Data Exploration

Cam Tu Nguyen

阮锦绣

Software Institute, Nanjing University
nguyenct@lamda.nju.edu.cn
ncamtu@gmail.com

Outline

- Data Exploration
  - Data Objects and Attribute Types
  - Basic Statistical Descriptions of Data
  - Data Visualization
  - Measuring Data Similarity and Dissimilarity
  - Useful thing to know: Document representation and similarity measure between documents.

Data Objects and Attribute Types

- Data sets are made of data objects, such as person, customer, items.
  - Data objects can also be referred to as samples, instances, data points, etc.
- An attribute represents a characteristic or feature of a data object.
  - Example: eye color of a person, marital status, etc.
  - Attributes can also be called dimensions, features, variables.

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 Objects and Attribute Types

- **Types of attributes**
  - Nominal attribute: the values are symbols or names of things
    - Example: hair_color; marital_status
  - Binary attribute is a nominal attribute with only two categories or states (0 or 1)
    - Example: gender; medical_test_result (positive/negative)
    - Symmetric vs Asymmetric
  - Ordinal attribute is an attribute with possible values that have a meaningful order or **ranking** among them.
    - Example: grade (A+, A, A-, B+, etc.)
    - Ordinal attributes are useful for registering subjective assessments of qualities that are difficult to be measured objectively.

Data Objects and Attribute Types

- **Types of attributes**
  - Numeric Attribute is quantitative, i.e., it is a measurable quantity represented in integer or real values.
    - **Interval-scaled attributes**:
      - Example: temperature in Celsius, the unit of temperature is 1/100 of the difference between the melting temperature and the boiling temperature.
    - **Ratio-scaled attributes** is a numeric attribute with an inherent zero-point.
      - Example: temperature in Kelvin, length, time, time, counts

Data Objects and Attribute Types

- Properties of Attribute Values
  - The type of an attribute depends on which of the following properties it possesses:
    - Distinctness: $= \; \neq$
    - Order: $< \; >$
    - Addition: $+ \; -$
    - Multiplication: $\times \; \div$
  - Nominal attribute: distinctness.
  - Ordinal attribute: distinctness & order.
  - Interval attribute: distinctness, order and addition.
  - Ratio attribute: all 4 properties.

Data Objects and Attribute Types

- **Discrete vs Continuous Attribute**
  - Discrete attributes have finite or countably infinite number of values
    - Example: hair_color; smoker, medical_test, drink_size, zip codes, customer_id
  - Continuous attributes: if an attribute is not discrete, it is continuous.
    - Numeric attribute and continuous attribute are often used interchangeably in the literature.

Outline

- Data Exploration
  - Data Objects and Attribute Types
  - Basic Statistical Descriptions of Data
  - Data Visualization
  - Measuring Data Similarity and Dissimilarity
  - Useful thing to know: Document representation and similarity measure between documents.

Basic Statistical Descriptions of Data

- Basic statistical descriptions can be used to identify properties of the data.
  - Measures of central tendency
  - The Dispersion of the data.

Basic Statistical Descriptions of Data

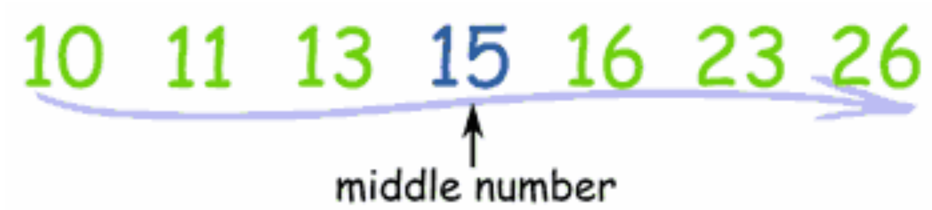
- **Measuring the Central Tendency: Mean, Median and Mode**
  - Mean: Let *X* be some attribute with *N* observed values $x_1, x_2, \ldots, x_N$

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

    - Means are not robust measurements, i.e. they are sensitive to noises or outliers. Trimmed means can be obtained by removing extreme values.
    - Means are used only for numeric attributes (or continuous attributes).

Basic Statistical Descriptions of Data

- **Measuring the Central Tendency: Mean, Median and Mode**
  - Median
    - More robust than means, can be applied to numeric, and may extend to use for ordinal attributes.



    - If the number of values is even, then the median is *the two middlemost values and any value in between*. If the attribute is numeric, the median is taken as *the average of the two middlemost values.*
    - The median is expensive to compute when we have a large number of observations.

Basic Statistical Descriptions of Data

- Approximate median on large dataset:
  - Suppose that data are grouped into **intervals with known frequencies.**
  - Let the interval that contains the median frequency be the **median interval:**

$$median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}} \right) width,$$

L_1 is the lower boundary of the median interval
N is the number of values in the entire data set.

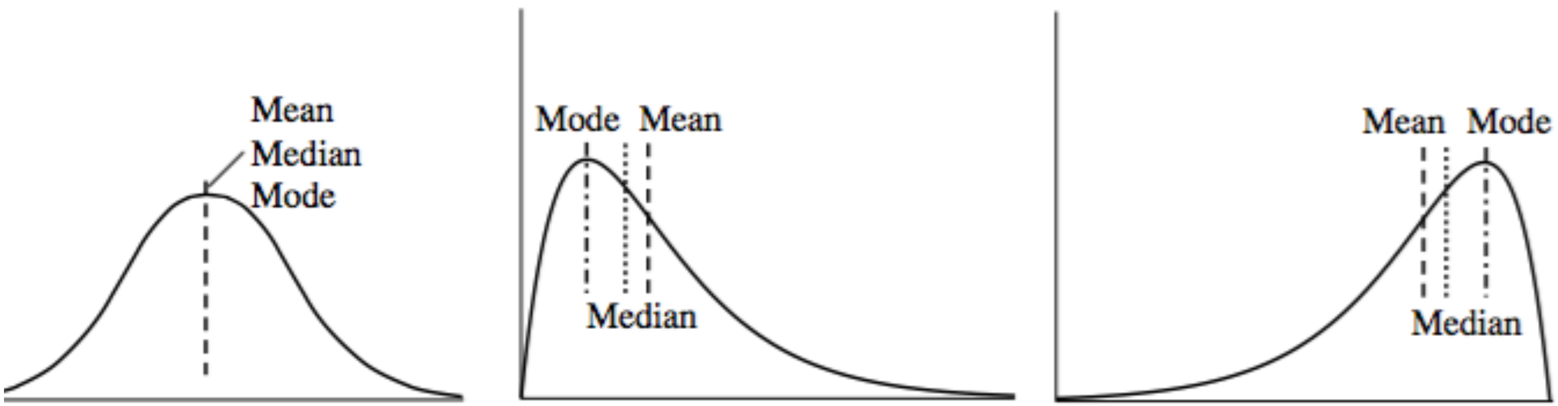
$(\sum freq)_l$ is the sum of the frequencies of all the intervals that are lower than the median interval.

