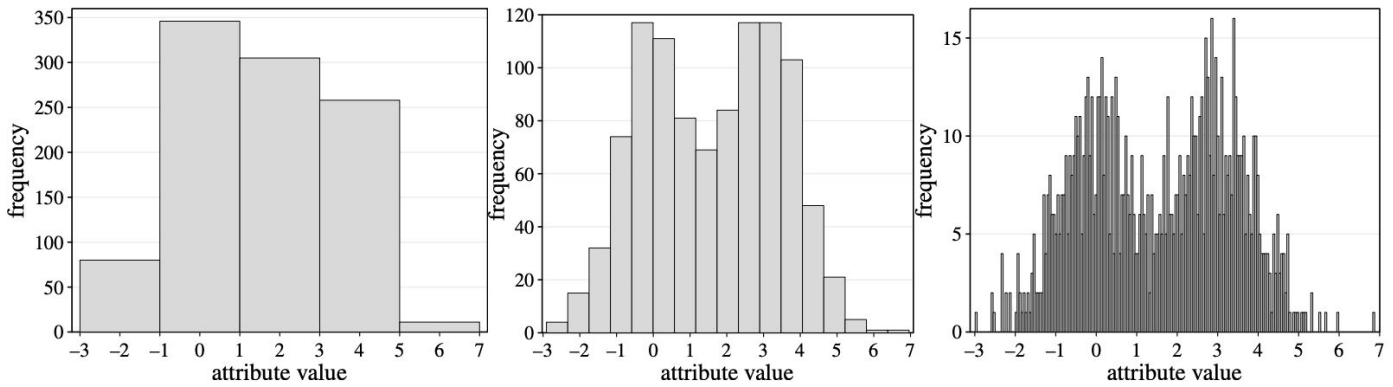
- A **histogram** shows the frequency distribution for a numerical attribute.
- The range of the numerical attribute is **discretized** into a fixed number of intervals (**bins**)
- For each interval the (absolute) **frequency** of values falling into it is indicated by the height of a bar.



Histograms: Number of bins

3 histograms with 5, 17 and 200 bins for a sample from the same bimodal distribution.





Number of bins

- Number of bins according to **Sturges' rule:**

$$k = \lceil \log_2(n) + 1 \rceil$$

where n is the sample size

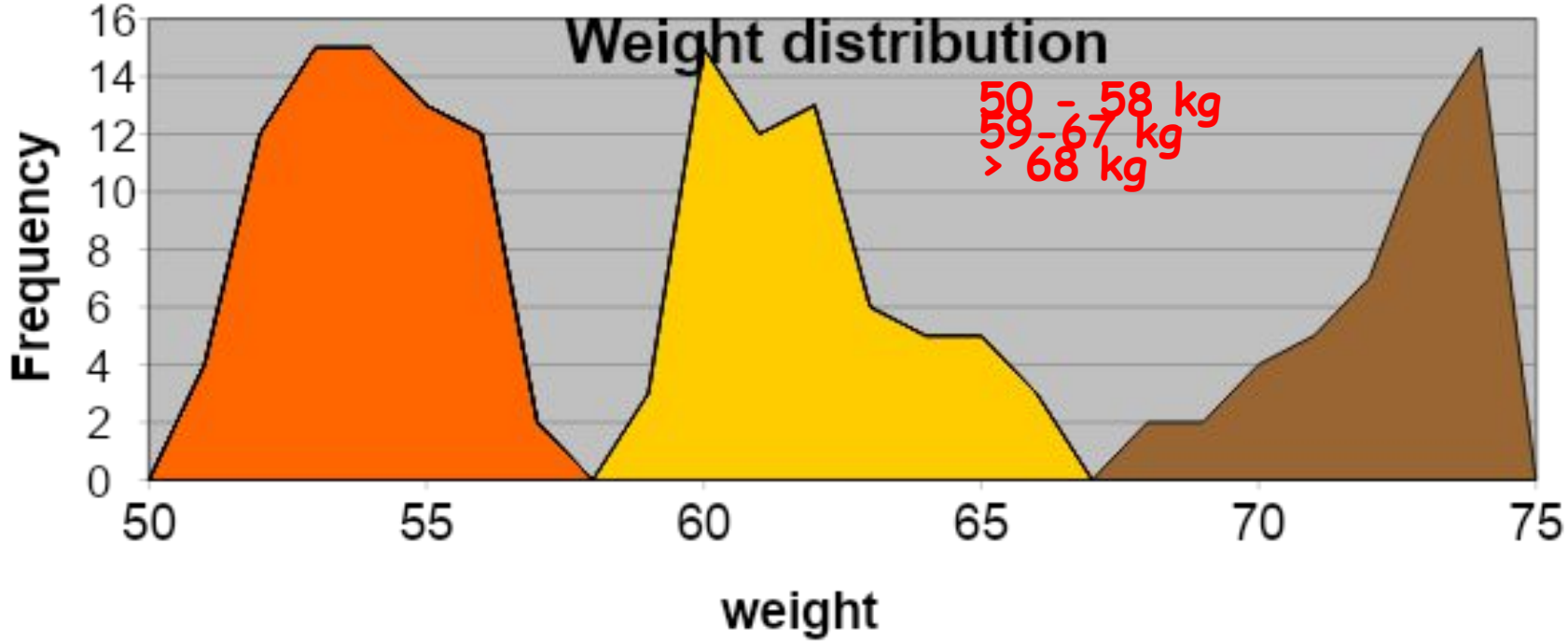
- Sturges' rule is suitable for data from normal distributions and from data sets of moderate size.

How to choose intervals?

1. Interval with a fixed “reasonable” granularity
   Ex. intervals of 10 cm for height.

2. Interval size is defined by some domain dependent criterion
   Ex.: 0-20ML, 21-22ML, 23-24ML, 25-26ML, >26ML

3. Interval size determined by analyzing data, studying the distribution and find breaks or using clustering



Natural Binning

- Simple
- Sort of values, subdivision of the range of values in *k* parts with the same size

$$\delta = \frac{x_{\max} - x_{\min}}{k}$$

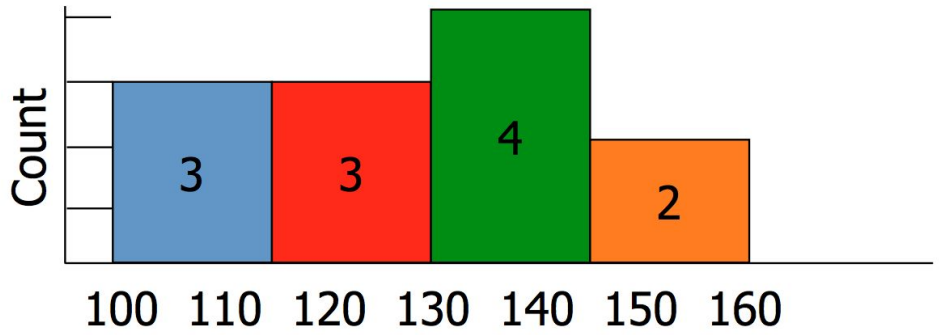
- Element x_j belongs to the class i if

$$x_j \in [x_{min} + i\delta,\ x_{min} + (i+1)\delta)$$

- It can generate distribution very unbalanced

Example

- Histogram for Price
- δ =(160-100)/4 = 15
- bin 1: [100,115)
- bin 2: [115,130)
- bin 3: [130,145)
- bin 4: [145, 160]



Bar	Beer	Price
A	Bud	100
A	Becks	120
C	Bud	110
D	Bud	130
D	Becks	150
E	Becks	140
E	Bud	120
F	Bud	110
G	Bud	130
H	Bud	125
H	Becks	160
I	Bud	135

Equal Frequency Binning

- Sort and count the elements, definition of k intervals of f, where:

$$f = \frac{N}{k}$$

(N = number of elements of the sample)

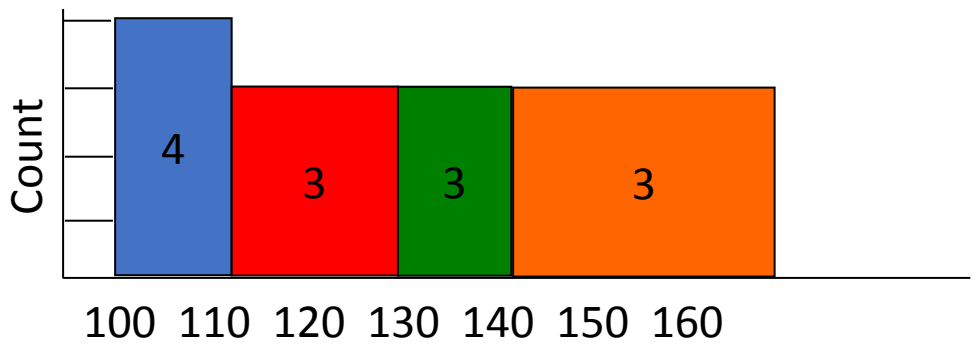
- The element x_i belongs to the class j if

$$j \times f \leq i < (j+1) \times f$$

- It is not useful for highlighting interesting distribution

Example

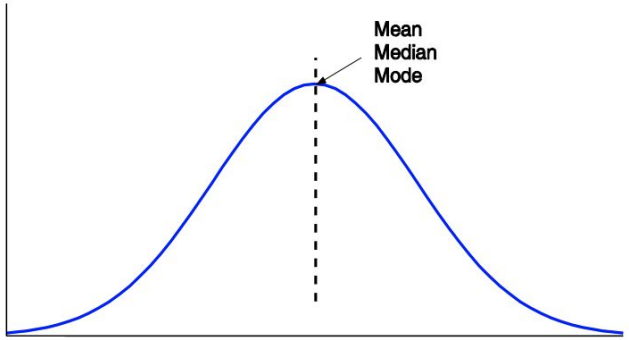
- Histogram for Price
- $f = 12/4 = 3$
- class 1: {100,110,110}
- class 2: {120,120,125}
- class 3: {130,130,135}
- class 4: {140,150,160}



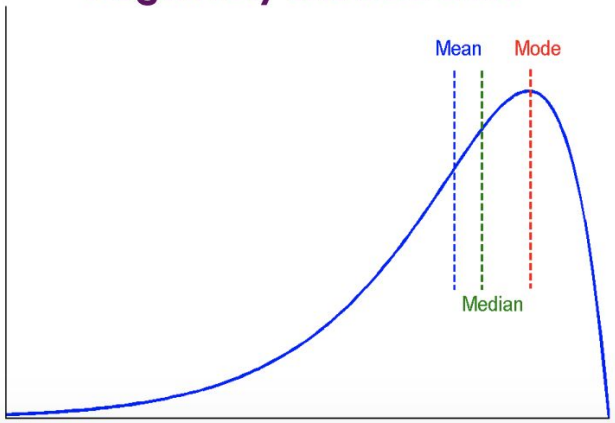
Bar	Beer	Price
A	Bud	100
A	Becks	120
C	Bud	110
D	Bud	130
D	Becks	150
E	Becks	140
E	Bud	120
F	Bud	110
G	Bud	130
H	Bud	125
H	Becks	160
I	Bud	135

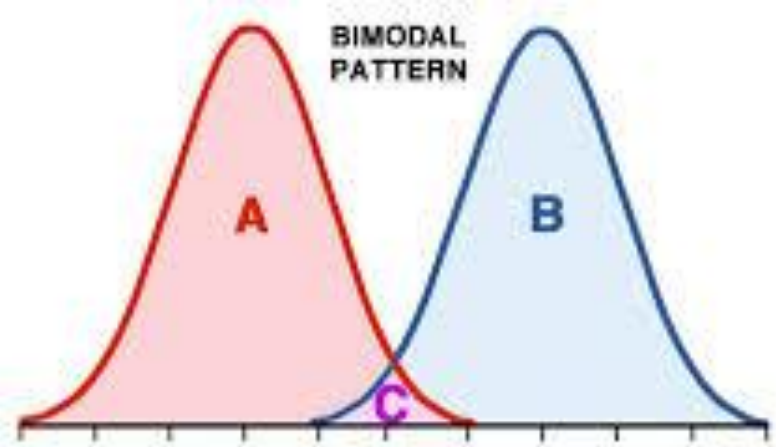
Observing Data Distribution

Symmetric data

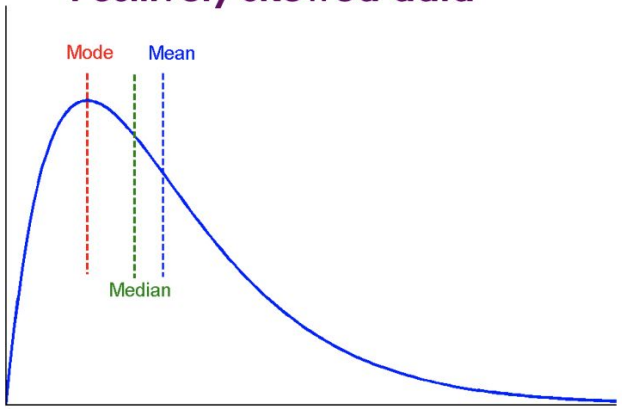


Negatively skewed data





Positively skewed data



Example

Give an example of something having a positively skewed distribution

- **income** is a good example of a positively skewed variable: there will be a few people with extremely high incomes, but most people will have incomes bunched together below the mean.

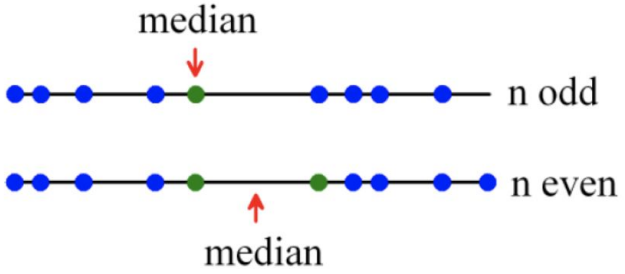
Give an example of something having a bimodal distribution

- bimodal distribution has some kind of underlying binary variable that will result in a separate mean for each value of this variable.
- One example can be **human weight** – the gender is binary and is a statistically significant indicator of how heavy a person is.

