INTRODUCTORY APPLIED MACHINE LEARNING

Yan-Fu Kuo
Dept. of Bio-industrial Mechatronics Engineering
National Taiwan University

Today:

- Types of data
- Data visualization

About Your Project…

- UCI machine learning repository: http://archive.ics.uci.edu/ml/datasets.html

Outline

- Goal of the lecture
- Types of data
- Data preprocessing
- Measures of similarity
- The Iris data set
- Descriptive statistics
- Visualization

Goals

- After this, you should be able to:
  - Understand data types and data acronyms
  - Calculate similarity between data points
  - Use basic descriptive statistics
  - Visualize 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 or feature
- A collection of attributes describe an object
  - Object is also known as 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

Attribute Values

- Attribute values are numbers or symbols assigned to an attribute
- Discrete attribute
  - Has only a finite or countably infinite set of values
  - Examples: zip codes or counts
  - Often represented as integer variables
- 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

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

Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

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

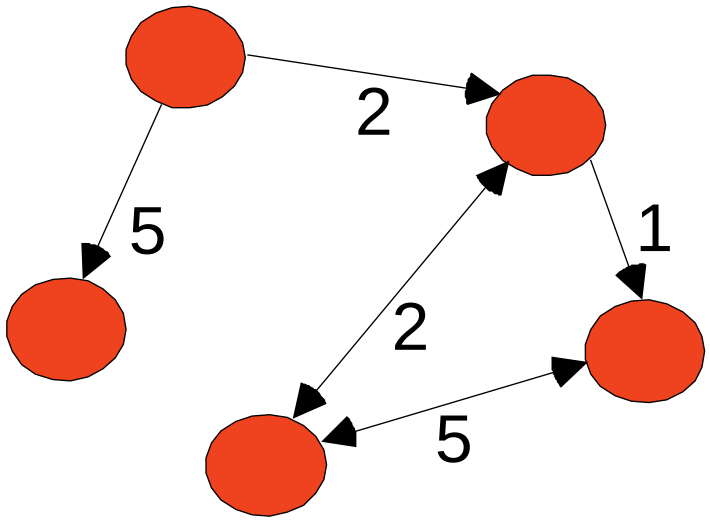
Record Data Matrix

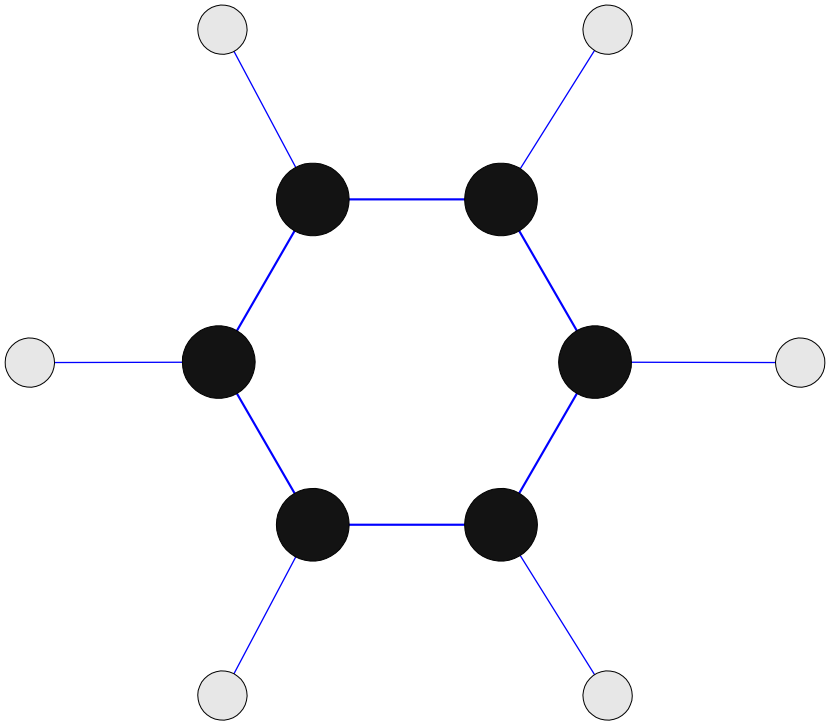
- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute
- Such data set can be represented by a matrix $\boldsymbol{A} \in \Re^{N \times M}$, where there are N rows, one for each object, and M 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