Basic Statistical Descriptions of Data

- **Measuring the Central Tendency: Mean, Median and Mode**
  - Mode: the mode for a set of data is the value that occurs most frequently in the set.
    - Modes can be used for both nominal and numeric attributes.
    - Data sets with one, two, or three modes are respectively called **unimodal**, **bimodal**, and **trimodal**.

Basic Statistical Descriptions of Data

Unimodal: symmetric vs asymmetric



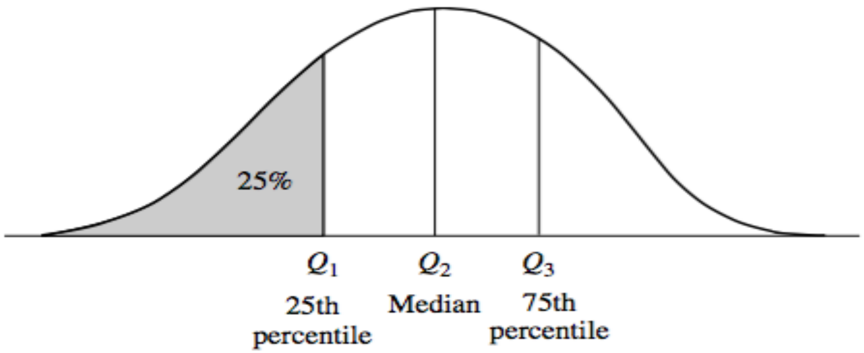
(a) Symmetric data (b) Positively skewed data (c) Negatively skewed data

Basic Statistical Descriptions of Data

- Measuring the Dispersion of Data:
  - Range, Quartiles, and Interquartile Range
    - Let $x_1, x_2, ..., x_N$ be a set of observations for some numeric attribute, X.
    - Range**:** the difference between the largest (max) and smallest (min)
    - Quantiles: sort values of X in increasing order.
      - The k-th q-quantiles for a given data distribution is the value x such that at most k/q of the data values are less than x, and at most (q-k)/q of the data values are more than x (0 < k < q)

Basic Statistical Descriptions of Data

- Measuring the Dispersion of Data:
  - Range, Quartiles, and Interquartile Range
    - **Percentiles**: 100-quantiles
    - **Quartiles**: 4-quanties.
    - **Interquartile range** (IQR):
      - IQR = Q3 – Q1
      - Q3: third quartile, is the 75^{th} percentile; and Q1 is the first quartile, or 25% percentile.



Basic Statistical Descriptions of Data

- Measuring the Dispersion of Data:
  - Variance: The variance of N observations for a numeric attribute X is:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \bar{x}^2,$$

where $\bar{x}$ is the mean value. The **standard deviation,** σ is the square root of the variance.

Outline

- Data Exploration
  - Data Objects and Attribute Types
  - Basic Statistical Descriptions of Data
  - Data Visualization
  - Measuring Data Similarity and Dissimilarity
  - Useful thing to know: Document representation and similarity measure between documents.

