Data Mining Fundamentals

Topics

- Data and Data Types
- Data Quality
- Data Preprocessing
- Similarity and Dissimilarity
- Data Exploration and Visualization

Topics

- **Data and Data Types**
- Data Quality
- Data Preprocessing
- Similarity and Dissimilarity
- Data Exploration and Visualization

What is Data?

Collection of **objects** defined by **attributes**

An **attribute** is a property or characteristic of an object

- Examples: eye color of a person, temperature, etc.
- Other names: variable, field, characteristic, feature, predictor, etc.

A collection of attributes describe an **object**

- Other names: record, point, case, sample, entity, entry, instance, etc.

Attributes

Objects

Pid	Sex	Age	Pclass	Survived
2	Female	38	1	Yes
3	Female	26	3	Yes
5	Male	35	3	No
7	Male	54	1	No
13	Male	20	3	No
14	Male	39	3	No
21	Male	35	2	No
24	Male	28	1	Yes
34	Male	66	1	No
54	Female	29	2	Yes

Attribute Values

Each attribute has a set of values objects draw from.

The same attribute can be mapped to different attribute values

Example: height can be measured in feet or meters

Different attributes can be mapped to the same set of values

Example: Attribute values for ID and age are integers

Attribute Classification

Discrete Attribute

Has a finite or countably infinite set of values

Examples: zip codes, click counts, set of words in a collection of documents

Often represented as integer variables

Binary attributes are a special case of discrete attributes

Continuous Attribute

Has real numbers as attribute values

Examples: temperature, height, or weight

Continuous attributes are typically represented as floating-point variables

Attribute Classification

Important attribute classes

Categorical

- Nominal
  - **Examples:** ID numbers, eye color, zip codes
- Ordinal
  - **Examples:** Rankings (e.g., place in competition), grades, clothing sizes ({XL, L, M, S, XS})

Interval

Examples: Temperatures in Celsius or Fahrenheit

Ratio

Examples: Temperature in Kelvin, length, time, counts

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

Pid	Sex	Age	Pclass	Survived
2	Female	38	1	Yes
3	Female	26	3	Yes
5	Male	35	3	No
7	Male	54	1	No
13	Male	20	3	No
14	Male	39	3	No
21	Male	35	2	No
24	Male	28	1	Yes
34	Male	66	1	No
54	Female	29	2	Yes

Record: Data Matrix

Data objects with only numeric attributes 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

The data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute.

Record: 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

Record: Transaction Data

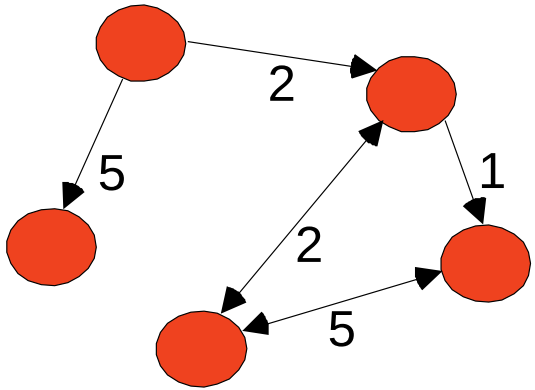
A special type of record data where each record (transaction) involves a set of items

Consider a grocery store. The set of products purchased by a customer during one shopping trip constitutes 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: HTML Data