Graph Data

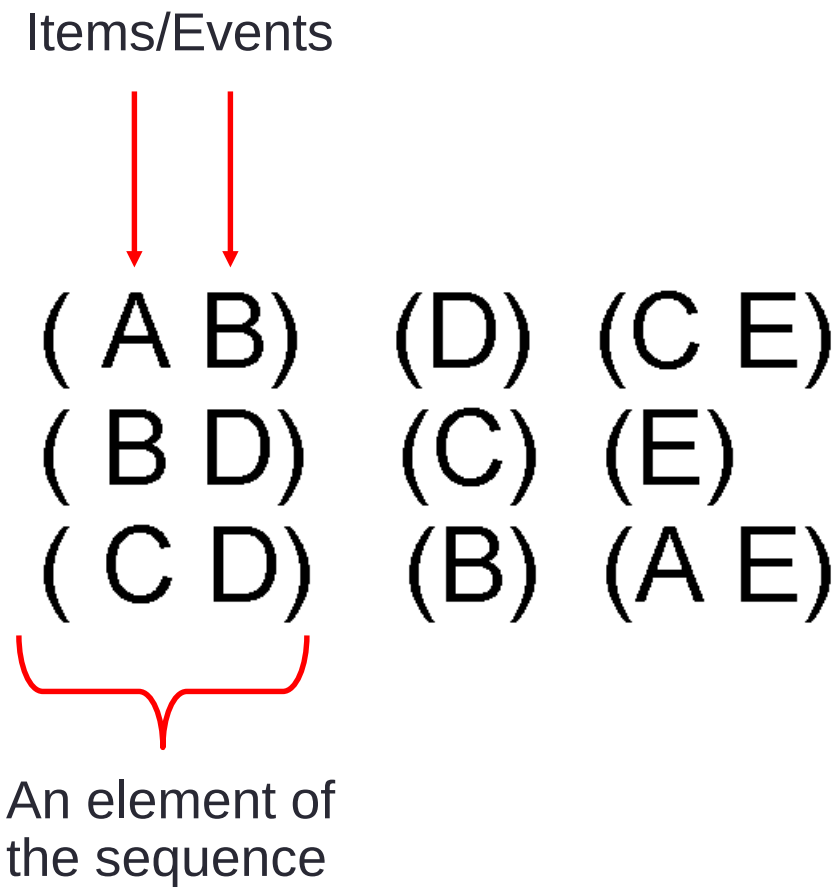
- Examples: Generic graph and chemical structure (C_6H_6)





Ordered Data

- Sequences of transactions



Ordered Data

- Genomic sequence data

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

Data Preprocessing – Sampling

- Sampling is the main technique employed for data selection
- It is often used for both the preliminary investigation of the data and the final data analysis
- Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming
- Sampling is used in machine learning because processing the entire set of data of interest is too expensive or time consuming

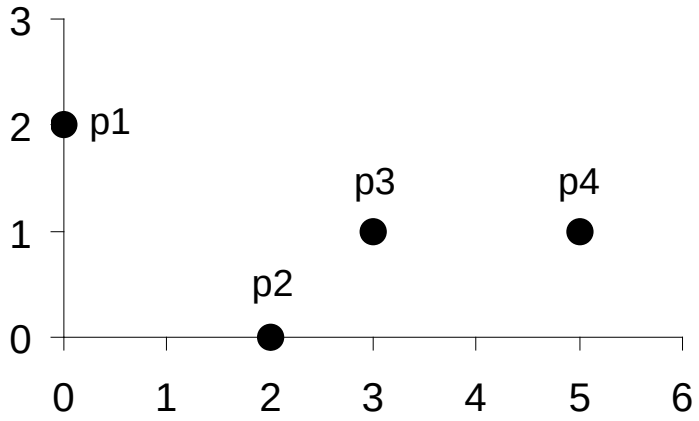
Measures of Similarity

- The Euclidean distance between two data points $\boldsymbol{p} = [p_k]$ and $\boldsymbol{q} = [q_k]$ is defined as

$$d(\boldsymbol{p}, \boldsymbol{q}) = \sqrt{\sum_{k=1}^{M} (p_k - q_k)^2} \in \Re$$

where $M \in \aleph$ is the number of dimensions (attributes) and p_k and q_k are, respectively, the kth attributes (components) of data objects $\boldsymbol{p}$ and $\boldsymbol{q}$