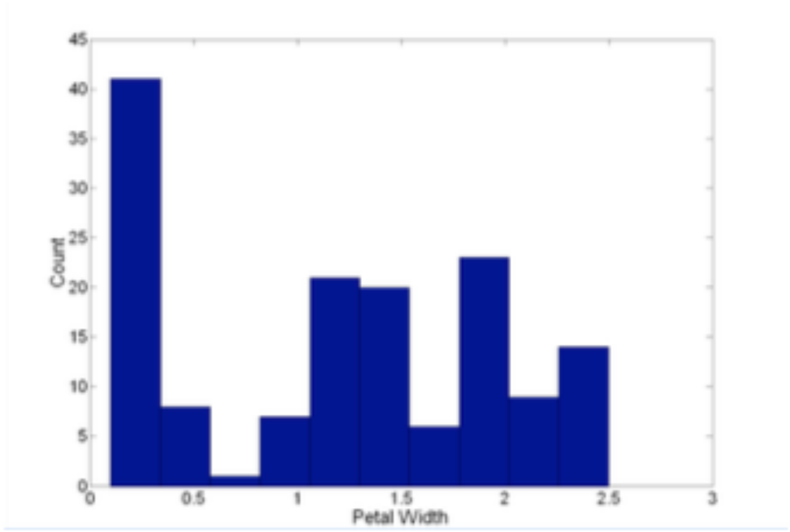
Data Visualization

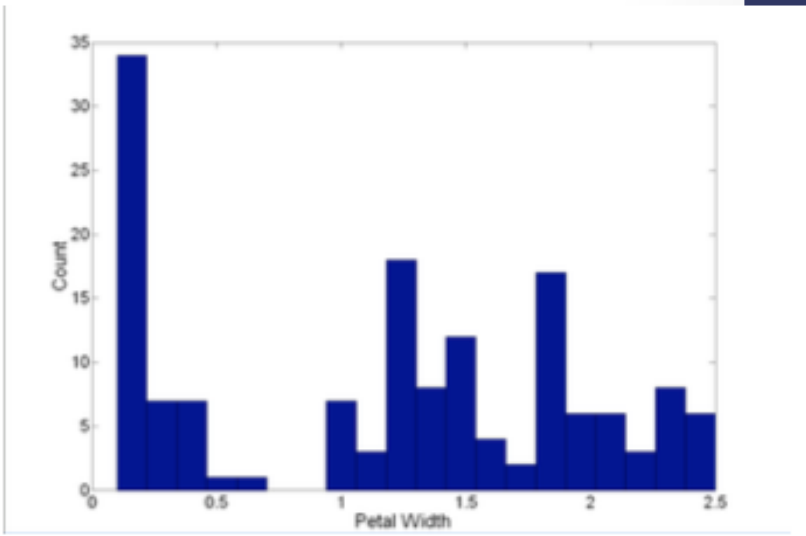
- **Iris Sample Data set**
  - Data Visualization is illustrated with the Iris Plant data set.
  - Can be obtained from the UCI Machine Learning Repository
    http://www.ics.uci.edu/~mlearn/MLRepository.html
  - From the statistician Sir. Ronal Fisher.
  - Three flower types (classes):
    - Setosa
    - Virginica
    - Versicolour
  - Four (non-class) attributes
    - Sepal width and length
    - Petal width and length

Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

Data Visualization

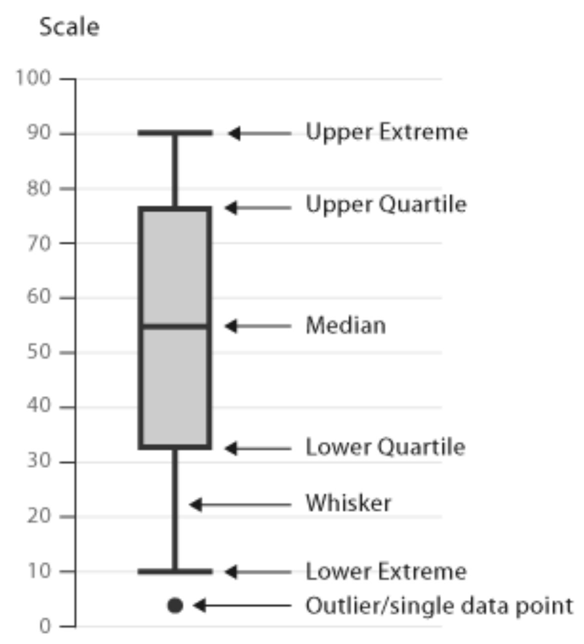
- Histogram
  - Usually shows the distribution of values of a single attribute.
  - Divide the values into bins and show a bar plot of the objects in each bin.
  - The height of each bar indicates the number of objects
  - Shape of histogram depends on the number of bins
- Example: Petal Width (10 and 20 bins, respectively)