Measuring the Central Tendency

- Mean
  - *m* is the sample size
  - A distributive measure can be computed by partitioning the data into smaller subsets
  - However, the mean is very sensitive to outliers
  - The median or a trimmed mean are also commonly used

$$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$



- Median
  - Middle value if odd number of values, or average of the middle two values otherwise
- Mode
  - Value that occurs m**ost frequently** in the data
  - It is possible that several different values have the greatest frequency: Unimodal, bimodal, trimodal, multimodal
  - If each data value occurs only once then **there is no mode**

Measuring the Dispersion of Data

- The degree in which data tend to spread is called the **dispersion**, or **variance** of the data

- The most common measures for data dispersion are **range**, **standard deviation,** the **five-number summary** (based on quartiles), and the **inter-quartile range**

- **Range**: The distance between the largest and the smallest values

Measuring the Dispersion of Data

- **Variance**

$$\text{variance}(x) = s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2$$

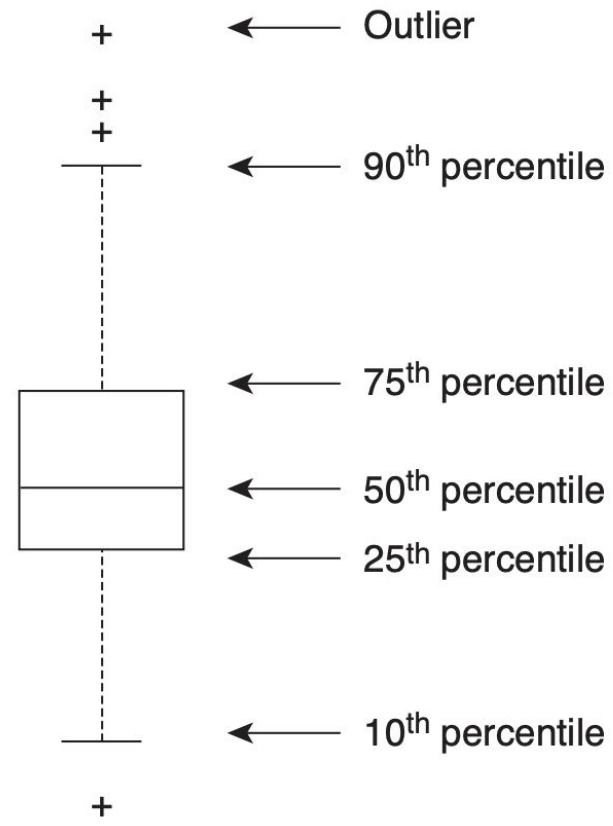
- **Standard deviation** σ is the square root of variance σ^2
  - σ measures spread about the mean ands should be used only when the mean is chosen as the measure of the center
  - σ=0 only when there is no spread, that is, when all observations have the same value. Otherwise σ>0
- **Because of outliers**, other measures are often used:
  - **absolute average deviation** (AAD)
  - **median average deviation** (MAD)

$$\text{AAD}(x) = \frac{1}{m}\sum_{i=1}^{m}|x_i - \overline{x}|$$

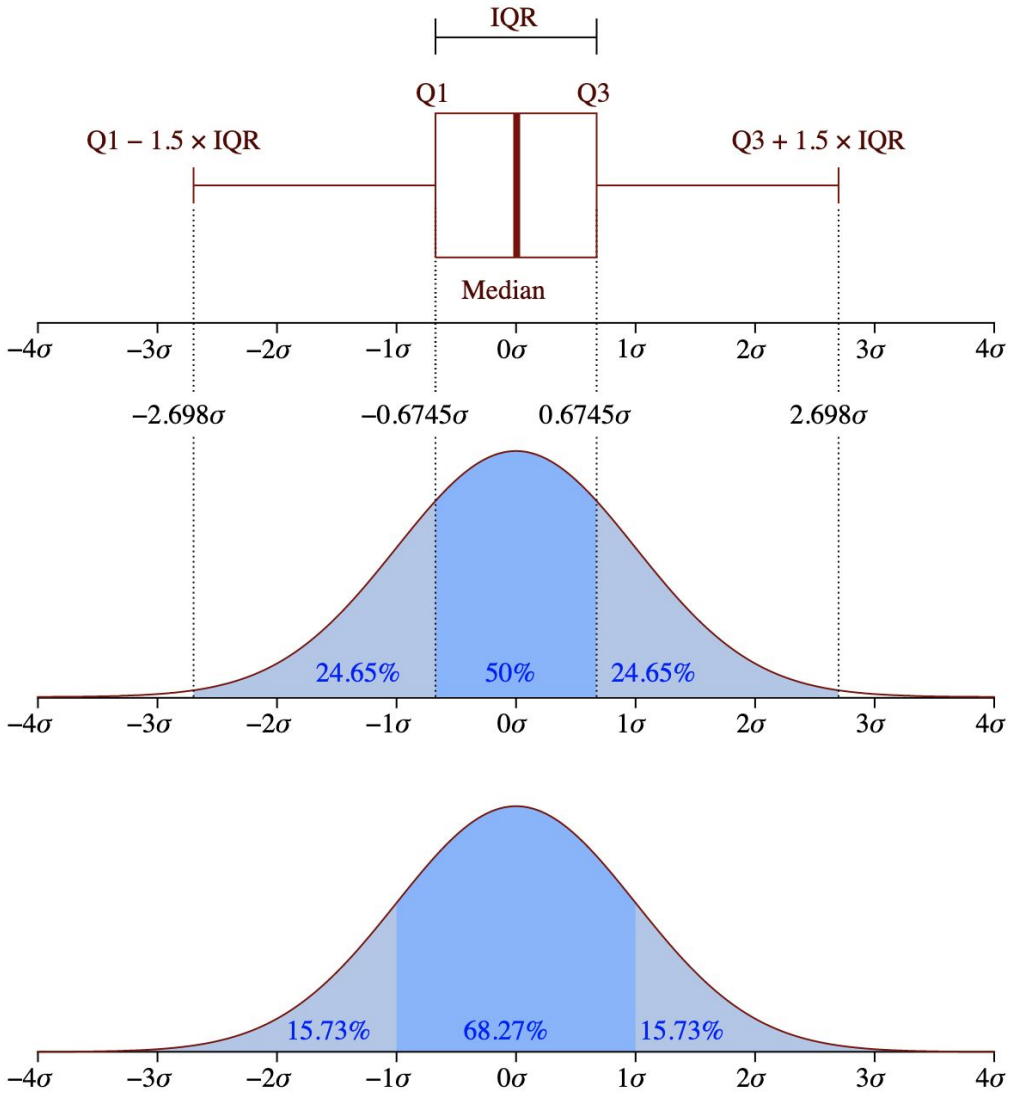
$$\text{MAD}(x) = median\left(\{|x_1 - \overline{x}|, \ldots, |x_m - \overline{x}|\}\right)$$

Box Plot: Five-number summary of a distribution

- Data represented with a **box**
- The ends of the box are at the
  - **Q1: 1st quartiles** (25%-quantile or 25th percentile)
  - **Q3: 3rd quartiles** (75%-quantile or 75th percentile)
- **Median**: value in the middle is the **Q2**: **2nd quartile** (50%-qua,, 50th perc.)
- The height of the box is **Interquartile range** (**IQR**): Q3 - Q1
- **Whiskers**: two lines outside the box extended from:
  - 1st, or 5th, or 10th percentile, or Q1 – *k* IQR (with k = 1.5)
  - 99th, or 95th, or 90th percentile, or Q3 + *k* IQR (with k = 1.5)
- **Outliers**: are points beyond whiskers
- In general, p%-quantile (0 < p < 100): Is the value *x* s.t. p% of the values are smaller and 100-p% are larger.



Relationship Between Box-Plot and Histogram



Example Data Set: Iris data

iris setosa

iris versicolor

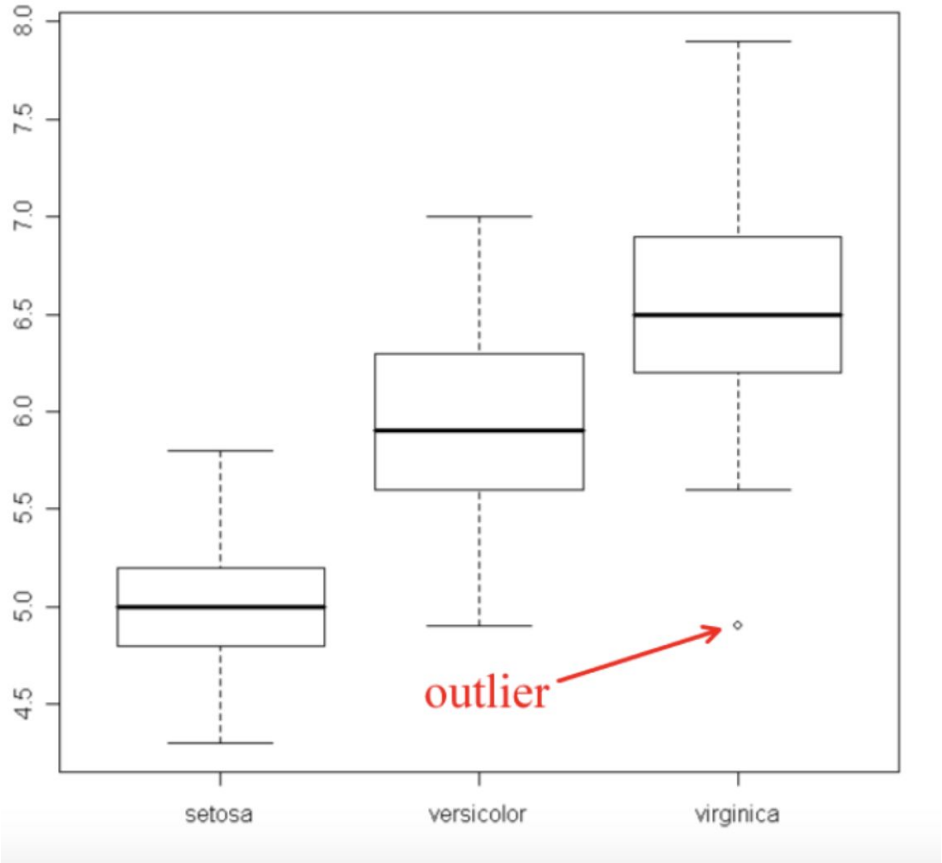
iris virginica

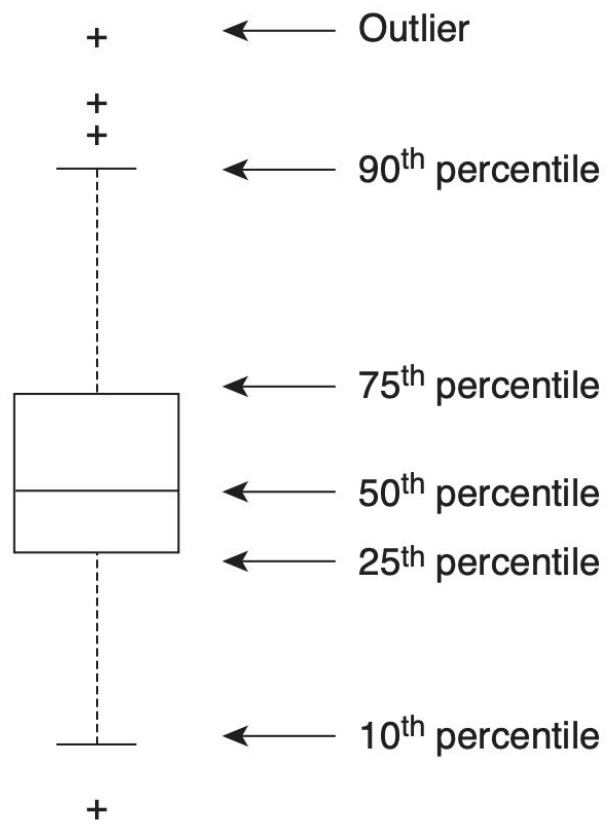
- collected by E. Anderson in 1935
- contains measurements of four real-valued variables:
    - sepal length, sepal widths, petal lengths and petal width of 150 iris flowers of types *Iris Setosa*, *Iris Versicolor*, *Iris Virginica* (50 each)
- The fifth attribute is the name of the flower type.

Example data set: Iris data

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
5.1	3.5	1.4	0.2	Iris-setosa
…				
…				
5.0	3.3	1.4	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor
…				
…				
5.1	2.5	3.0	1.1	Iris-versicolor
5.7	2.8	4.1	1.3	Iris-versicolor
…				
…				
5.9	3.0	5.1	1.8	Iris-virginica