Examples: Generic graph and HTML Links




```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

Ordered: Medical Data

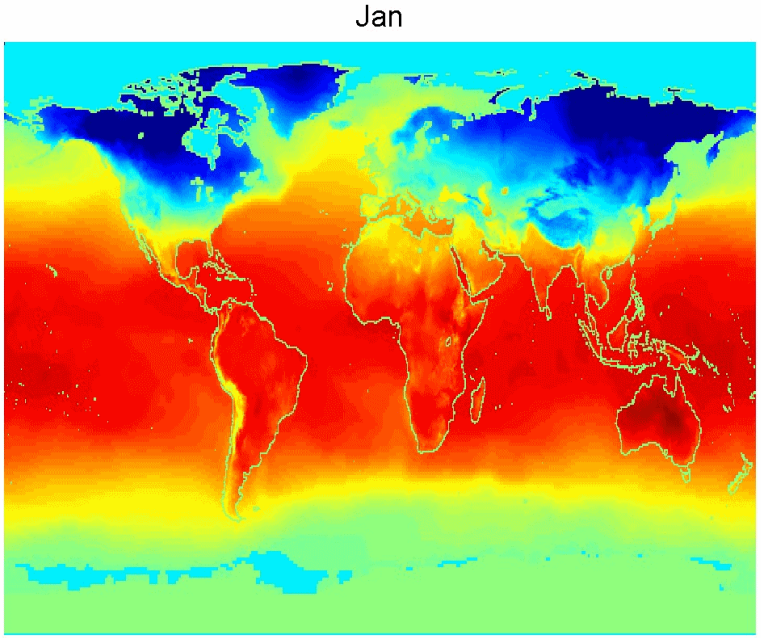
Genomic sequence data

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

Ordered: Climate Data

Spatial-Temporal Data

- Average Monthly Temperature of land and ocean



Topics

- Data and Data Types
- **Data Quality**
- Data Preprocessing
- Similarity and Dissimilarity
- Data Exploration and Visualization

Data Quality

What problems should we worry about?
How can we detect problems with the data?
What can we do about these problems?

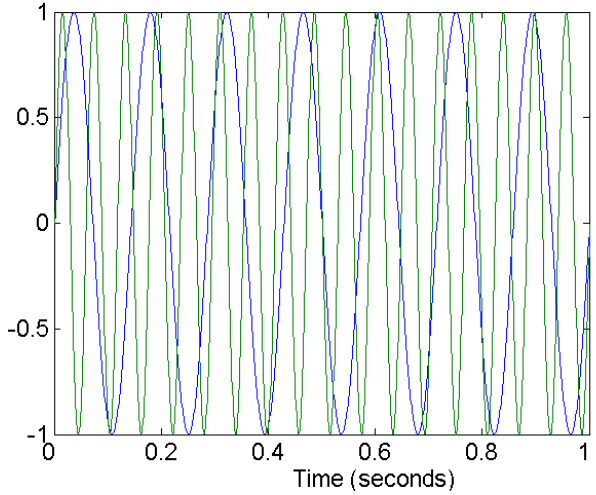
Examples of data quality problems:

- Noise and outliers
- Missing values
- Duplicate data

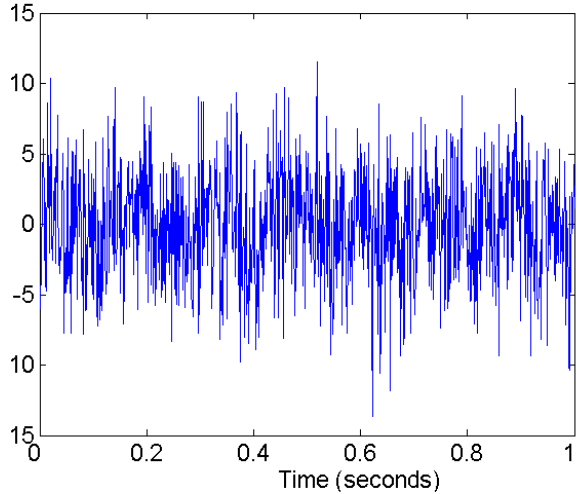
Noise

Noise: An invalid signal overlapping valid data

Examples: distortion of a person's voice over the phone; "snow" on a television screen; human inconsistency in labeling



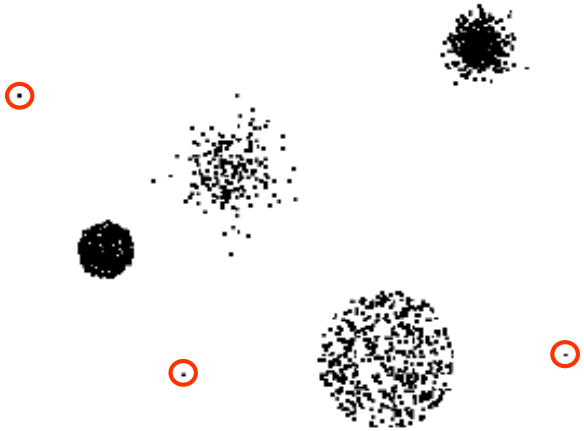
Two Sine Waves



Two Sine Waves + Noise

Outliers

Outliers: data objects with characteristics that are considerably different than most of the other data objects in the data set



Missing Values

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)

Handling missing values

- Eliminate Data Objects
- Estimate Missing Values
- Ignore the Missing Value During Analysis
- Replace with all possible values (weighted by their probabilities)

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