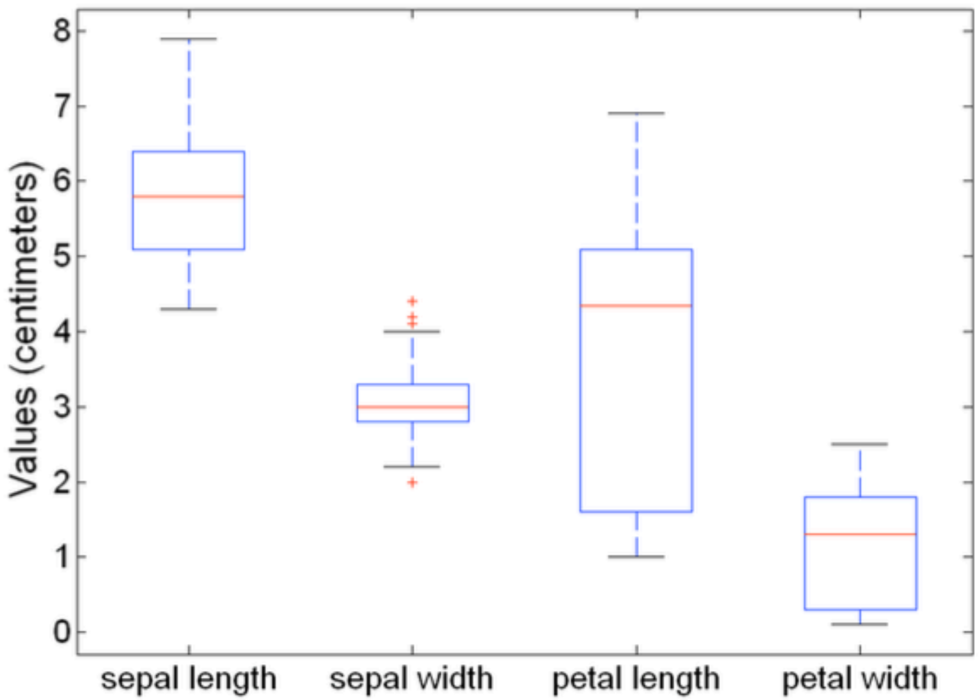
Data Visualization

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



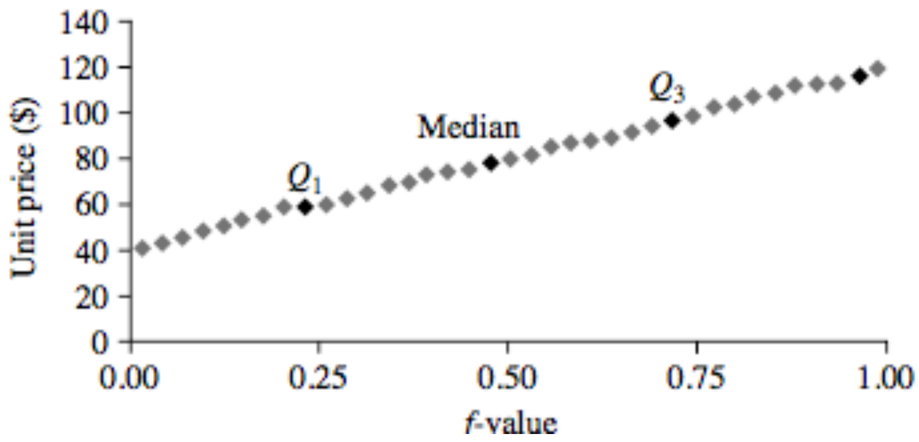
Example of Box Plots

- Box Plots can be used to compare attributes



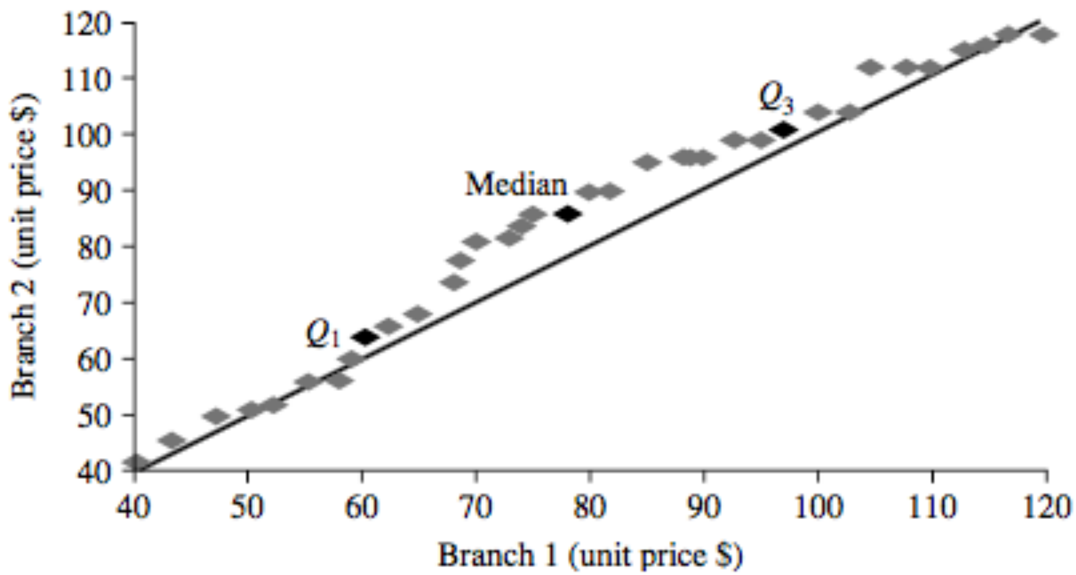
Data Visualization

- Quantile Plot
  - Let **x_i, for i=1,…, N** be the data sorted in increasing order.
  - Each data point is associated with a percentage f_i, which indicates that **f_i x 100%** of the data point is below x_i.
  - Note: 0.25 percentile is Q1, 0.5 is the median, …



Data Visualisation

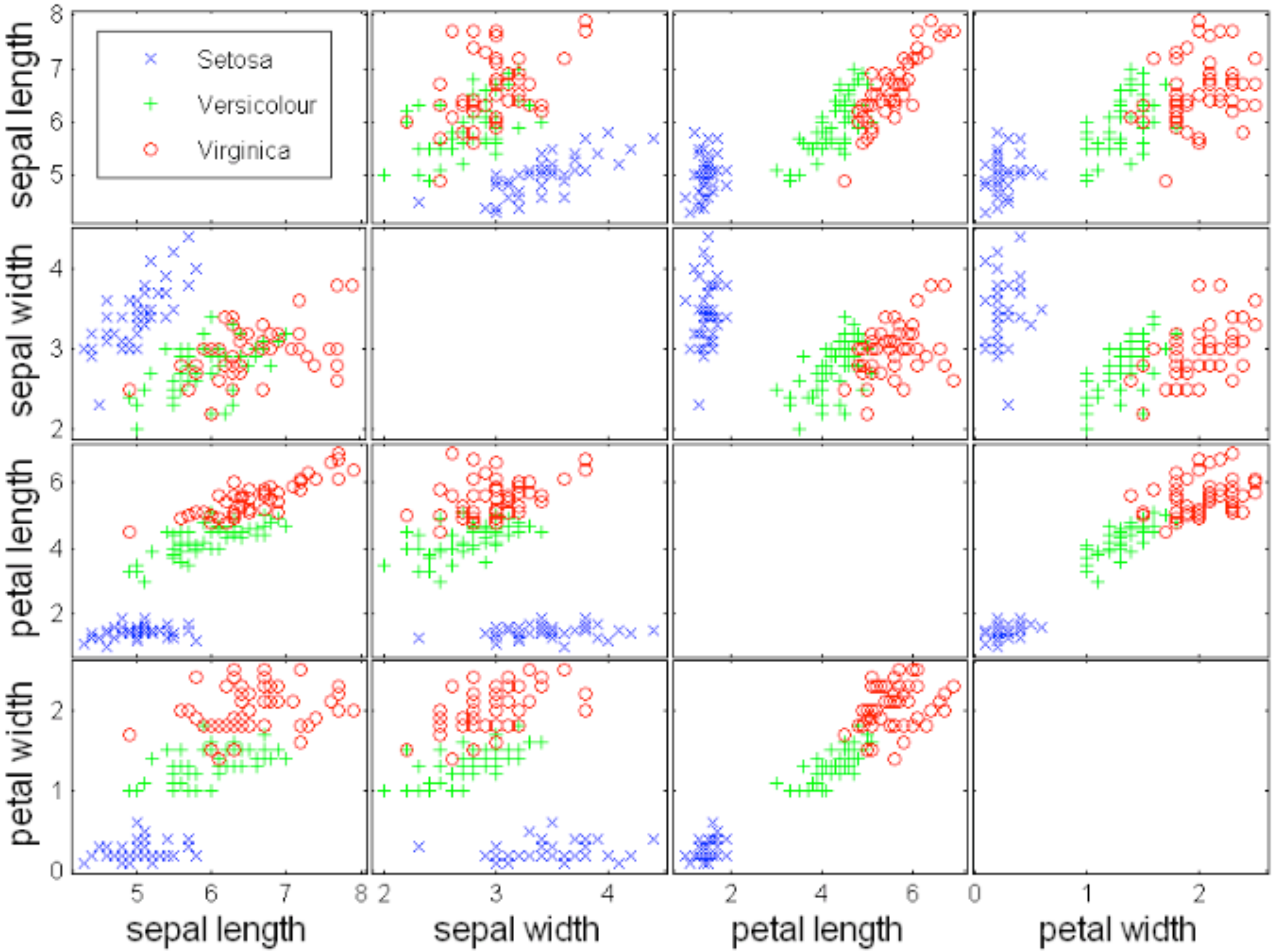
- A **quantile-quantile plot**, or **q-q plot**, graphs the quantiles of one univariate distribution against the corresponding quantiles of another.