Examples of Euclidean Distance



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

Distance Matrix

	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

Minkowski Distance

- Minkowski distance is a generalization of Euclidean distance

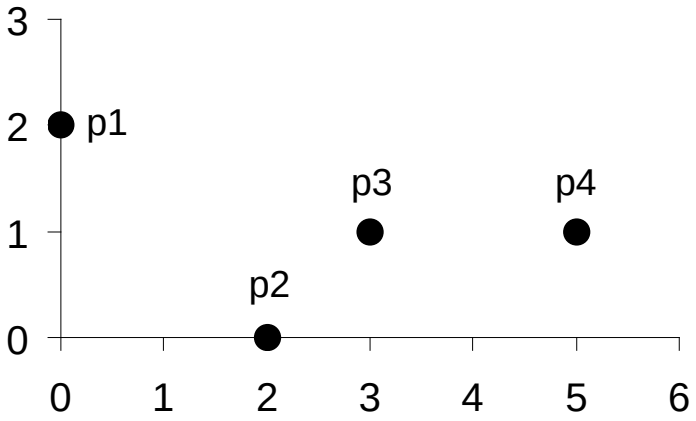
$$d(\boldsymbol{p}, \boldsymbol{q}) = \left(\sum_{k=1}^{M} |p_k - q_k|^r \right)^{\frac{1}{r}} \in \Re$$

where $r \in \aleph$ is a parameter, M is the number of dimensions (attributes), and p_k and q_k are, respectively, the kth attributes (components) data objects $\boldsymbol{p}$ and $\boldsymbol{q}$

Common Minkowski Distance

- $r = 1$: City block (L_1 norm)
- $r = 2$: Euclidean distance (L_2 norm)
- $r \to \infty$: “supremum” (L_∞ norm)

Examples of Minkowski Distance



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

Distance Matrix

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

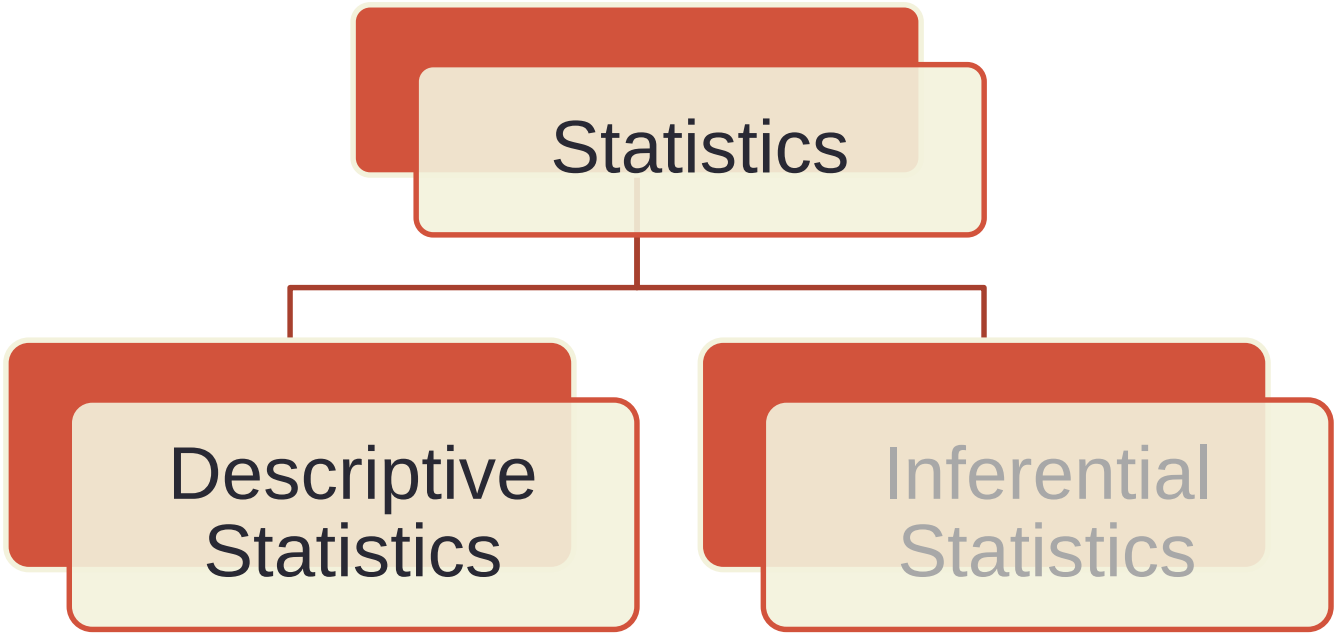
L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