Example:

Same person with multiple email addresses

Topics

- Data and Data Types
- Data Quality
- **Data Preprocessing**
- Similarity and Dissimilarity
- Data Exploration and Visualization

Data Preprocessing

Aggregation

Sampling

Dimensionality Reduction

Feature Subset Selection

Feature Creation

Discretization and Binarization

Attribute Transformation

Aggregation

Combining two or more attributes (or objects) into a single attribute (or object)

Results

Data reduction

Reduce the number of attributes or objects

Change of scale

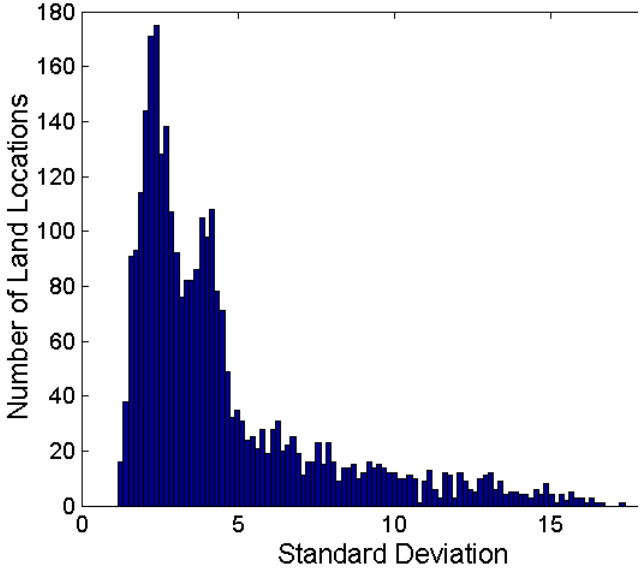
Cities aggregated into regions, states, countries, etc.

More “stable” data

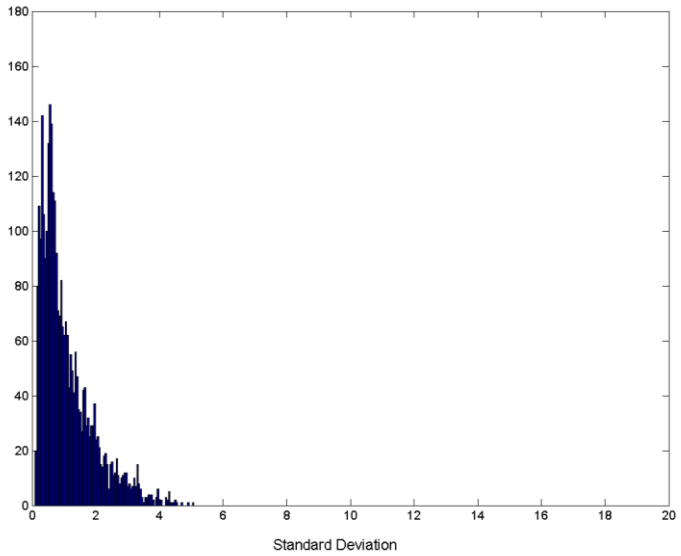
Aggregated data tends to have less variability

Aggregation

Variation of Precipitation in Australia



Standard Deviation of Average Monthly Precipitation



Standard Deviation of Average Yearly Precipitation

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

Data miners sample because processing the entire set of data of interest is too expensive or time consuming

Sampling

The key principle for effective sampling is:

A sample will work almost as well as using the entire data set **if the sample is representative.**

Types of Sampling

Simple Random Sampling

There is an equal probability of selecting any particular item

Stratified sampling

Split the data into several partitions; draw random samples from each partition

Types of Sampling

Sampling without replacement