Data Visualization

- **Scatter Plots**
  - Attributes values determine the position
  - Two-dimensional scatter plots most common, but can have three-dimensional scatter plots
  - Often additional attributes can be displayed by using the size, shape and color of the markers that represent the objects
  - It is useful to have arrays of scatter plots can compactly summarize the relationships of several pairs of attributes
    - See example on the next slide.

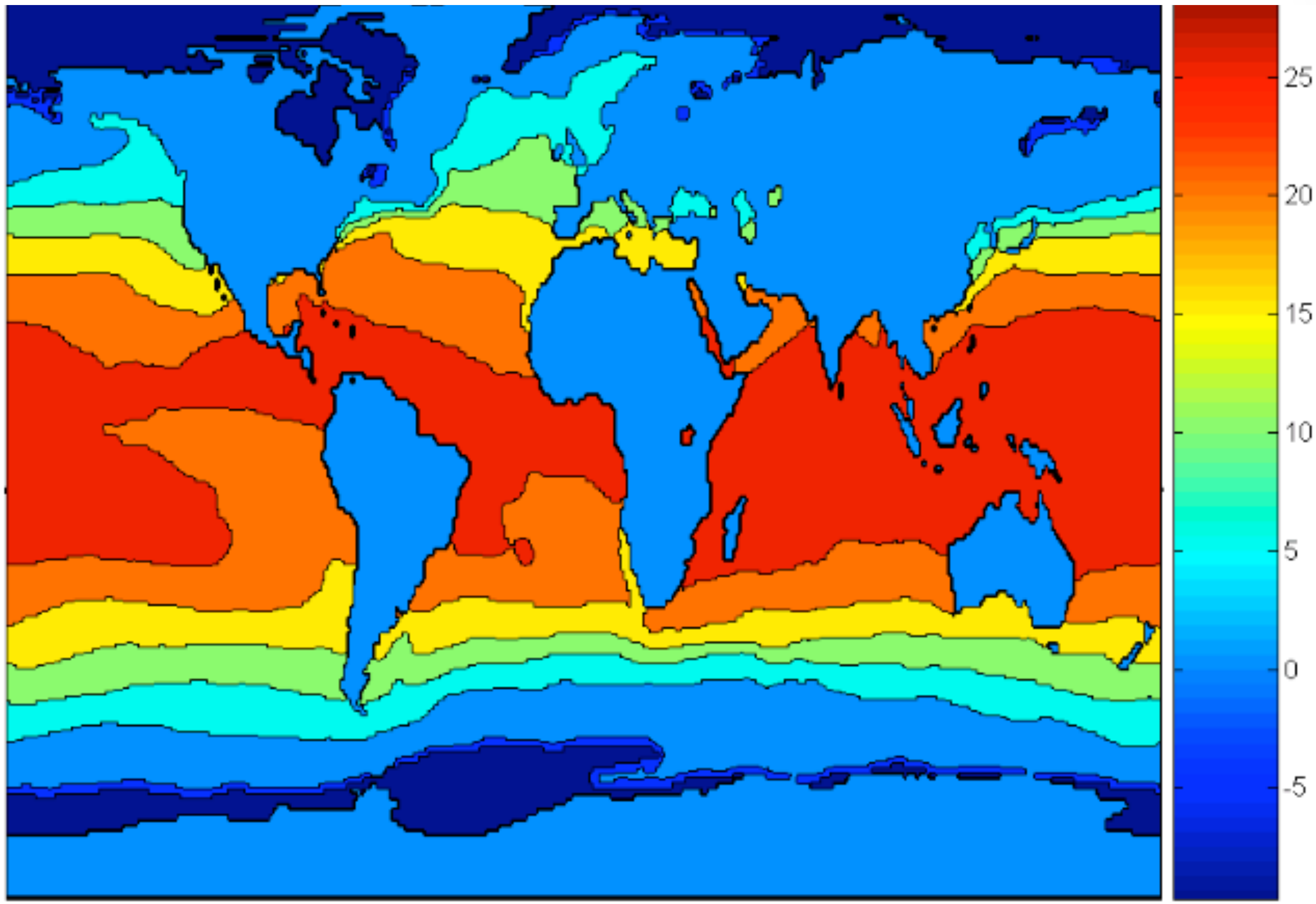
Scatter Plot Array of Iris Attributes



Data Visualization

- Contour plots
  - Useful when a continuous attribute is measured on a spatial grid.
  - They partition the plane into regions of similar values
  - The contour lines that form the boundaries of these regions connect points with equal values.
  - The most common example is contour maps of evaluation
  - Can also display temperature, rainfall, air pressure, etc.
    - An example for Sea Surface Temperature (SST) is provided on the next slide