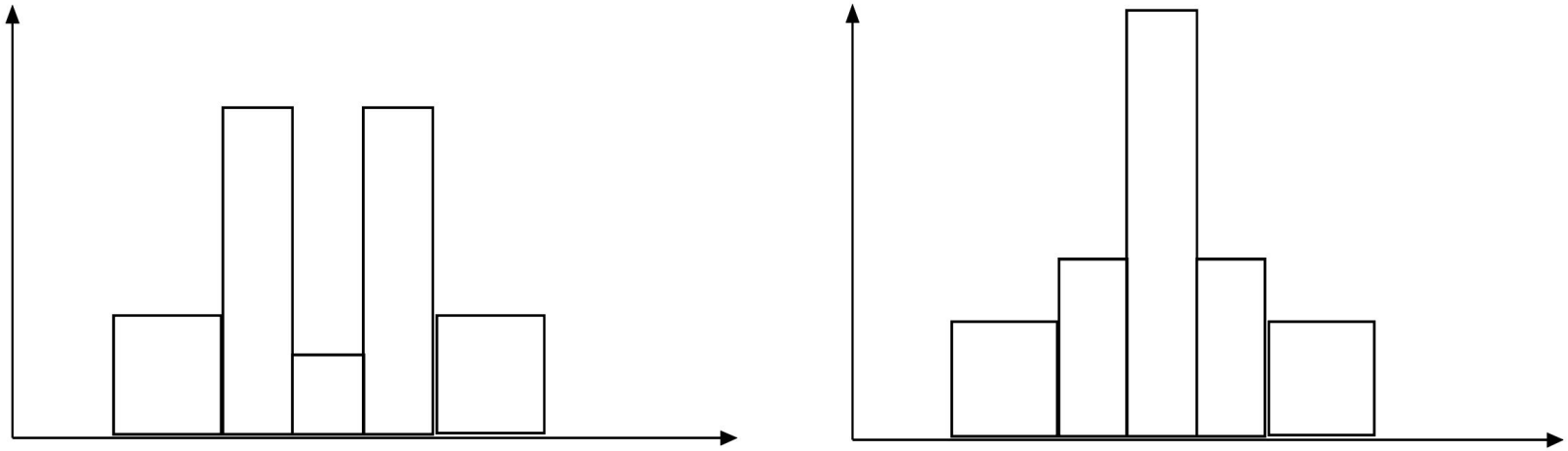
Example of Conditional Box Plot





Histograms Often Tell More than Boxplots

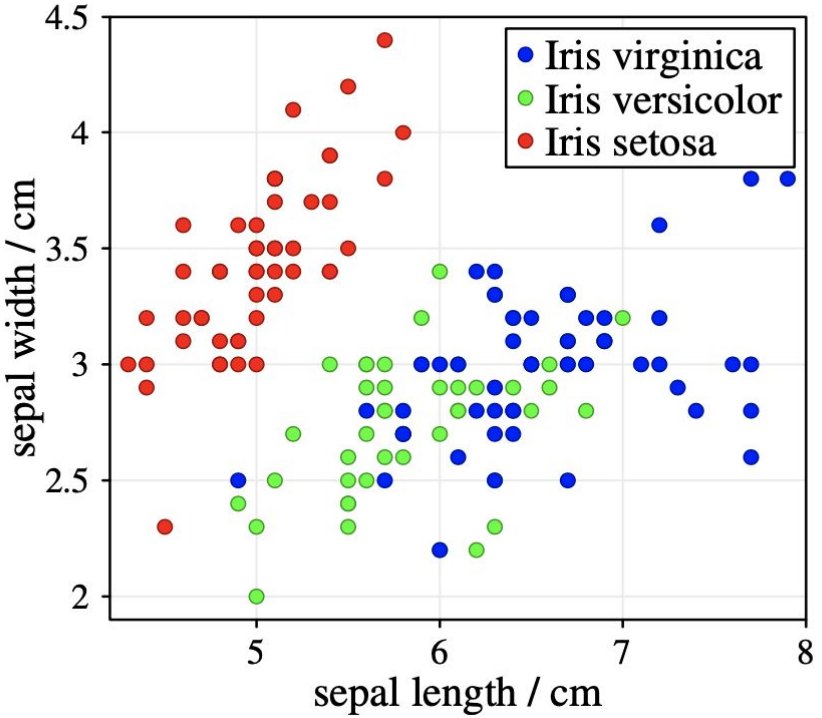
- The two histograms may have the same boxplot representation
  - **The same values** for: **min, Q1, median, Q3, max**
  - But they have rather **different data distributions**



Scatter Plot

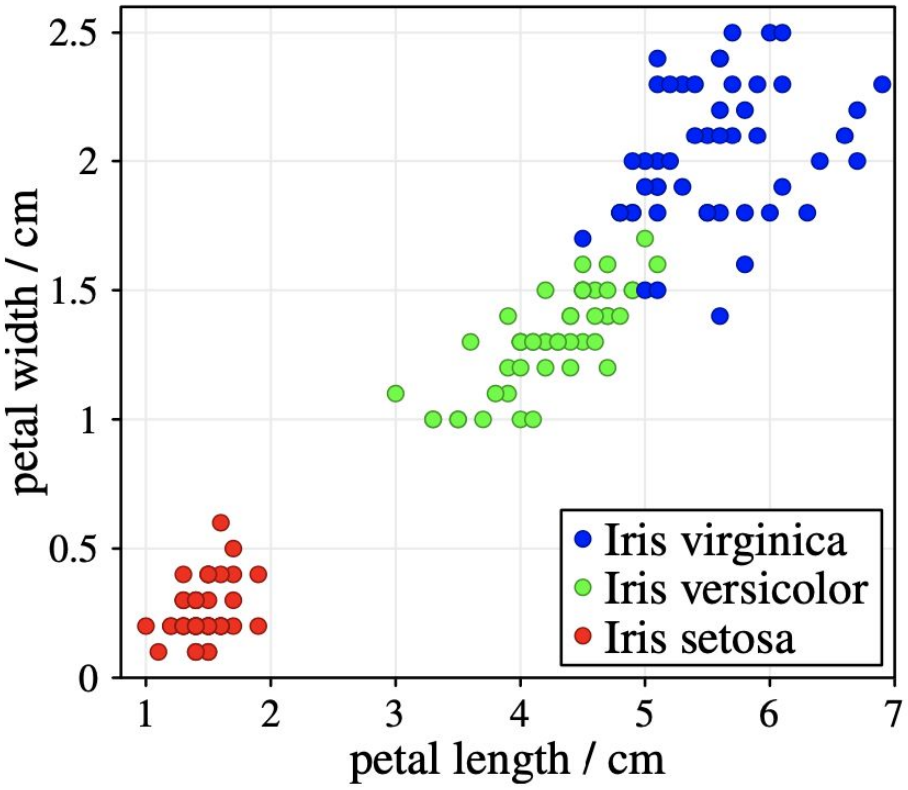
- Provides a first look at **bivariate data** to see **clusters** of points, **outliers**, **correlations**

- Each pair of values is treated as a **pair of coordinates** and plotted as points in the plane



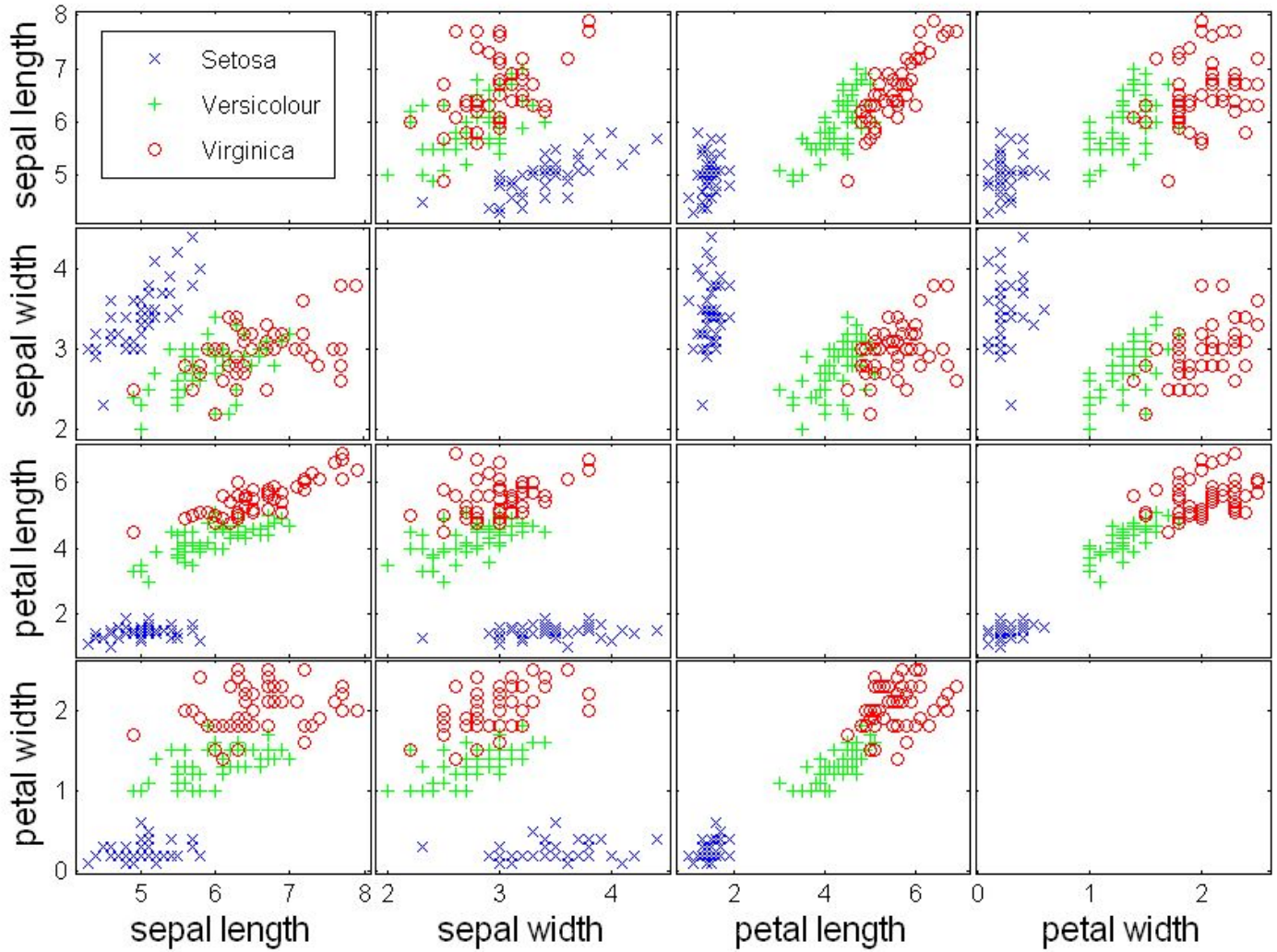
Scatter plots can be enriched with additional information: **Colour** or **different symbols** to **incorporate a third attribute** in the scatter plot.

Scatter Plot



The two attributes petal length and width provide a **better separation of the classes** Iris versicolor and Iris virginica than the sepal length and width.

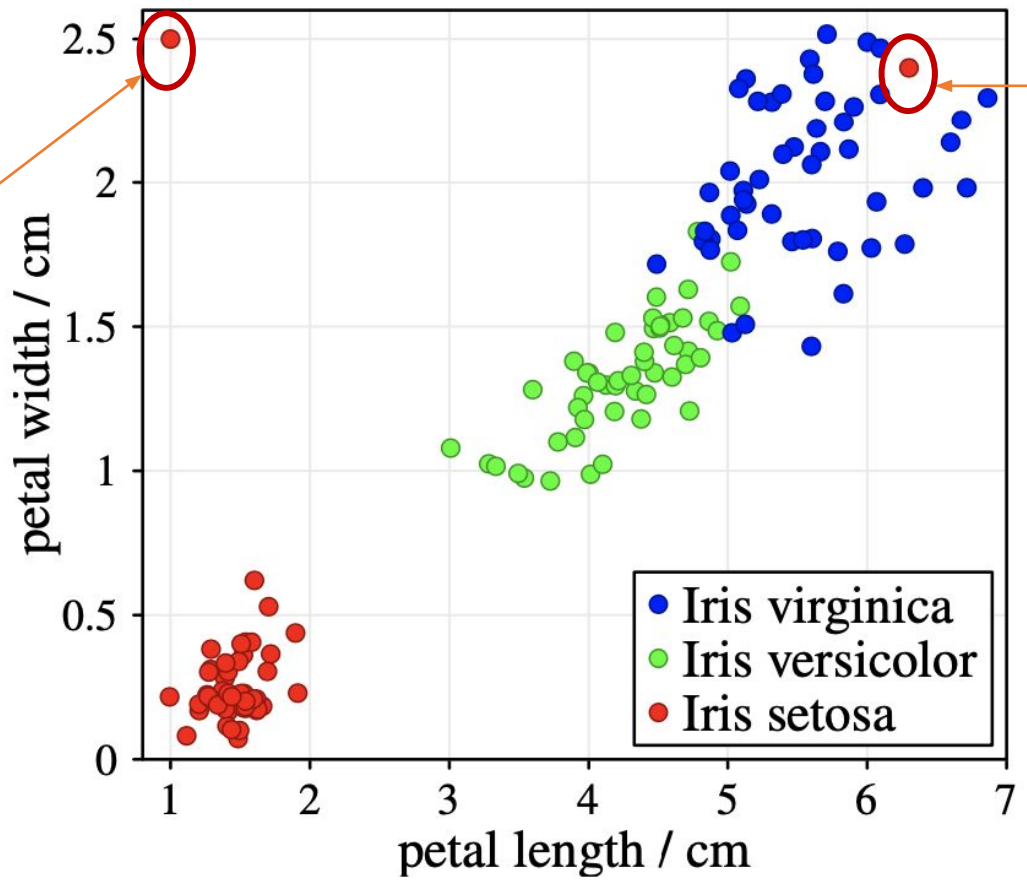
Scatter Matrix of Iris Attributes



Scatter Plot & Outliers

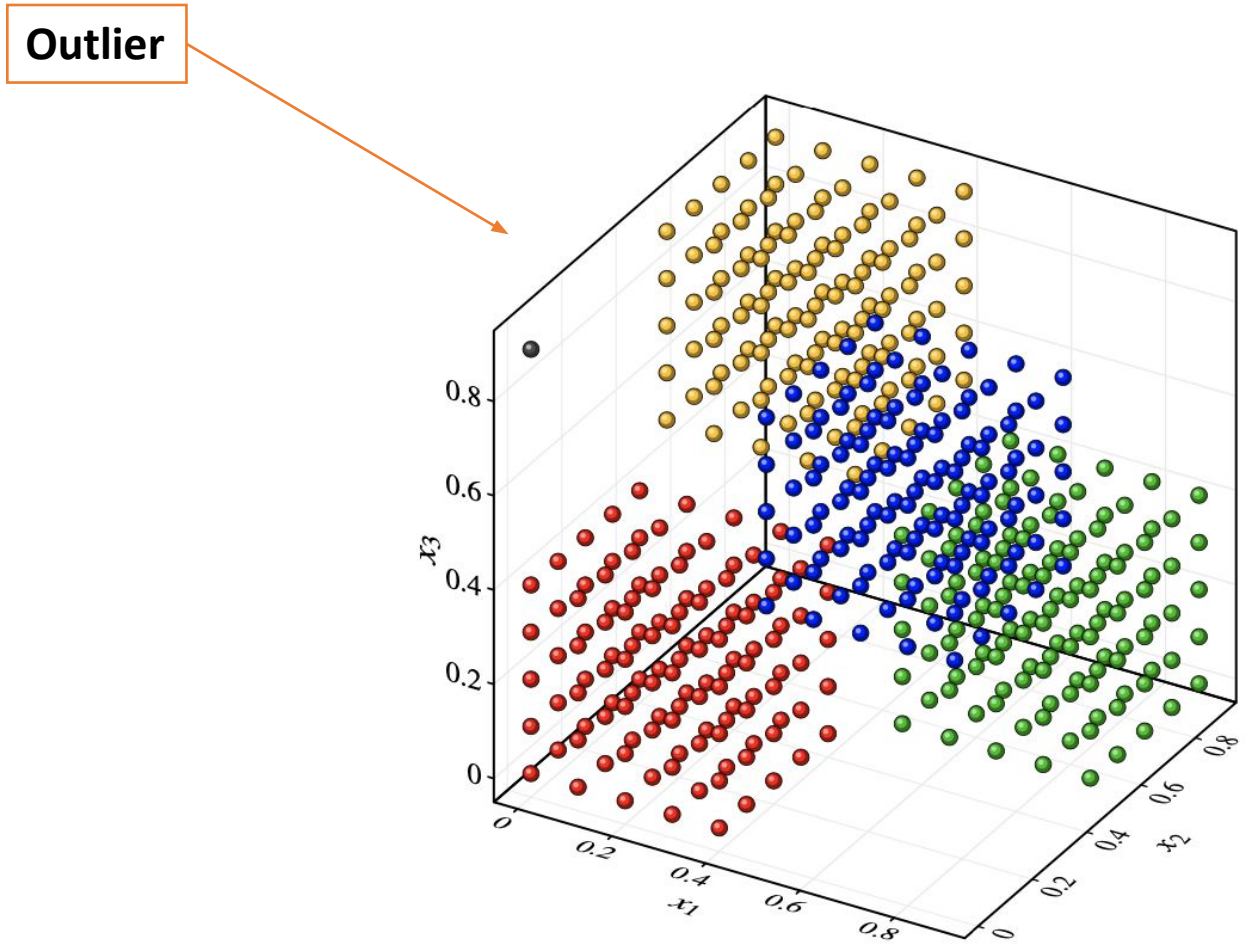
The Iris data set with two (additional artificial) outliers

outlier for the **whole dataset**



outlier for the **class Iris setosa.**

3D Scatter Plot



Visualization as a Test

- When visualisations reveal patterns or exceptions, then there is “something” in the data set.