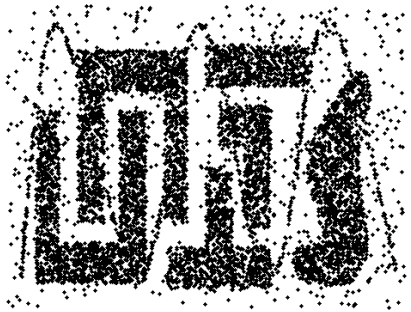
As each item is selected, it is removed from the population

Sampling with replacement

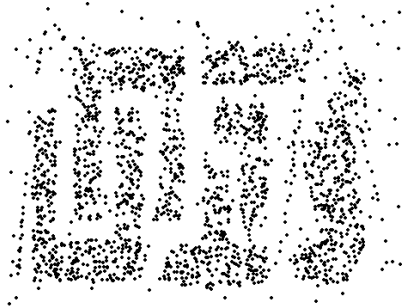
Objects are not removed from the population as they are selected for the sample

- In sampling with replacement, the same object can be picked up more than once

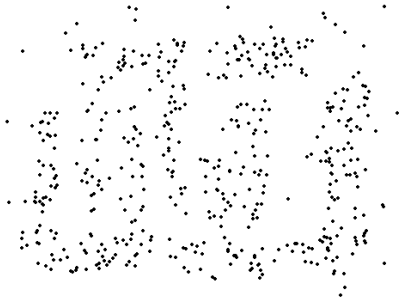
Sample Size



8000 points



2000 Points



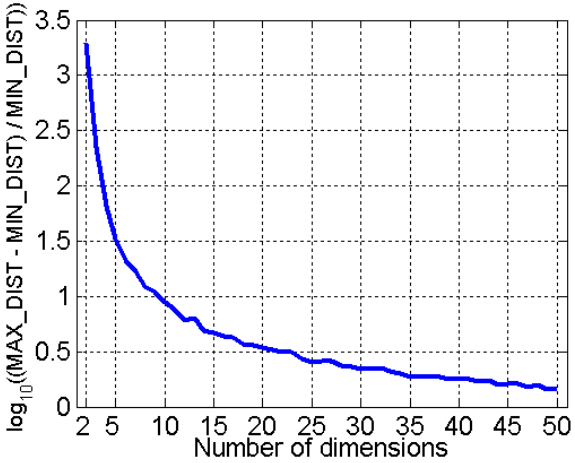
500 Points

Curse of Dimensionality

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

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

- Randomly generate 500 points
- Compute difference between max and min distance between any pair of points



Dimensionality Reduction

Purpose:

Avoid curse of dimensionality
Reduce time and memory required
Allow data to be more easily visualized
May help to eliminate irrelevant features or reduce noise

Techniques:

Principle Component Analysis
Singular Value Decomposition
Others: various supervised and non-linear techniques

Feature Subset Selection

Another way to reduce dimensionality of data

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

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

Feature Subset Selection

Techniques:

Brute-force approach:

Try all possible feature subsets as input to data mining algorithm

Embedded approach:

Feature selection occurs naturally as part of the data mining algorithm

Filter approach:

Features are selected before data mining algorithm is run