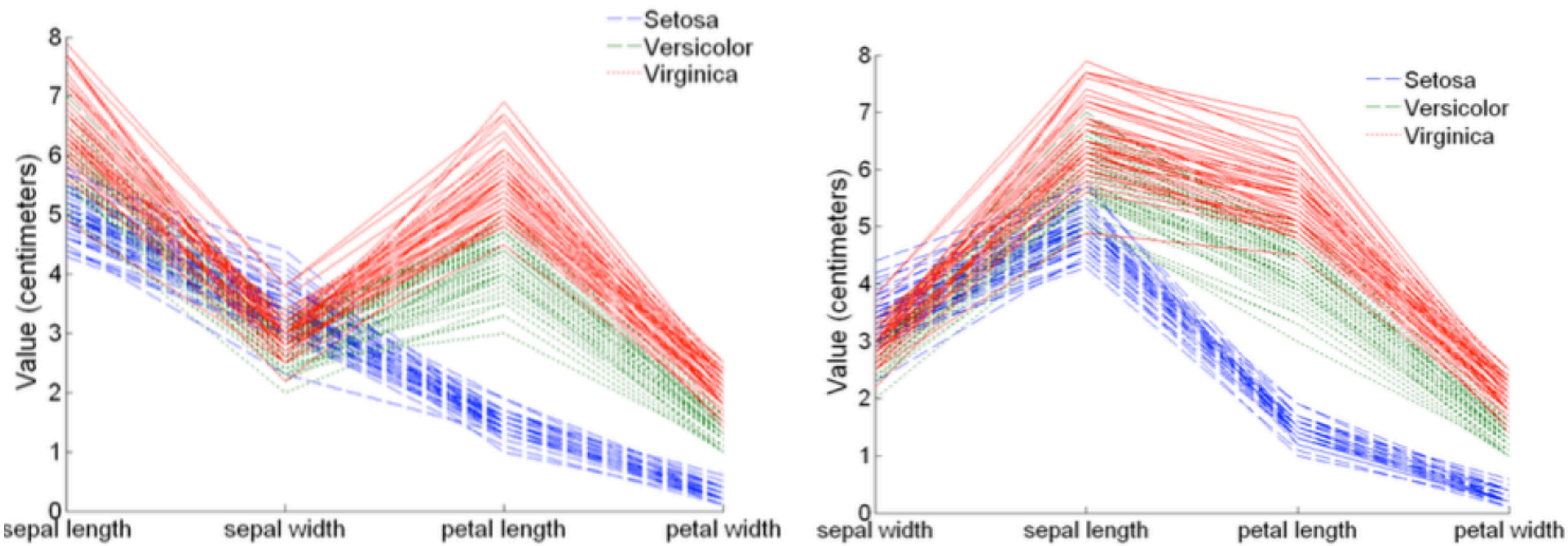
Contour Plot Example: SST Dec. 1998



Data Visualization

- 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

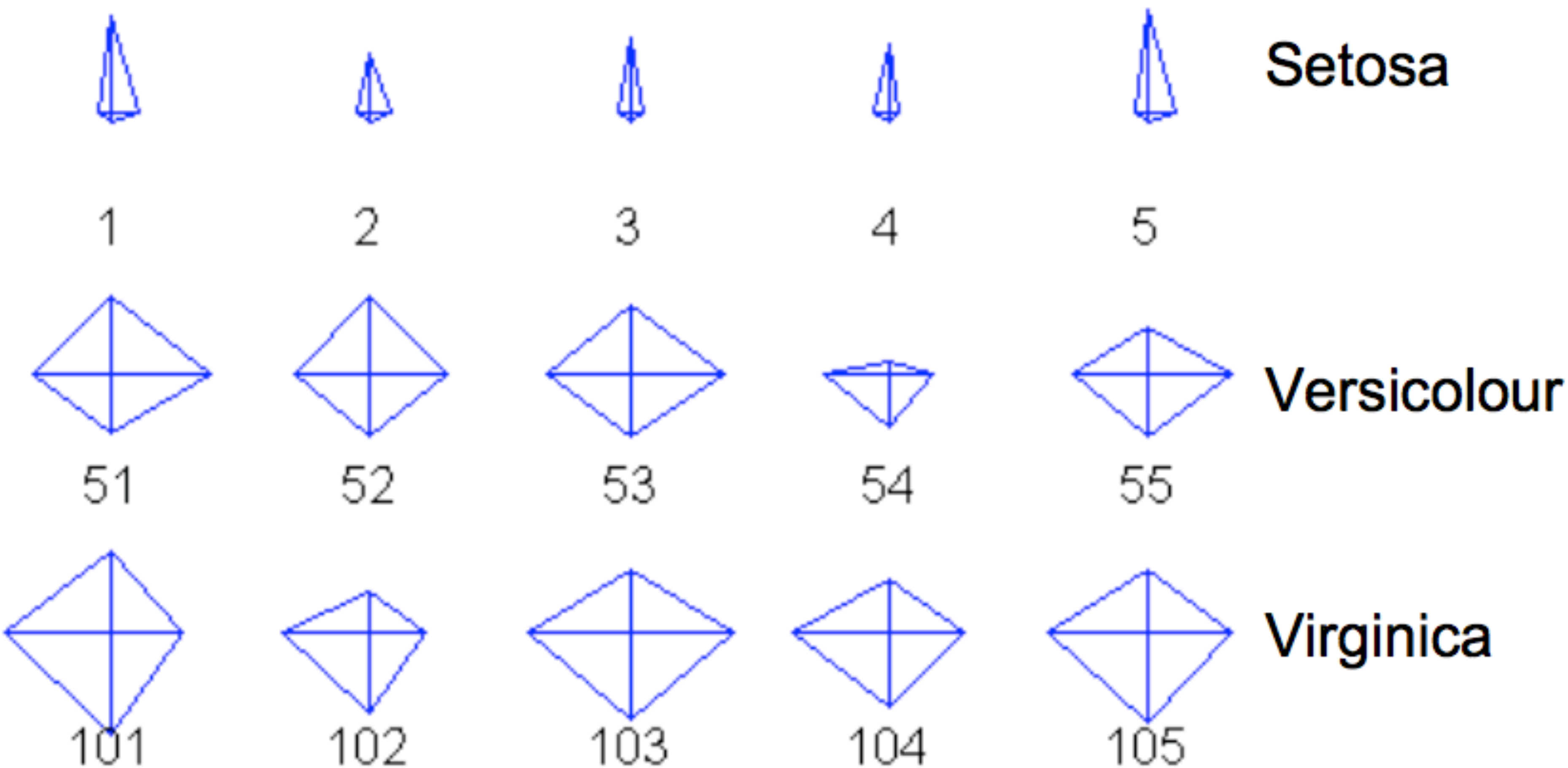
Parallel Coordinates Plots for Iris Data.