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

Common Properties of a Distance

- Positive definiteness: $d(\boldsymbol{p}, \boldsymbol{q}) \geq 0$ for all $\boldsymbol{p}$ and $\boldsymbol{q}$ and $d(\boldsymbol{p}, \boldsymbol{q}) = 0$ only if $\boldsymbol{p} = \boldsymbol{q}$
- Symmetry: $d(\boldsymbol{p}, \boldsymbol{q}) = d(\boldsymbol{q}, \boldsymbol{p})$ for all $\boldsymbol{p}$ and $\boldsymbol{q}$
- Triangle inequality: $d(\boldsymbol{p}, \boldsymbol{t}) \leq d(\boldsymbol{p}, \boldsymbol{q}) + d(\boldsymbol{q}, \boldsymbol{t})$ for all points $\boldsymbol{p}$, $\boldsymbol{q}$, and $\boldsymbol{t}$, where $d(\boldsymbol{p}, \boldsymbol{q})$ is the distance (dissimilarity) between points $\boldsymbol{p}$ and $\boldsymbol{q}$
- A distance that satisfies these properties is a metric

Statistical Methodologies



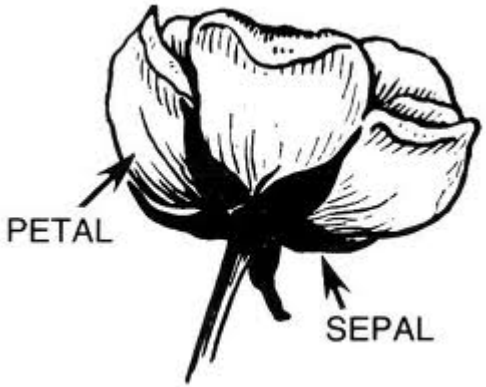
Numerical and graphical methods to look for patterns, to summarize the information in a data set

The Iris Data Set

- Can be obtained from the UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
- Matlab command:

```
>load fisheriris.mat
```

- From the statistician Douglas Fisher
- Three flower types (classes): Setosa, Virginica, Versicolour
- Four (non-class) attributes: Sepal width and length, Petal width and length



Mean, Median, and Variance

- The mean is the most common measure of the location of a set of points, though it is very sensitive to outliers

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

- The median is also commonly used

$$median(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}\left(x_{(r)} + x_{(r+1)}\right) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