- When visualisations do not indicate anything specific, there might still be patterns or structures in the data that cannot be revealed by the corresponding (simple) visualisation techniques.

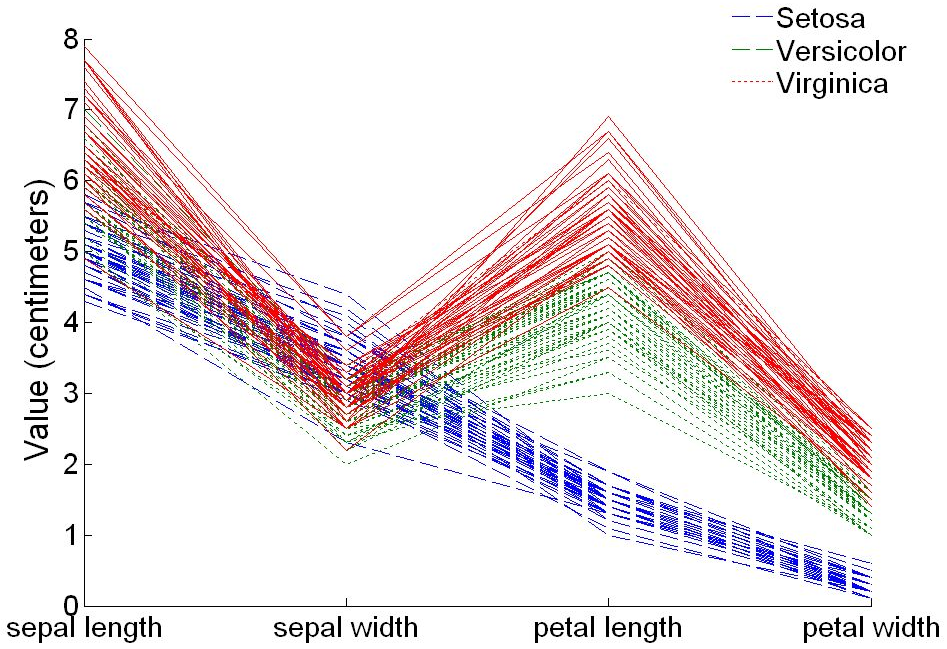
Parallel Coordinates

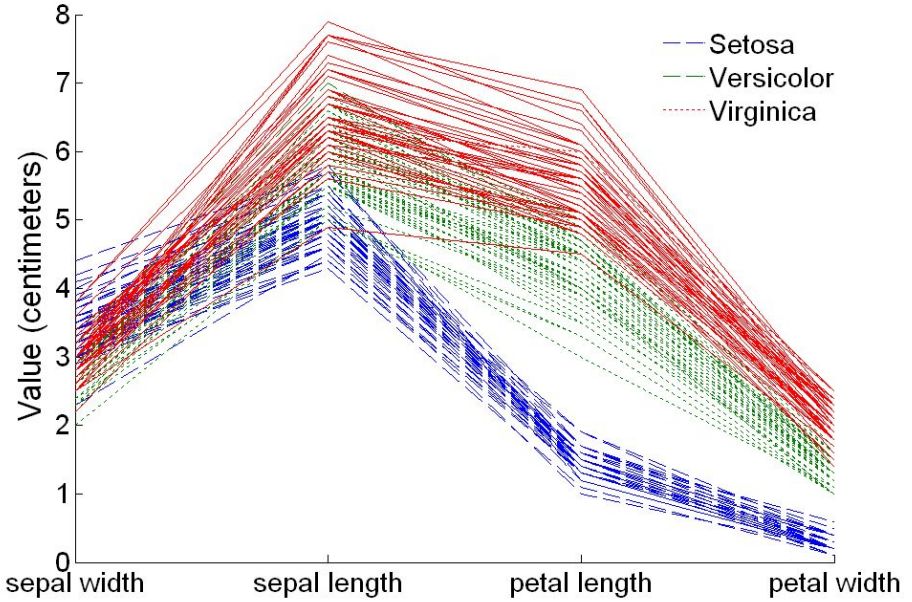
- Used to plot the attribute values of high-dimensional data
- Instead of using perpendicular axes, use a set of parallel axes

The attribute values of each object are plotted as a point on each corresponding coordinate axis and the points are connected by a line

- Thus, each object is represented as a line
- Often, the lines representing a distinct class of objects group together, at least for some attributes
- Ordering of attributes is important in seeing such groupings

Parallel Coordinates Plots for Iris Data

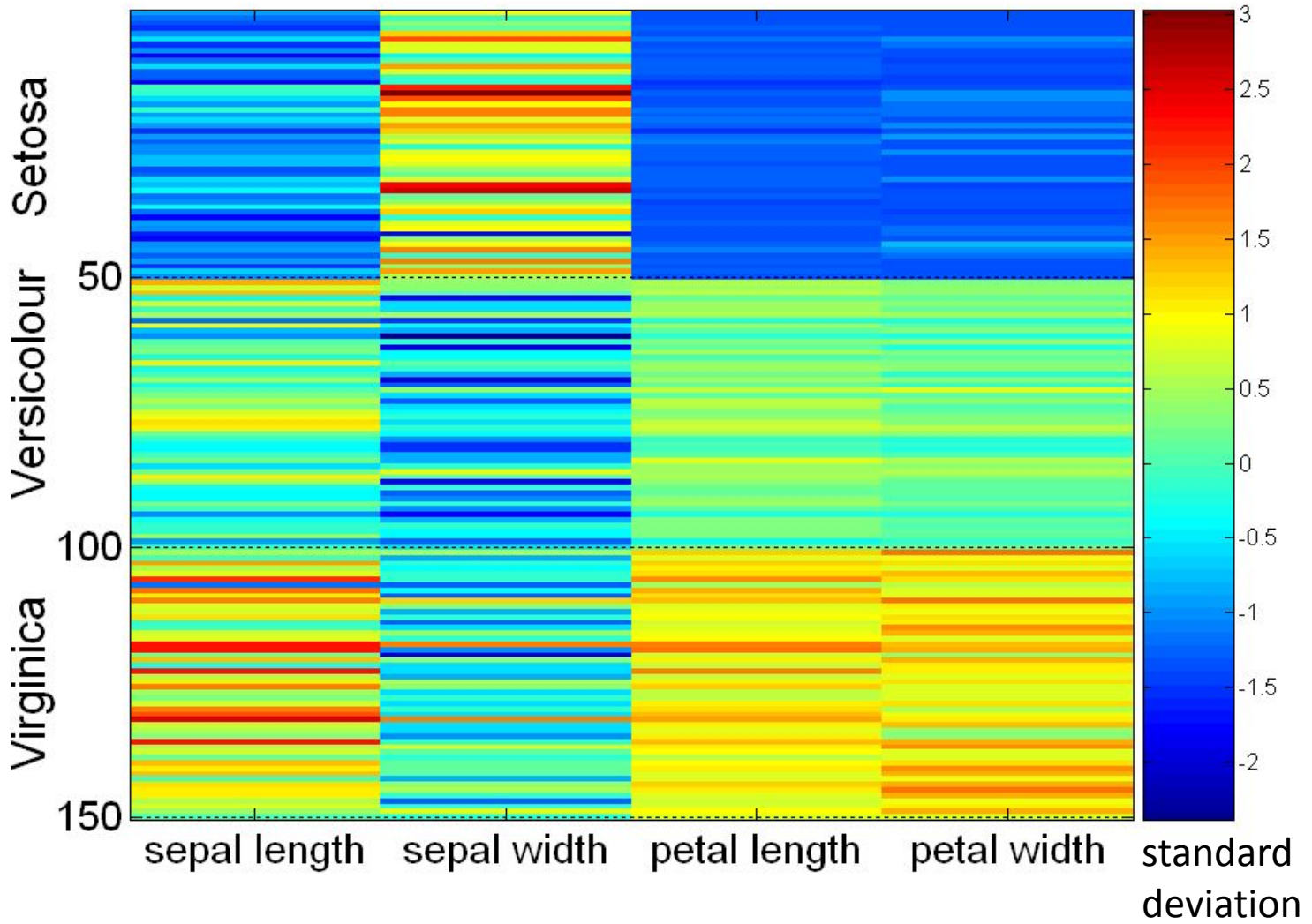




Matrix Plots

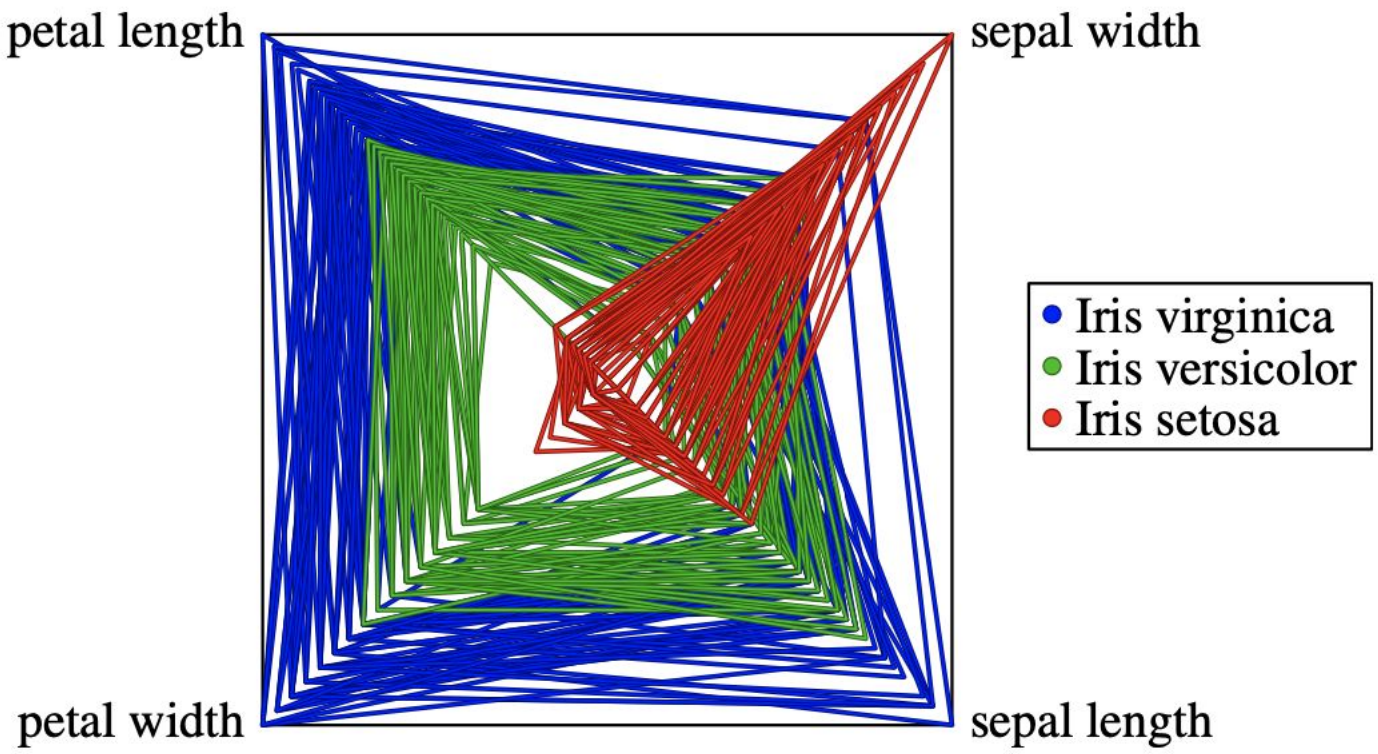
- Can plot the data matrix
- This can be useful when objects are sorted according to class
- Typically, the attributes are normalized to prevent one attribute from dominating the plot
- **Plots of similarity or distance matrices** can also be useful for visualizing the relationships between objects