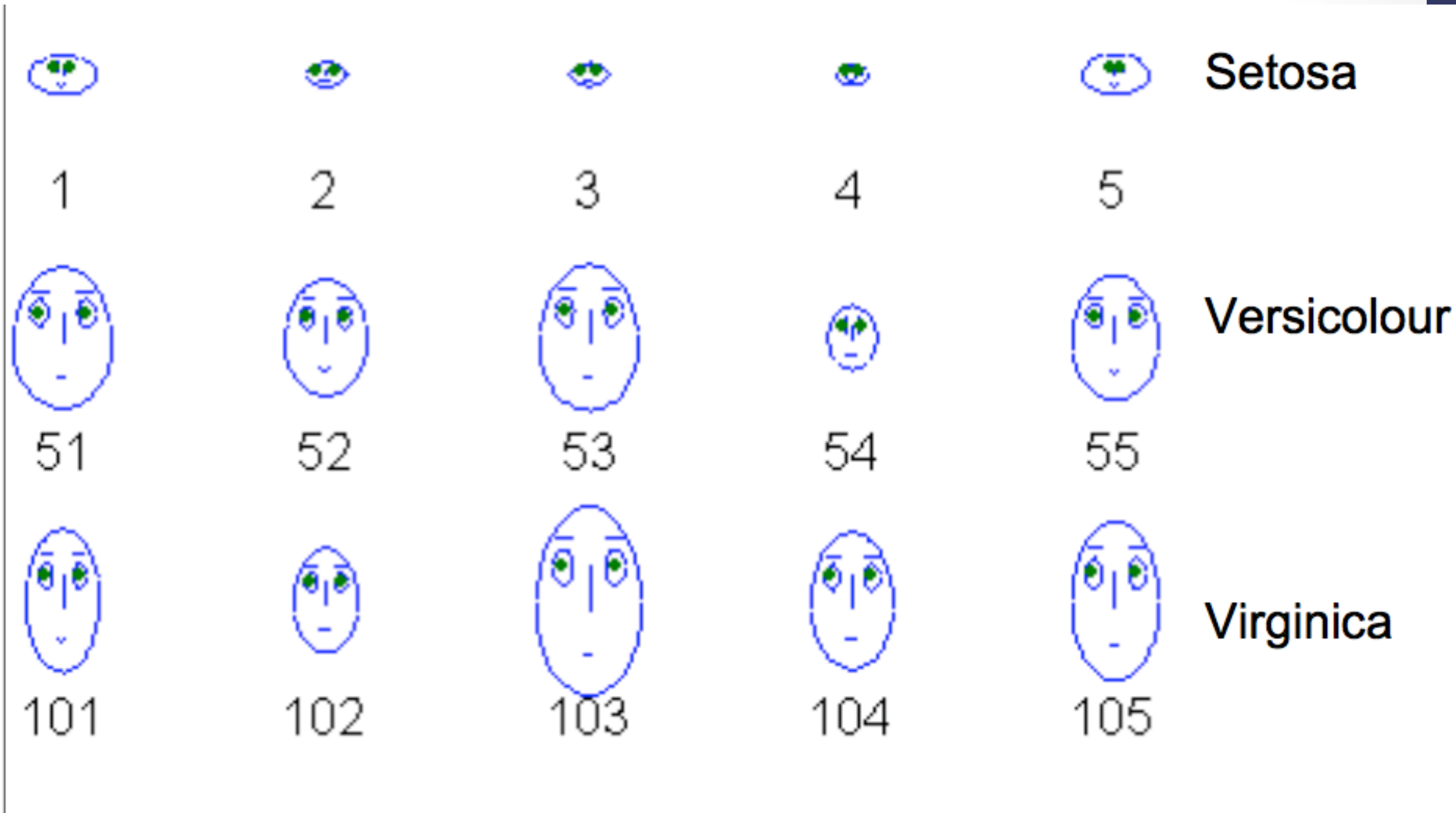
Other Visualization Techniques

- Star Plots
  - Similar approach to parallel coordinates, but axes radiate from a central point
  - The line connecting the values of an object is a polygon
- Chernoff Faces
  - Approach created by Herman Chernoff
  - This approach associates each attribute with a characteristic of a face
  - The values of each attribute determine the appearance of the corresponding facial characteristic
  - Each object becomes a separate face
  - Relies on human's ability to distinguish faces

Star Plots for Iris Data



Chernoff Faces for Iris Data



Outline

- Data Exploration
  - Data Objects and Attribute Types
  - Basic Statistical Descriptions of Data
  - Data Visualization
  - Measuring Data Similarity and Dissimilarity
  - Useful thing to know: Document representation and similarity measure between documents.

Data Matrix and Dissimilarity matrix

- Objects described by **multiple attributes**
  - An object (e.g. a person) x_i has **p** attributes

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$$

- A data set with **n** objects can be represented by a **n-by-p** **data matrix**:

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}.$$

Data Matrix and Dissimilarity matrix

- Dissimilarity matrix (or object-by object structure) stores a collections of proximities for all pairs of n objects, which is represented by **n-by-n matrix**:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Proximity Measures for Nominal Attributes

- A nominal attribute can take on **M** states
  - Example: **map_color** has 5 states (red, yellow, green, pink, and blue).

- Dissimilarity between two objects xi and xj with nominal attributes can be computed based ***on the ratio of mismatches***:

$$d(i, j) = \frac{p - m}{p}$$

p: the total number of attributes
m: the number of attributes **k** that $x_{ik} = x_{jk}$

Proximity Measures for Nominal Attributes

- Example:

A Sample Data Table Containing Attributes of Mixed Type

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

- Dissimilarity Matrix based on **test-1** attribute:

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}.$$

Proximity Measures for Binary Attributes

- Recall:
  - Binary attributes have 2 states (0, 1)
  - A binary attribute can be symmetric or asymmetric.

- Assume that all binary attributes have the same weights, we have 2x2 contingency table:

		Object j		
		1	0	sum
Object i	1	q	r	$q+r$
	0	s	t	$s+t$
	sum	$q+s$	$r+t$	p

q: the number of attributes that equal 1 for both i,j
r (s): the number of attributes that equal 1 for i (j) ; and 0 for j (i)
t: the number of attributes that equal 0 for both i and j.