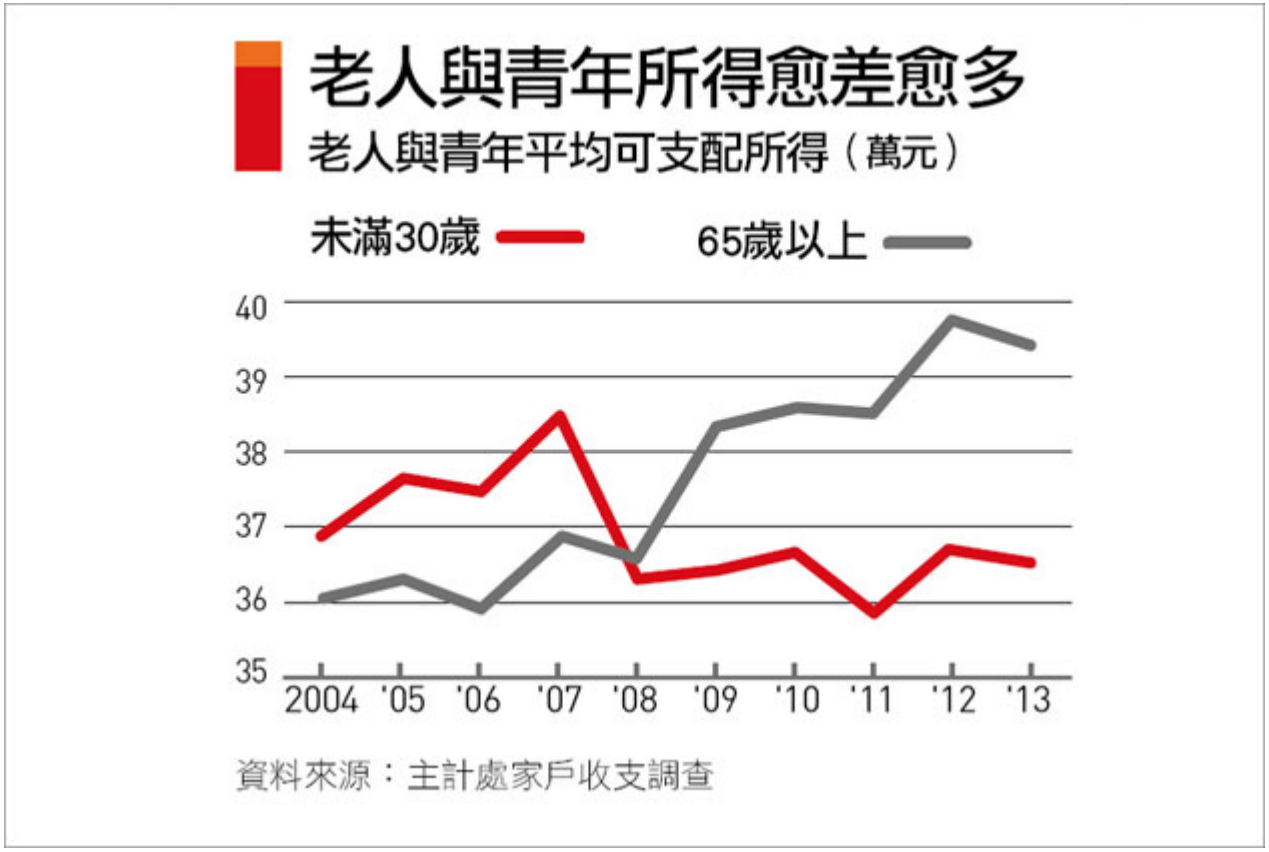
- The variance is the most common measure of the spread of a set of points

$$\text{Var}(x) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2$$

Visualization

- The conversion of data into a visual or tabular format, so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported
- Visualization of data is one of the most powerful and appealing techniques for data exploration
  - Humans have a well developed ability to analyze large amounts of information that is presented visually
  - Can detect general patterns and trends
  - Can detect outliers and unusual patterns