Visualization of the Iris Data Matrix

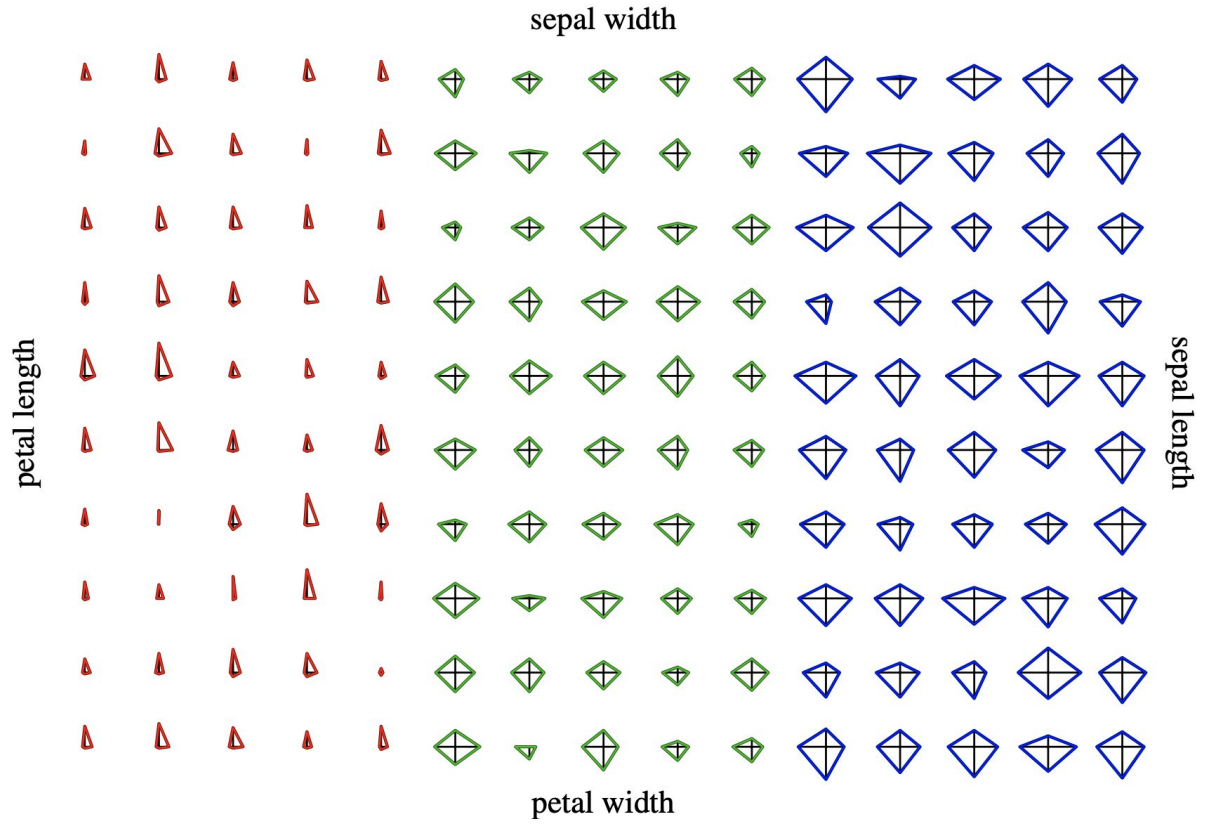


Radar Plot for Iris Data

- Similar idea as parallel coordinates
- Coordinate axes are drawn as parallel lines, but in a star-like fashion intersecting in one point
- Axes radiate from a central point
- The line connecting the values of an object is a polygon



Star Plots for Iris Data



Star plots are the same as radar plots where **each data object is drawn separately**.

Correlation Analysis

- Correlation measures the linear relationship between objects
- Captures similar behaviour of two attributes

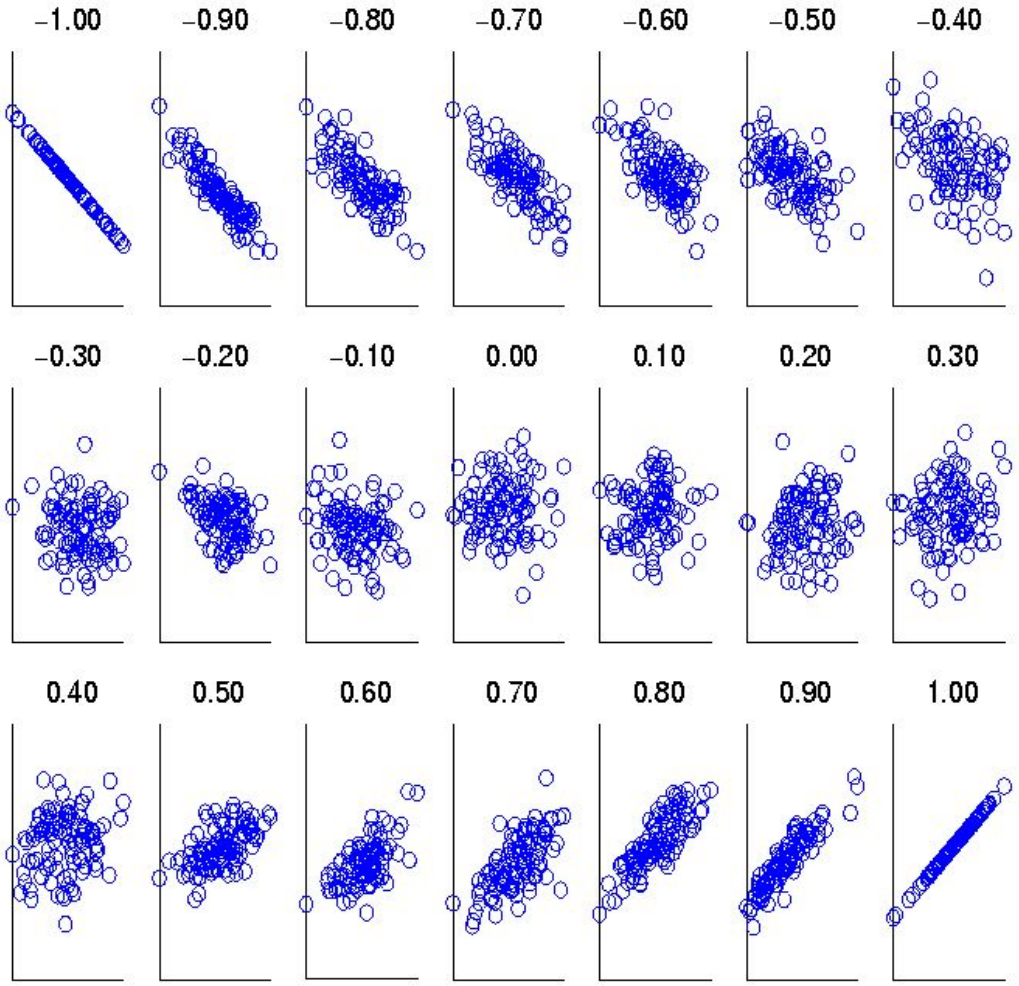
Pearson's Correlation Coefficient

The (sample) Pearson's correlation coefficient is a measure for a linear relationship between two numerical attributes X and Y and is defined as

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \qquad -1 \leq r_{xy} \leq 1$$

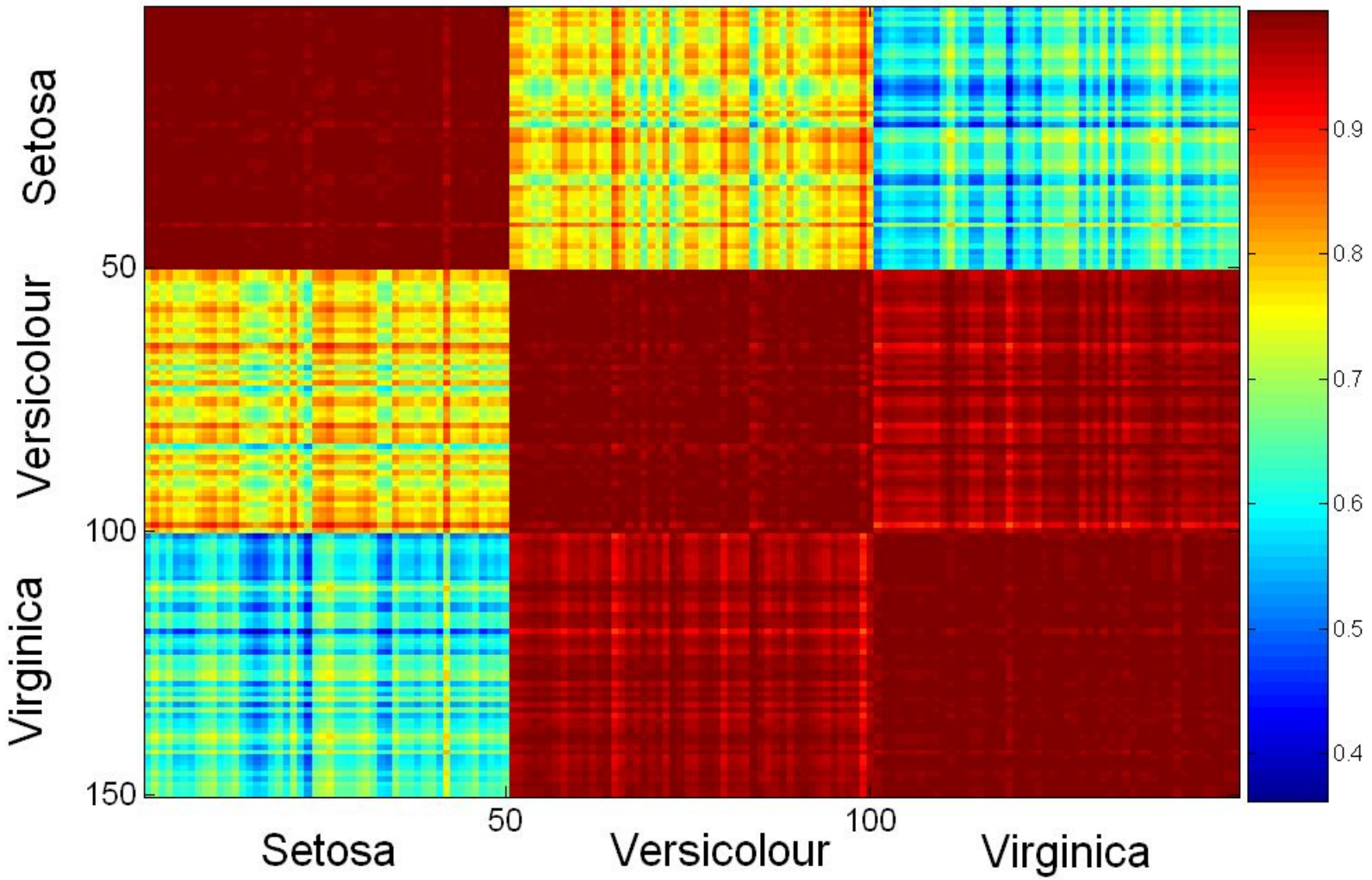
- where $\bar{x}$ and $\bar{y}$ are the mean values of the attributes X and Y, respectively. s_x and s_y are the corresponding (sample) standard deviations.
- The larger the absolute value of the Pearson correlation coefficient, the stronger the linear relationship between the two attributes.
- For $|r_{xy}| = 1$ the values of X and Y lie exactly on a line.
- Positive (negative) correlation indicates a line with positive (negative) slope.

Visually Evaluating Correlation



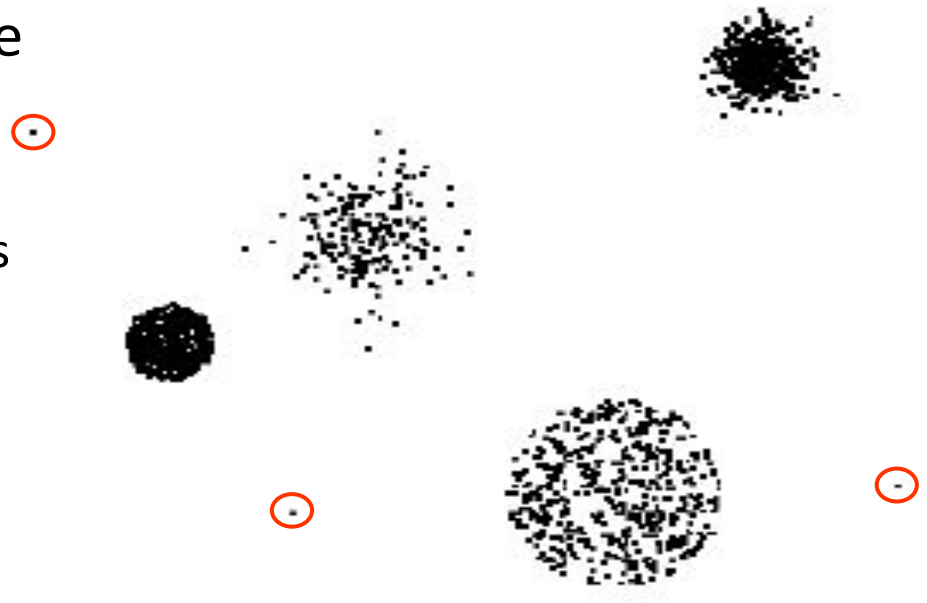
Scatter plots showing the similarity from –1 to 1.

Visualization of the Iris Correlation Matrix



Outliers

- **Outliers** are data objects with characteristics that are considerably different than most of the other data objects in the data set
  - **Case 1:** Outliers are noise that interferes with data analysis
  - **Case 2:** Outliers are the goal of our analysis
    - Credit card fraud
    - Intrusion detection
- **Causes:**
  - Data quality problems (erroneous data coming from wrong measurements or typing mistakes)
  - Exceptional or unusual situations/data objects.



Outliers as Noise

- Outliers coming from erroneous data should be excluded from the analysis
- Even if the outliers are correct (exceptional data), it is sometime useful to exclude them from the analysis.
- For example, a single extremely large outlier can lead to completely **misleading values for the mean value**.

Outlier Detection

- **Single attribute:**
  - **Categorical** attributes: An outlier is a value that occurs with a frequency extremely lower than the frequency of all other values.
  - **Numerical** attributes: box plots (points outside whiskers)
- **Multidimensional attribute:**
  - Scatter plots for (visually detecting) outliers w.r.t. two attributes
  - PCA plots for (visually detecting) outliers
  - Cluster analysis techniques: Outliers are those points which cannot be assigned to any cluster.

Missing Values

- For some instances values of single attributes might be missing
- Reasons for missing values
  - **Information is not collected**
    (e.g., people decline to give their age and weight)
  - **Attributes may not be applicable to all cases**
    (e.g., annual income is not applicable to children)
  - **Broken sensors**
  - **Refusal to answer a question**
- Missing value might not necessarily be indicated as missing (instead: zero or default values).

Checklist for Data Understanding

- Determine the quality of the data. (e.g. syntactic accuracy)
- Find outliers. (e.g. using visualization techniques)
- Detect and examine missing values. Possible hidden by default values.
- Discover new or confirm expected dependencies or correlations between attributes.
- Check specific application dependent assumptions (e.g. the attribute follows a normal distribution)
- Compare statistics with the expected behaviour.

DATA MINING 1
Data Preparation

Dino Pedreschi, Riccardo Guidotti

Revisited slides from Lecture Notes for Chapter 2 "Introduction to Data Mining", 2nd Edition by Tan, Steinbach, Karpatne, Kumar

UNIVERSITÀ DI PISA

Data Understanding vs Data Preparation

Data understanding provides general information about data

- the existence of **missing values**
- the existence of **outliers**
- the character of attributes
- **dependencies** between attributes.