Proximity Measures for Binary Attributes

- The dissimilarity between i and j is:

$$d(i,j) = \frac{r+s}{q+r+s+t}.$$

- The asymmetric binary dissimilarity (for asymmetric binary attributes):

$$d(i,j) = \frac{r+s}{q+r+s}.$$

- The similarity between the objects i and j can be computed as:

$$sim(i,j) = 1 - d(i,j)$$

(Jaccard coefficient)

Proximity Measures for Binary Attributes

- Example: Relational Table Where Patients Are Described by Binary Attributes

name	***gender***	***fever***	***cough***	***test-1***	***test-2***	***test-3***	***test-4***
Jack	M	Y	N	P	N	N	N
Jim	M	Y	Y	N	N	N	N
Mary	F	Y	N	P	N	P	N
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

忽略性别

- The distances between each pair of the three patients:

$$d(Jack, Jim) = \frac{1+1}{1+1+1} = 0.67,$$

$$d(Jack, Mary) = \frac{0+1}{2+0+1} = 0.33,$$

$$d(Jim, Mary) = \frac{1+2}{1+1+2} = 0.75.$$

Dissimilarity of Numeric Data

- Distance measures that are commonly used for numeric attributes. Let $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$and $x_j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ be two objects described by **p** numeric attributes.
- The Euclidean distance between i and j is:

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}.$$

- The Manhattan (or city block) distance:

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|.$$

- The Minkowski distance is the generalization of the two above distances:

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

Dissimilarity of Numeric Data

- Minkowski distance
  - When h=2, we have Euclidean distance (or L2 distance)
  - When h=1, we have Manhattan distance (L1 distance/ L1 norm)
  - The **Supremum distance** (also L_{max} or L_{∞}, or Chebyshev distance):

$$d(i,j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if} - x_{jf}|.$$

Proximity Measures for Ordinal Attributes

- The values of ordinal attribute have a meaningful order or ranking among them
  - Example: *size* attribute with values (small, medium, large)

- Suppose that *f* is an attribute from a set of ordinal attributes describing *n objects; f* has M_f values with ranks 1, …, M_f
  - Replace each x_{if} by its corresponding rank r_{if}
  - Normalize to the range [0,1] to make all the ordinal attributes have the same weights:

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}.$$

  - Dissimilarity can be computed on normalized values z_{if} using distances for numeric attributes

Proximity Measures for Ordinal Attributes

- Example:

A Sample Data Table Containing Attributes of Mixed Type

Object Identifier	*test-1 (nominal)*	*test-2 (ordinal)*	*test-3 (numeric)*
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

- Dissimilarity Matrix based on **test-2** attribute:

$$\begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}.$$

Dissimilarity for Attributes of Mixed Types

- How to compute the dissimilarity between objects described by mixed attributes?
- Suppose that the data set contains p attributes of mixed type. The dissimilarity between i and j is:

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

- The indicator $\delta_{ij}^{(f)}$
  - 0 if x_{if} or x_{jf} are missing; or $x_{if} = x_{jf} = 0$ and attribute f is asymmetric binary;
  - 1 otherwise.

Dissimilarity for Attributes of Mixed Types

- The contribution of attribute f to the dissimilarity is computed dependent on its type.
  - If f is numeric: $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{max_h x_{hf} - min_h x_{hf}}$, where h runs over all nonmissing objects for attribute f.
  - If f is nominal or binary: $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$; otherwise, $d_{ij}^{(f)} = 1$.
  - If f is ordinal: compute the ranks and normalized, then treat the resulted value as numeric.

Dissimilarity for Attributes of Mixed Types

- Example:

A Sample Data Table Containing Attributes of Mixed Type

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

- Dissimilarity Matrix based on **test-3 attribute**:

$$\begin{bmatrix} 0 & & & \\ 0.55 & 0 & & \\ 0.45 & 1.00 & 0 & \\ 0.40 & 0.14 & 0.86 & 0 \end{bmatrix}.$$

$$d(1,2) = \frac{|45-22|}{|64-22|} \approx 0.55$$

$$d(1,3) = \frac{|64-45|}{|64-22|} \approx 0.45$$

Dissimilarity for Attributes of Mixed Types

- Note that $\delta_{ij}^{(f)} = 1$ for all i,j, f
- Recall:
  - Dissimilarity matrix based on test-1, test-2 and test-3 attribute

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}. \quad \begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}. \quad \begin{bmatrix} 0 & & & \\ 0.55 & 0 & & \\ 0.45 & 1.00 & 0 & \\ 0.40 & 0.14 & 0.86 & 0 \end{bmatrix}.$$

  - Dissimilarity based on 3 attributes:

$$\begin{bmatrix} 0 & & & \\ 0.85 & 0 & & \\ 0.65 & 0.83 & 0 & \\ 0.13 & 0.71 & 0.79 & 0 \end{bmatrix}.$$

Outline

- Data Exploration
  - Data Objects and Attribute Types
  - Basic Statistical Descriptions of Data
  - Data Visualization
  - Measuring Data Similarity and Dissimilarity
  - Useful thing to know: Document representation and similarity measure between documents.

Document Representation: TF.IDF

- TF.IDF (Term Frequency times Inverse Document Frequency)
  - Let f_{ik} to be the frequency (number of occurances) of term (word) k in document i; define the term frequency to be:

$$TF_{ik} = \frac{f_{ik}}{\max_l f_{lk}}$$

  - The IDF for a term is defined as follows:
    - Suppose term k appears in n_k of N documents. Then

$$IDF_k = \log_2(N/n_k)$$

Measuring similarity between documents

- Term vector (using TF or TFIDF) are sparse
- Cosine similarity:

$$sim(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}||\,||\boldsymbol{y}||},$$

- $||\boldsymbol{x}||$ = $\sqrt{x_1^2 + x_2^2 + \cdots + x_p^2}.$

Summary

- Data Exploration
  - Data Objects and Attribute Types
  - Basic Statistical Descriptions of Data
  - Data Visualization
  - Measuring Data Similarity and Dissimilarity
  - Useful thing to know: Document representation and similarity measure between documents.