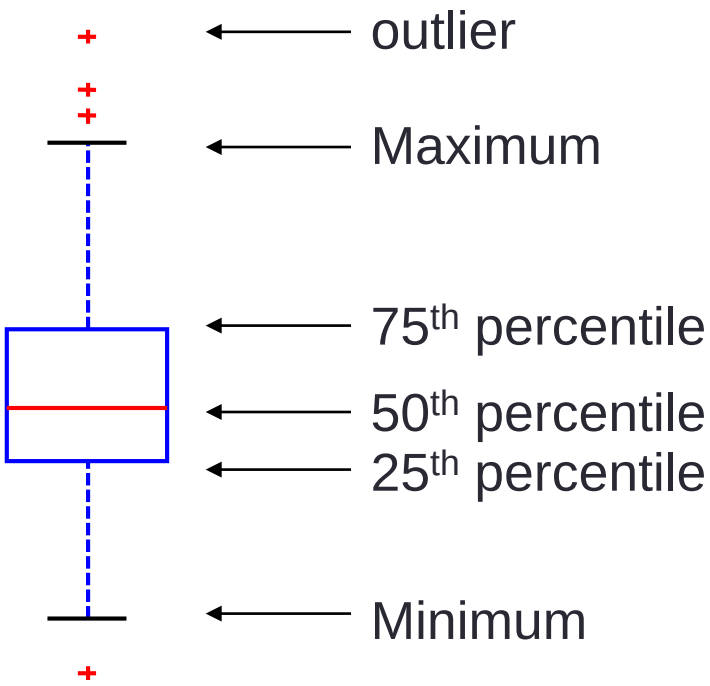
The Powerfulness of Visualization



天下雜誌 554期

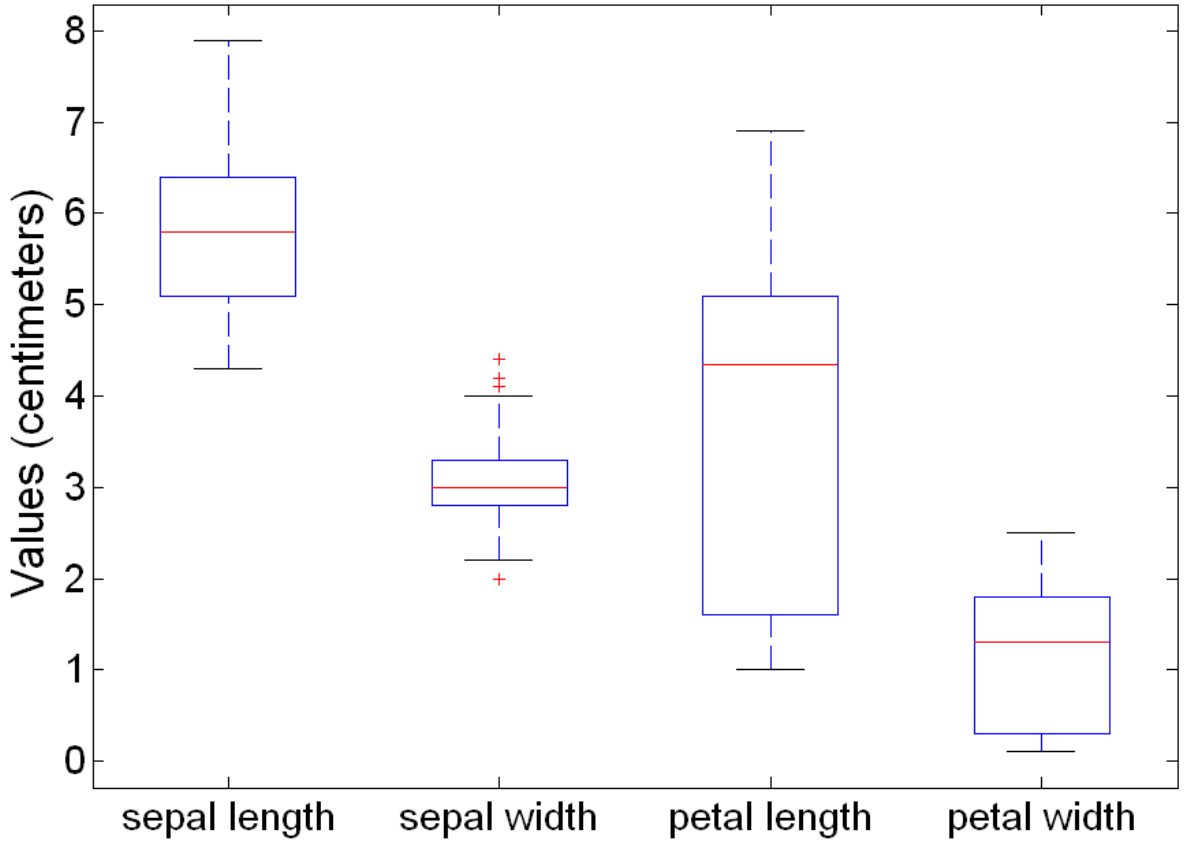
Visualization Techniques: Box Plots

- Invented by J. Tukey
- A way of displaying the distribution of data
- Following figure shows the basic part of a box plot