Data preparation uses this information to

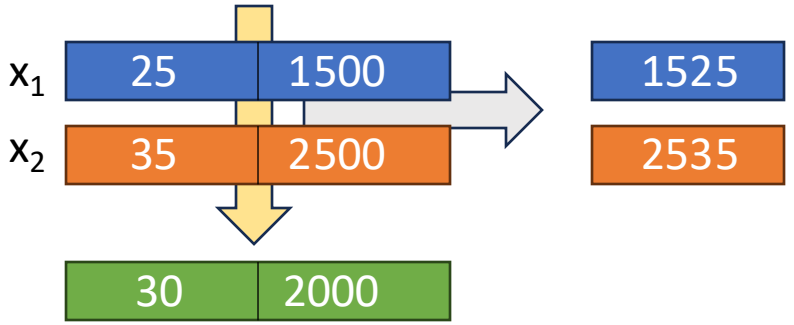
- select attributes
- reduce the data dimension
- select records
- treat missing values
- treat outliers
- integrate, unify and transform data
- improve data quality

Data Preparation

- Aggregation
- Data Reduction: Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation

Aggregation

- Combining two or more attributes (or points) into a single attribute (or point)
- Purpose
  - **Data reduction**
    - Reduce the number of attributes or objects
  - **Change of scale**
    - Cities aggregated into regions, states, countries, etc.
    - Days aggregated into weeks, months, or years
  - **More “stable” data**
    - Aggregated data tends to have less variability



Example: Precipitation in Australia

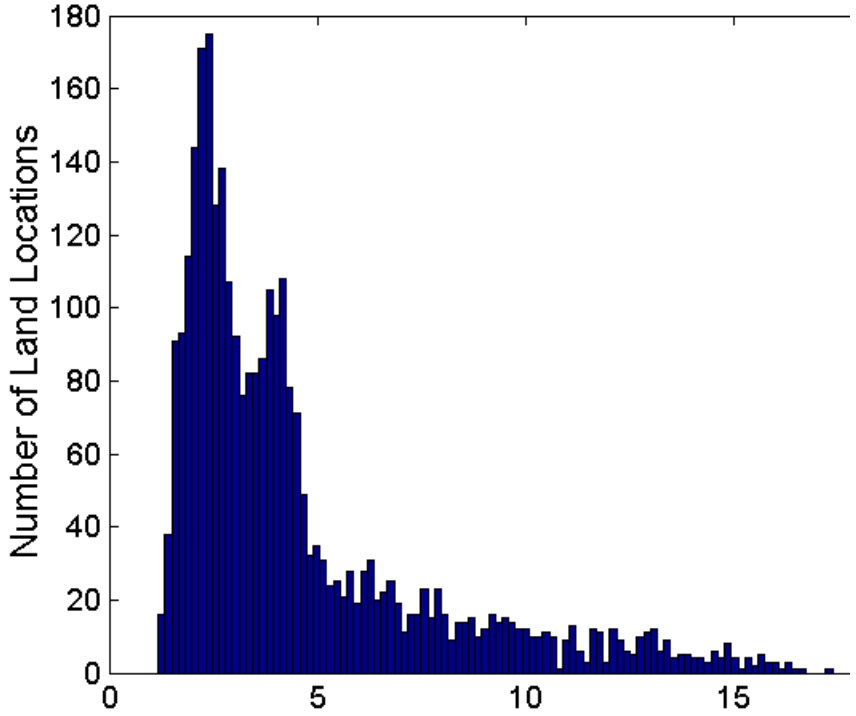
- This example is based on precipitation in Australia from 1982 to 1993.

  The next slide shows

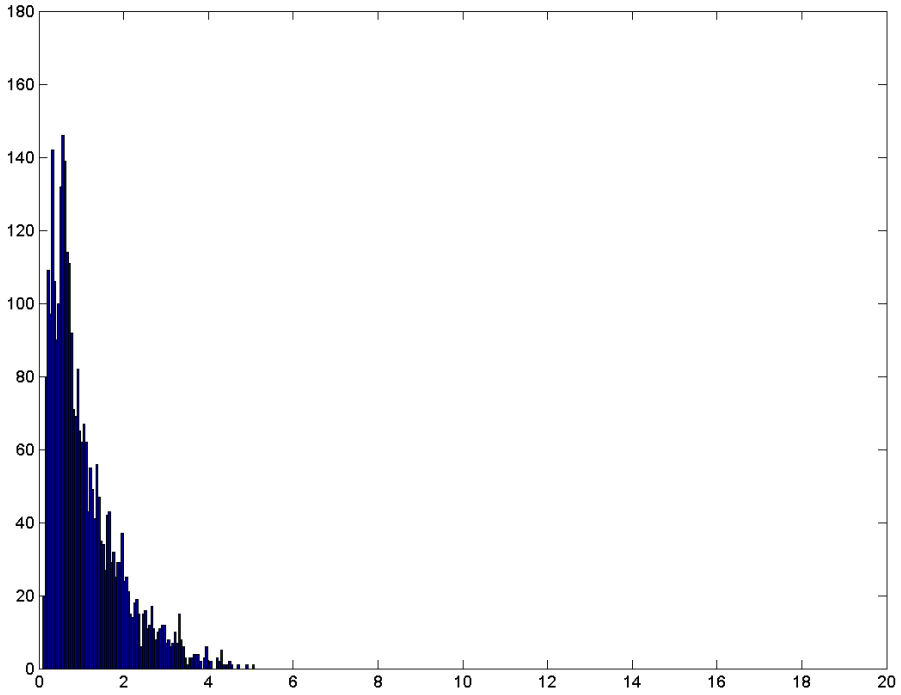
  - A histogram for the standard deviation of average monthly precipitation for specific locations in Australia, and
  - A histogram for the standard deviation of the average yearly precipitation for the same locations.
- The average **yearly precipitation has less variability** than the **average monthly precipitation.**
- All precipitation measurements (and their standard deviations) are in centimeters.

Example: Precipitation in Australia …

Variation of Precipitation in Australia



Standard Deviation of Average Monthly Precipitation



Standard Deviation of Average Yearly Precipitation

Data Reduction

Reducing the amount of data

- Reduce the number of **records** (rows)
  - Data Sampling
  - Clustering
- Reduce the number of **attributes** (columns)
  - Select a subset of attributes
  - Generate a new (a smaller) set of attributes

Sampling

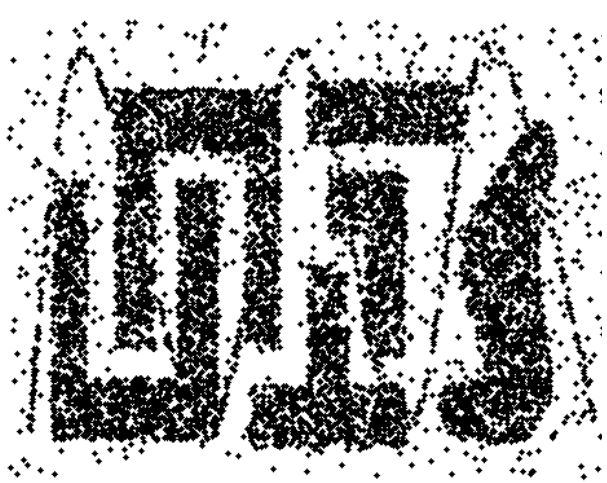
- Sampling is the main technique employed for data reduction.
  - It is often used for both the preliminary investigation of the data and the final data analysis.

- Sampling is typically used in data mining because processing the entire set of data of interest is too expensive or time consuming.

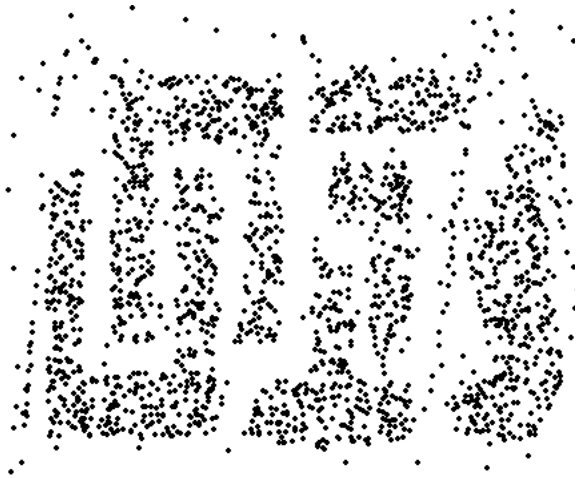
Sampling …

- The key principle for effective sampling is the following:
  - Using a sample will work almost as well as using the entire data set**, if the sample is representative**
  - A sample is representative if it has approximately the **same properties** (of interest) as the original set of data

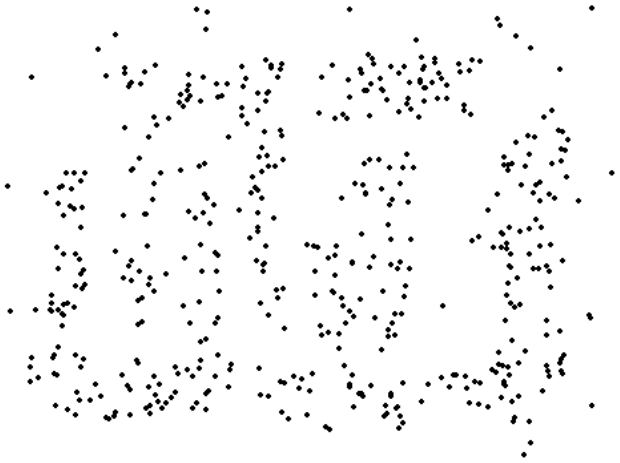
Sample Size



8000 points



2000 Points



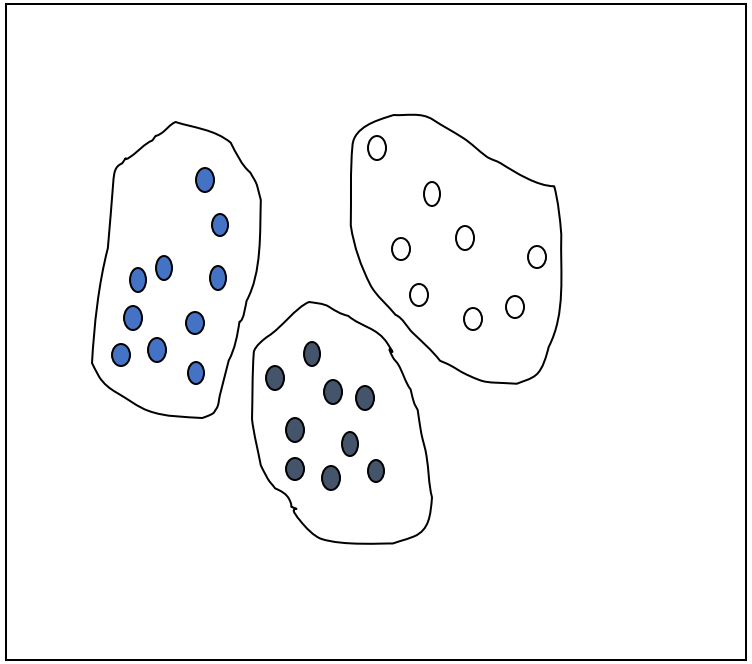
500 Points

Types of Sampling

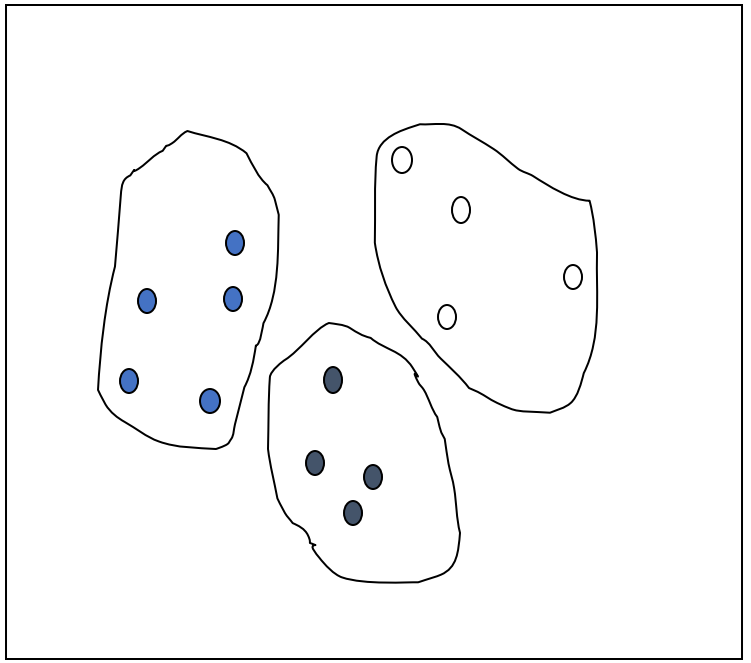
- Simple Random Sampling
  - There is an **equal probability** of selecting any particular item
  - **Sampling without replacement**
    - As each item is selected, it is removed from the population
  - **Sampling with replacement**
    - Objects are not removed from the population as they are selected for the sample.
    - In sampling with replacement, the same object can be picked up more than once
- Stratified sampling
  - Split the data into several partitions; then draw random samples from each partition
  - Approximation of the percentage of each class
  - Suitable for distribution with peaks: each peak is a **layer**