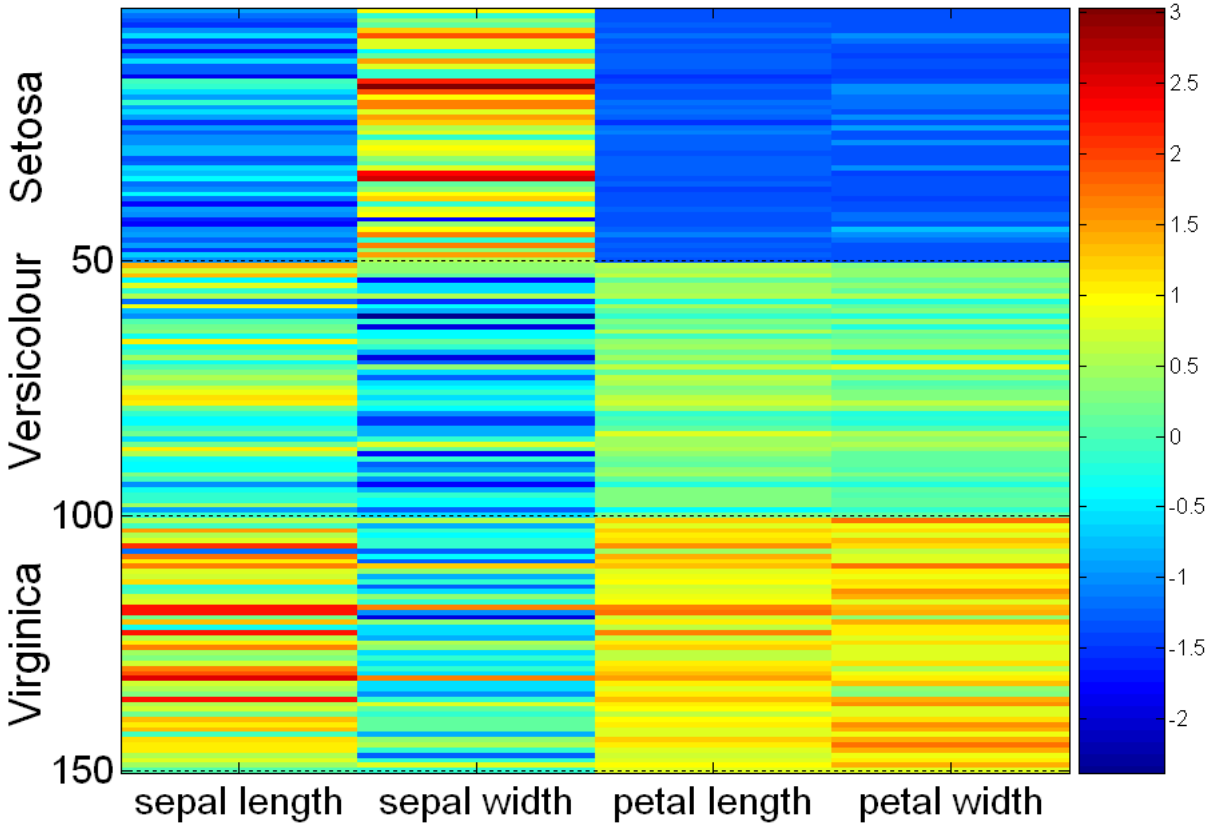
Example of Box Plots

- Box plots can be used to compare attributes



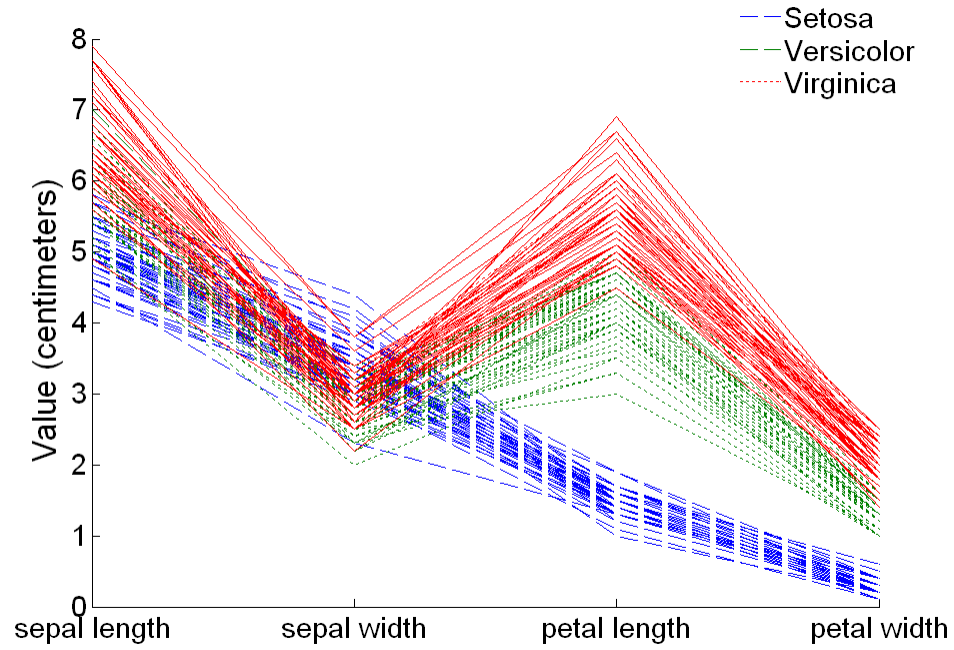
Visualization Techniques: Matrix Plots

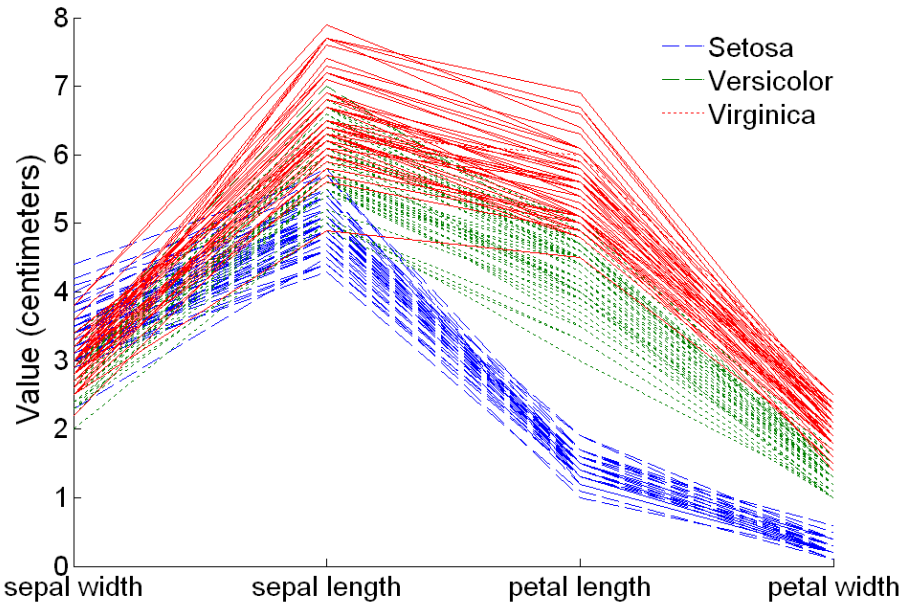
- Display three variables on a 2D plot
- This can be useful when objects are sorted according to class



Visualization Techniques: Parallel Coordinates

- Used to plot the attribute values of high-dimensional data
- The attribute values of each object are plotted as a point on each corresponding coordinate axis and the points are connected by a line





Summary

- For many machine-learning applications, a first step is identifying data type
- Norm is a metric to measure distance between data points
- Data visualization makes data analytics more effective

References

- P. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*