Stratified Sampling

Raw Data



Cluster/Stratified Sample



Reduction of Dimensionality

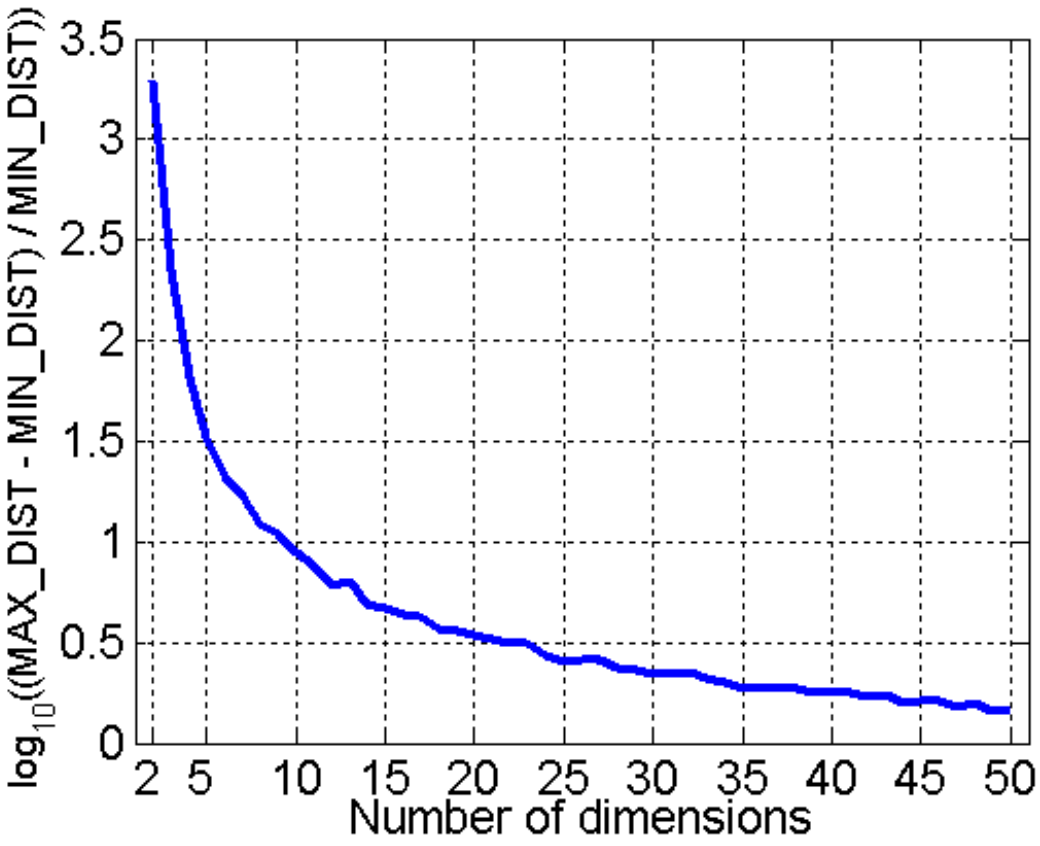
Selection of a subset of attributes that is as small as possible and sufficient for the data analysis.

- removing (more or less) **irrelevant** features
  - Contain **no information** that is **useful** for the data mining task at hand
  - **Example**: students' ID is often irrelevant to the task of predicting students' GPA
- removing **redundant** features
  - **Duplicate** much or all of the **information** contained in one or more other attributes
  - **Example**: purchase price of a product and the amount of sales tax paid

Curse of Dimensionality

- When dimensionality increases, data becomes increasingly sparse in the space that it occupies

- Definitions of density and distance between points, which are critical for clustering and outlier detection, become less meaningful



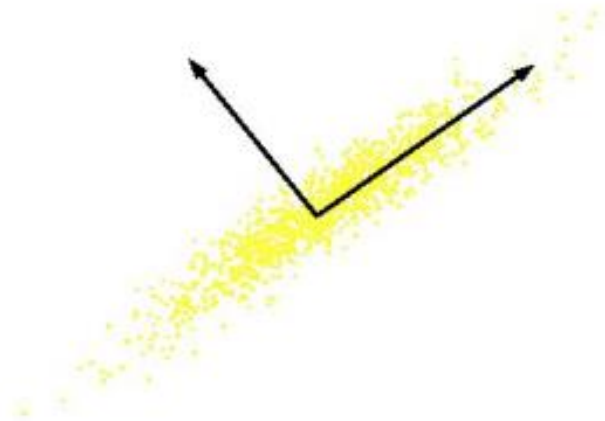
Dimensionality Reduction

- Purpose:
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise

- Techniques
  - **Principal Components Analysis (PCA)**
  - Singular Value Decomposition
  - Others: supervised and non-linear techniques

Principal Component Analysis

- The goal of PCA is to **find a new set of dimensions** (attributes or features) that better **captures the variability of the data.**
- The **first dimension** is chosen to capture as **much of the variability** as possible.
- The **second dimension** is orthogonal to the first and, subject to that constraint, captures as much of **the remaining variability** as possible, and so on.
- It is a **linear transformation** that chooses a new coordinate system for the data set

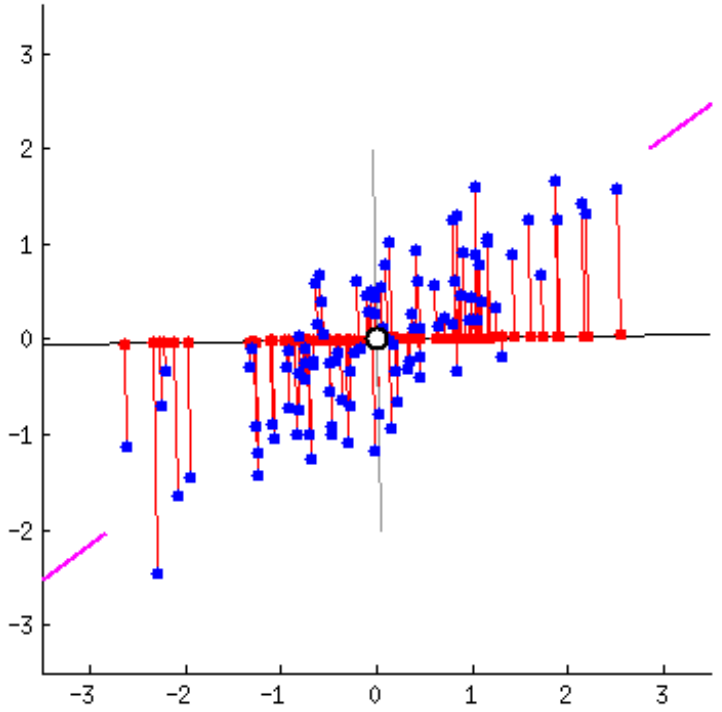


Steps of the approach

- **Step 1:** Standardize the dataset.
- **Step 2:** Calculate the covariance matrix for the features in the dataset.
- **Step 3:** Calculate the eigenvalues and eigenvectors for the covariance matrix.
- **Step 4:** Sort eigenvalues and their corresponding eigenvectors and pick k eigenvalues and form a matrix of eigenvectors.
- **Step 5:** Transform the original matrix.

How to construct PC?

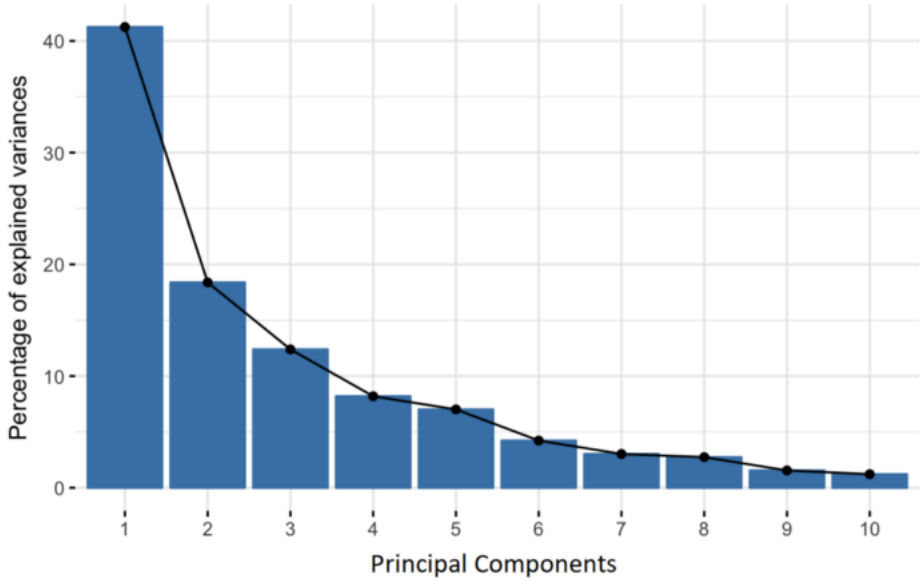
- The first principal component accounts for the **largest possible variance** in the data set.
- I want to fix the black line such that the spread on them of the red points, i.e., the original points projected on the black line, is maximised.
- The second principal component is calculated in the same way, with the condition that **it is uncorrelated with the first principal component** and that it accounts for the **next highest variance.**



- More details will be provided in DM2

Identify the Principal Components

Given 10-dimensional data you get 10 principal components but only the first PCs capture most of the variability of the data



Discarding the components with low information and considering the remaining components as your new variables.

Removing Irrelevant/Redundant Features

- For **removing irrelevant features**, a **performance measure** is needed that indicates how well a feature or subset of features performs w.r.t. the considered data analysis task

- For removing **redundant features**, either a performance measure for subsets of features or a correlation measure is needed.

Reduction of Dimensionality

Filter Methods

- Selection after analyzing the **significance** and **correlation** with other attributes
- Selection is independent of any data mining task
- The operation is a pre-processing

Wrapper Methods

- Selecting the top-ranked features using as reference a DM task
- Incremental Selection of the “best” attributes
- “Best” = with respect to a specific measure of statistical significance (e.g.: information gain)

Embedded Methods

- Selection as part of the data mining algorithm
- During the operation of the DM algorithm, the algorithm itself decides which attributes to use and which to ignore (e.g. Decision tree)

Wrapper Feature Selection Techniques

- **Selecting the top-ranked features**: Choose the features with the best evaluation when single features are evaluated.
- **Selecting the top-ranked subset:** Choose the subset of features with the best performance. This requires exhaustive search and is impossible for larger numbers of features. (For 20 features there are already more than one million possible subsets.)
- **Forward selection:** Start with the empty set of features and add features one by one. In each step, add the feature that yields the best improvement of the performance.
- **Backward elimination:** Start with the full set of features and remove features one by one. In each step, remove the feature that yields to the least decrease in performance.

Feature Creation

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes

- Three general methodologies:
  - Feature construction
    - Domain-dependent
    - Example: dividing mass by volume to get density
  - Feature Projection
    - Transforms the data from the high-dimensional space to a space of fewer dimensions

Feature Creation: features needed for task

Find the best workers in a company.

- Attributes :
  - the tasks, a worker has finished within each month,
  - the number of hours he has worked each month,
  - the number of hours that are normally needed to finish each task.
- These attributes *contain* information about the efficiency of the worker.
- But instead using these three "raw" attributes, it might be more useful to define a new attribute *efficiency*.
- $\text{efficiency} = \frac{\text{hours actually spent to finish the tasks}}{\text{hours normally needed to finish the tasks}}$

Feature Creation: features needed for task

- **Task**: face recognition in images
- Images are only set of contiguous pixels
- They are not suitable for many types of classification algorithms
- Process to provide **higher level features**
  - presence or absence of certain types of areas that are highly correlated with the presence of human faces
  - a much broader set of classification techniques can be applied to this problem

Feature Projection or Extraction

- It transforms the data in the high-dimensional space to a space of fewer dimensions.
- The data transformation may be linear, or nonlinear.
- Approaches:
  - Principal Component Analysis (PCA)
  - Singular Value Decomposition (SVD)
  - Non-negative matrix factorization (NMF)
  - Linear Discriminant Analysis (LDA)
  - Autoencoder

Data Cleaning

- How to handle anomalous values
- How to handle outliers
- Data Transformations

Anomalous Values

- **Missing values**
  - NULL, ?
- **Unknown Values**
  - Values without a real meaning
- **Not Valid Values**
  - Values not significant

Manage Missing Values

1. Elimination of records
2. Substitution of values

Note: it can influence the original distribution of numerical values

- Use mean/median/mode
- Estimate missing values **using the probability distribution** of existing values
- Data Segmentation and using mean/mode/median of each **segment**
- Data Segmentation and using **the probability distribution within the segment**
- Build a model of **classification/regression** for computing missing values

Discretization

- Discretization is the process of converting a continuous attribute into an ordinal attribute
  - A potentially infinite number of values are mapped into a small number of categories
  - Discretization is commonly used in classification
  - Many classification algorithms work best if both the independent and dependent variables have only a few values

  - When you make a histogram for a continuous attribute you are discretizing your data!

Discretization: Advantages

- Hard to understand the optimal discretization
  - We should need the real data distribution
- Original values can be **continuous** and **sparse**
- Discretized data can be **simple** to be interpreted
- Data distribution after discretization can have a **Normal shape**
- Discretized data can be too much **sparse yet**
  - Elimination of the attribute

Unsupervised Discretization

- Characteristics:
  - No label for the instances
  - The number of classes is unknown

- Techniques of *binning*:
  - **Natural binning** → Intervals with the same width
  - **Equal Frequency binning** → Intervals with the same frequency
  - **Statistical binning** → Use statistical information (Mean, variance, Quartile)

Discretization of Quantitative Attributes

Solution: each value is replaced by the interval to which it belongs.

height: 0-150cm, 151-170cm, 171-180cm, >180c

weight: 0-40kg, 41-60kg, 60-80kg, >80kg

income: 0-10ML, 11-20ML, 20-25ML, 25-30ML, >30ML

CID	**height**	**weight**	**income**
1	151-171	60-80	>30
2	171-180	60-80	20-25
3	171-180	60-80	25-30
4	151-170	60-80	25-30

Problem: the discretization may be useless (see **weight**).

How many groups (or classes)?

- If too few

  $\Rightarrow$Loss of information on the distribution

- If too many

  => Dispersion of values and does not show the form of distribution

- The optimal number of classes is function of N elements (Sturges, 1929)

$$C = 1 + \frac{10}{3}\log_{10}(N)$$

- The optimal width of the classes depends on the variance and the number of data (Scott, 1979)

$$h = \frac{3{,}5 \cdot s}{\sqrt{N}}$$

Binarization

- Binarization maps a continuous or categorical attribute into one or more binary variables

- Typically used for association analysis

- Often convert a continuous attribute to a categorical attribute and then convert a categorical attribute to a set of binary attributes
  - Association analysis needs asymmetric binary attributes
  - Examples: eye color and height measured as {low, medium, high}

Binarization

n = $\log_2$(m) binary digits are required to represent m integers.

It can generate some correlations

- **One variable for each possible value**
- Only presence or absence
- Association Rules requirements

Table 2.5. Conversion of a categorical attribute to three binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3
awful	0	0	0	0
poor	1	0	0	1
OK	2	0	1	0
good	3	0	1	1
great	4	1	0	0

Table 2.6. Conversion of a categorical attribute to five asymmetric binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3	x_4	x_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
OK	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

Data Transformation: Motivations

- Data with errors and incomplete
- Data not adequately distributed
  - Strong asymmetry in the data
  - Many peaks
- Data transformation can reduce these issues

Attribute Transformation

- An attribute transform is a function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
  - Simple functions: x^k, log(x), e^x, |x|
  - Normalization
    - Refers to various techniques to adjust to differences among attributes in terms of frequency of occurrence, mean, variance, range
    - Take out unwanted, common signal, e.g., seasonality
  - In statistics, standardization refers to subtracting off the means and dividing by the standard deviation

Properties of Trasformation

- Define a transformation T on the attribute X:

$$Y = \mathrm{T}(X)$$

such that :

  - Y preserve the **relevant** information of X
  - Y eliminates at least one of the problems of X
  - Y is more **useful** of X

Transformation Goals

- **Main goals:**
  - stabilize the variances
  - normalize the distributions
  - Make linear relationships among variables

- **Secondary goals:**
  - simplify the elaboration of data containing features you do not like
  - represent data in a scale considered more suitable

Why linear correlation, normal distributions, etc?

- Many statistical methods require
  - linear correlations
  - normal distributions
  - the absence of outliers

- Many data mining algorithms have the ability to automatically treat **non-linearity** and **non-normality**
  - The algorithms work still better if such problems are treated

Normalizations

- min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- z-score normalization

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

Transformation Functions

- Exponential transformation

$$T_p(x) = \begin{cases} ax^p + b & (p \neq 0) \\ c \log x + d & (p = 0) \end{cases}$$

- with a, b, c, d and p real values
  - Preserve the order
  - Preserve some basic statistics
  - They are continuous functions
  - They are derivable
  - They are specified by simple functions

Better Interpretation

- Linear Transformation

1€ = 1936.27 Lit.

- *p=1, a= 1936.27 ,b =0*

$$T_p(x) = \begin{cases} ax^p + b & (p \neq 0) \\ c \log x + d & (p = 0) \end{cases}$$

ºC= 5/9(ºF -32)

- *p = 1, a = 5/9, b = -160/9*

Stabilizing the Variance

- **Logarithmic Transformation**

$$T(x) = c \log x + d$$

  - Applicable to positive values
  - Makes homogenous the variance in log-normal distributions
    - E.g.: normalize seasonal peaks

Logarithmic Transformation: Example

Bar	*Birra*	*Ricavo*
A	Bud	20
A	Becks	10000
C	Bud	300
D	Bud	400
D	Becks	5
E	Becks	120
E	Bud	120
F	Bud	11000
G	Bud	1300
H	Bud	3200
H	Becks	1000
I	Bud	135

2300	Mean
2883,3333	Scarto medio assoluto
3939,8598	Standard Deviation
5	Min
120	1° Quartile
350	Median
1775	2° Quartile
11000	Max

Data are sparse!!!

Logarithmic Transformation: Example

Bar	*Birra*	*Ricavo (log)*
A	Bud	1,301029996
A	Becks	4
C	Bud	2,477121255
D	Bud	2,602059991
D	Becks	0,698970004
E	Becks	2,079181246
E	Bud	2,079181246
F	Bud	4,041392685
G	Bud	3,113943352
H	Bud	3,505149978
H	Becks	3
I	Bud	2,130333768

Media	2,585697
Scarto medio assoluto	0,791394
Deviazione standard	1,016144
Min	0,69897
Primo Quartile	2,079181
Mediana	2,539591
Secondo Quartile	3,211745
Max	4,041393

Stabilizing the Variance

$$T(x) = ax^p + b$$

- **Square-root Transformation**
- $p = 1/c$, c integer number
  - To make homogenous the variance of particular distributions e.g., Poisson Distribution
- **Reciprocal Transformation**
  - $p < 0$
  - Suitable for analyzing time series, when the variance increases too much wrt the mean

DATA MINING 1
Data Similarity

Dino Pedreschi, Riccardo Guidotti

Revisited slides from Lecture Notes for Chapter 2 "Introduction to Data Mining", 2nd Edition by Tan, Steinbach, Karpatne, Kumar

UNIVERSITÀ DI PISA

Similarity and Dissimilarity

- **Similarity**
  - Numerical measure of how alike two data objects are.
  - Is higher when objects are more alike.
  - Often falls in the range [0,1]
- **Dissimilarity**
  - Numerical measure of how different are two data objects
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- **Proximity refers to a similarity or dissimilarity**

Similarity/Dissimilarity for one Attribute

p and *q* are the attribute values for two data objects.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Table 5.1. Similarity and dissimilarity for simple attributes

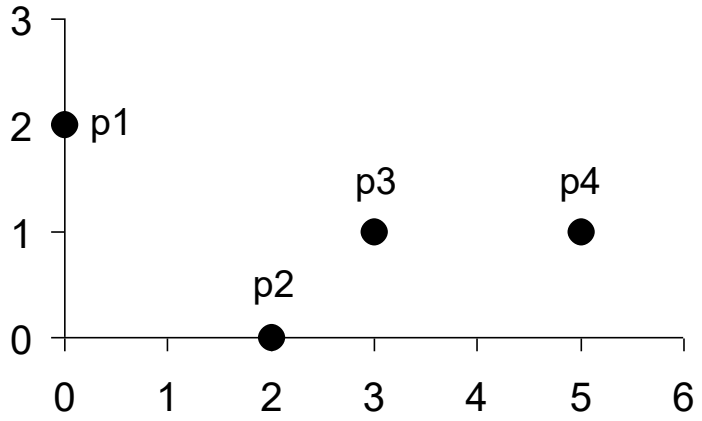
Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

where *n* is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects $\mathbf{x}$ and $\mathbf{y}$. Standardization is necessary, if scales differ.

- Standardization is necessary, if scales differ.

Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

Where r is a parameter, n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects $\boldsymbol{x}$ and $\boldsymbol{y}$.

Minkowski Distance: Examples

- $r = 1$. City block (Manhattan, taxicab, L_1 norm) distance.
  - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors
- $r = 2$. Euclidean distance
- $r \rightarrow \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.
  - This is the maximum difference between any component of the vectors
- Do not confuse *r* with *n*, i.e., all these distances are defined for all numbers of dimensions.

Minkowski Distance

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distance Matrix

Common Properties of a Distance

- Distances, such as the Euclidean, have some well-known properties.

  1. $d(\mathbf{x}, \mathbf{y}) \geq 0$ for all x and y and $d(\mathbf{x}, \mathbf{y}) = 0$ only if $\mathbf{x} = \mathbf{y}$. (Positive definiteness)
  2. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$. (Symmetry)
  3. $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ for all points $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$. (Triangle Inequality)

  where $d(\mathbf{x}, \mathbf{y})$ is the distance (dissimilarity) between points (data objects), $\mathbf{x}$ and $\mathbf{y}$.

- A distance that satisfies these properties is a metric

Common Properties of a Similarity

Similarities, also have some well-known properties.

1. $s(\mathbf{x}, \mathbf{y}) = 1$ (or maximum similarity) only if $\mathbf{x} = \mathbf{y}$.
   (does not always hold, e.g., cosine)
2. $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}$ and $\mathbf{y}$. (Symmetry)

where $s(\mathbf{x}, \mathbf{y})$ is the similarity between points (data objects), $\mathbf{x}$ and $\mathbf{y}$.

Binary Data

Categorical	insufficient	sufficient	good	very good	excellent
p1	**0**	0	1	0	0
p2	**0**	0	1	0	0
p3	**1**	0	0	0	0
p4	**0**	1	0	0	0

item	bread	butter	milk	apple	tooth-past
p1	**1**	1	0	1	0
p2	**0**	0	1	1	1
p3	**1**	1	1	0	0
p4	**1**	0	1	1	0

Similarity Between Binary Vectors

- Common situation is that objects, *p* and *q*, have only binary attributes
- Compute similarities using the following quantities

  M_{01} = the number of attributes where p was 0 and q was 1

  M_{10} = the number of attributes where p was 1 and q was 0

  M_{00} = the number of attributes where p was 0 and q was 0

  M_{11} = the number of attributes where p was 1 and q was 1

- Simple Matching and Jaccard Coefficients

  SMC = number of matches / number of attributes

  $= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

  J = number of 11 matches / number of not-both-zero attributes values

  $= (M_{11}) / (M_{01} + M_{10} + M_{11})$

SMC versus Jaccard: Example

p = 1 0 0 0 0 0 0 0 0 0

q = 0 0 0 0 0 0 1 0 0 1

M_{01} = 2 (the number of attributes where p was 0 and q was 1)

M_{10} = 1 (the number of attributes where p was 1 and q was 0)

M_{00} = 7 (the number of attributes where p was 0 and q was 0)

M_{11} = 0 (the number of attributes where p was 1 and q was 1)

SMC = $(M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00})$ = (0+7) / (2+1+0+7) = 0.7

J = (M_{11}) / $(M_{01} + M_{10} + M_{11})$ = 0 / (2 + 1 + 0) = 0

Document Data

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Cosine Similarity

- If d_1 and d_2 are two document vectors, then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2||$$

where • indicates vector dot product and $||d||$ is the length of vector d.

- Example:

d_1 = **3 2 0 5 0 0 0 2 0 0**

d_2 = **1 0 0 0 0 0 0 1 0 2**

$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

$||d_1|| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$||d_2|| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.245$

$\cos(d_1, d_2) = .3150$

Using Weights to Combine Similarities

- May not want to treat all attributes the same.
  - Use non-negative weights ω_k
  - $similarity(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^{n} \omega_k \delta_k s_k(\mathbf{x},\mathbf{y})}{\sum_{k=1}^{n} \omega_k \delta_k}$
- Can also define a weighted form of distance

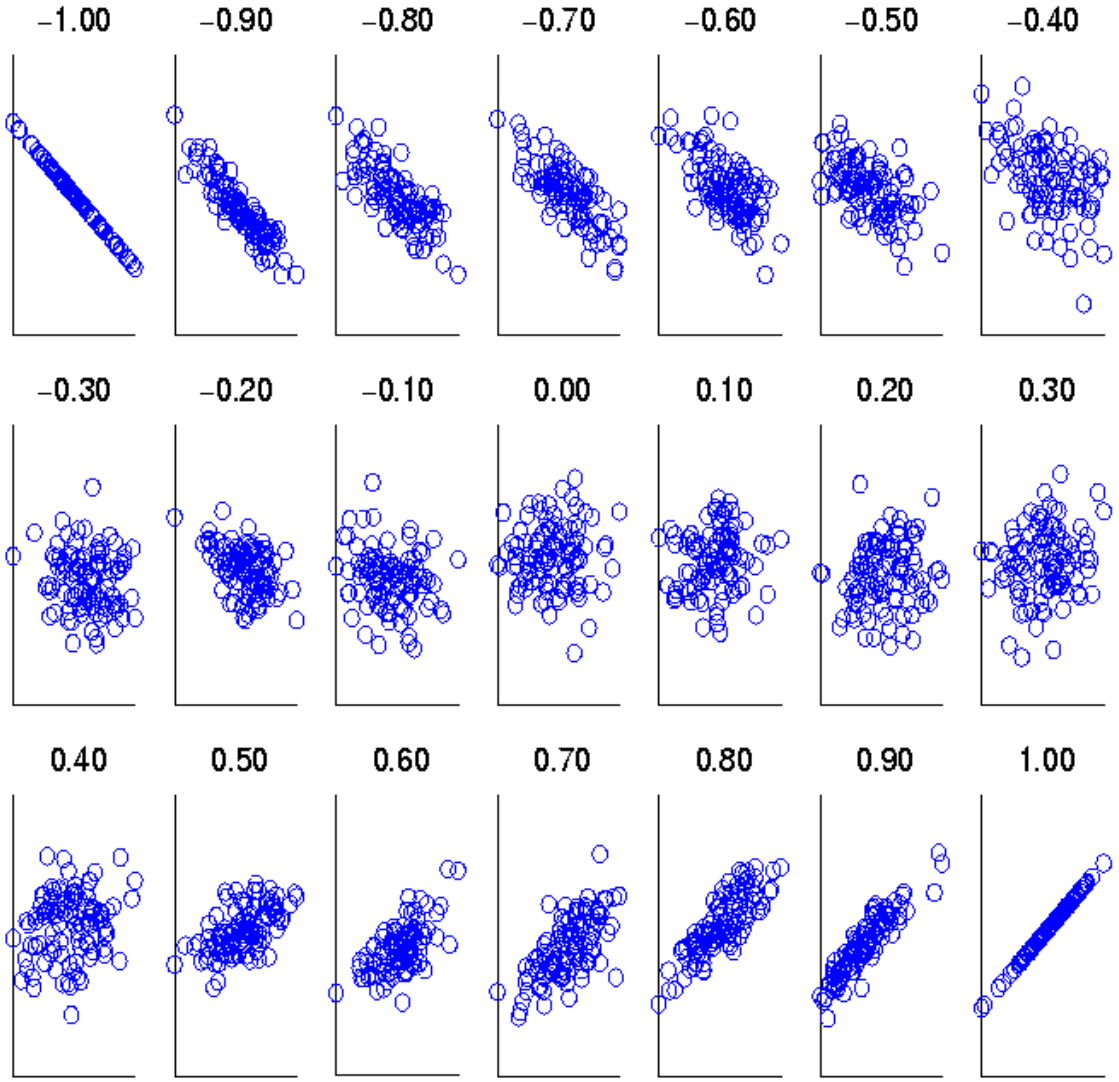
$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} w_k |x_k - y_k|^r \right)^{1/r}$$

Correlation

- Correlation measures the linear relationship between objects (binary or continuous)
- To compute correlation, we standardize data objects, p and q, and then take their dot product (covariance/standard deviation)

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{covariance}(\mathbf{x}, \mathbf{y})}{\text{standard\_deviation}(\mathbf{x}) * \text{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x \; s_y},$$

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

Mixed/Heterogenous Distances

- What happen if we have data with both continuous and categorical attributes?
- Option 1: discretize continuous attributes and use categorical distances like Jaccard, SMC, etc.
- ~~Option 2: pretend that categorical attributes can be represented with values and use continuous distances like Euclidean, Manhattan, etc.~~
- Option 3: define a new heterogenous distance like:
- $d(x, y) = n_{cat}/n \, d_{cat}(x_{cat}, y_{cat}) + n_{con}/n \, d_{con}(x_{con}, y_{con})$