Wrapper approach:

Use a data mining algorithm as a black box to find best subset of attributes

Feature Creation

Original attributes not always best representation of information

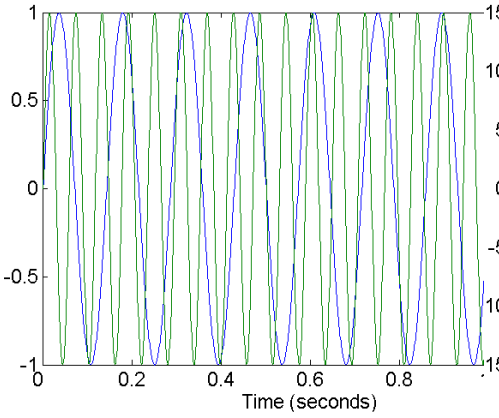
Create new features which are more efficient/focused

Three general methodologies:

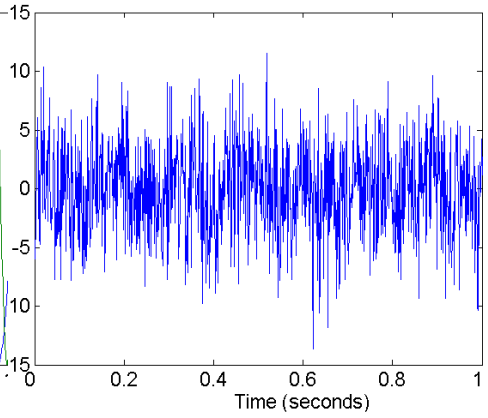
- Feature Extraction-domain specific
- Feature Construction-combining features
- Mapping Data to New Space

Mapping Data to a New Space

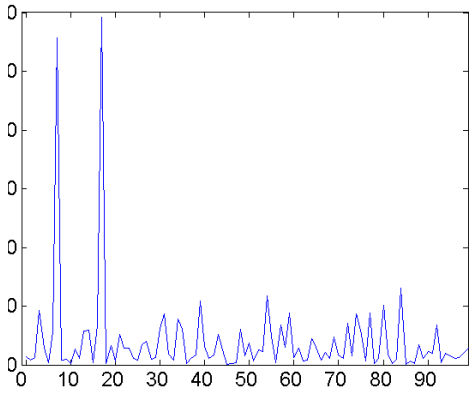
- Fourier transform
- Wavelet transform



Two Sine Waves



Two Sine Waves + Noise



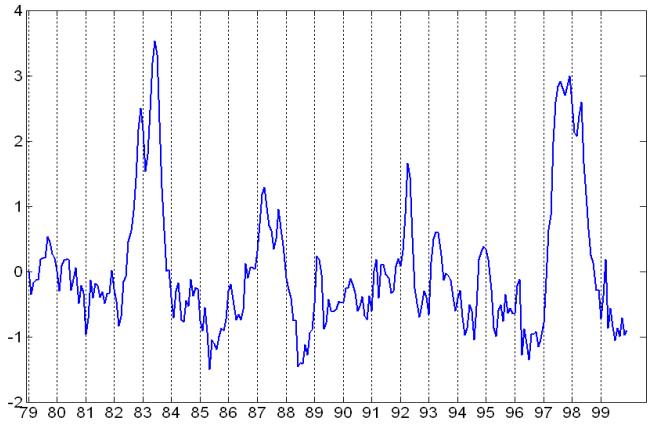
Frequency

Attribute Transformation

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|

Standardization and normalization



Topics

- Data and Data Types
- Data Quality
- Data Preprocessing
- **Similarity and Dissimilarity**
- Data Exploration and Visualization

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 either/both

Similarity/Dissimilarity for Single Attributes

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 $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Euclidean Distance

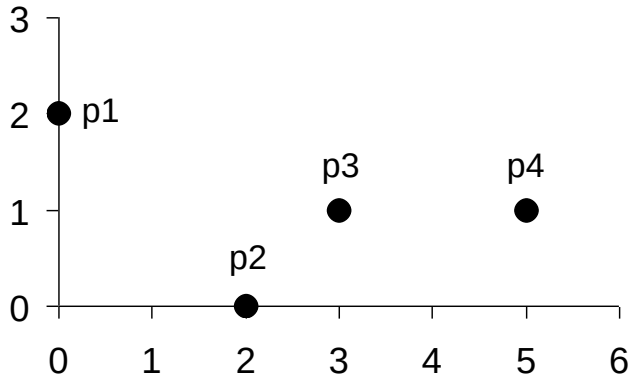
- Euclidean Distance:

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

n is the number of dimensions (attributes)

p_k and q_k are, respectively, the k^{th} attributes (components) of data objects p and q.

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

Correlation

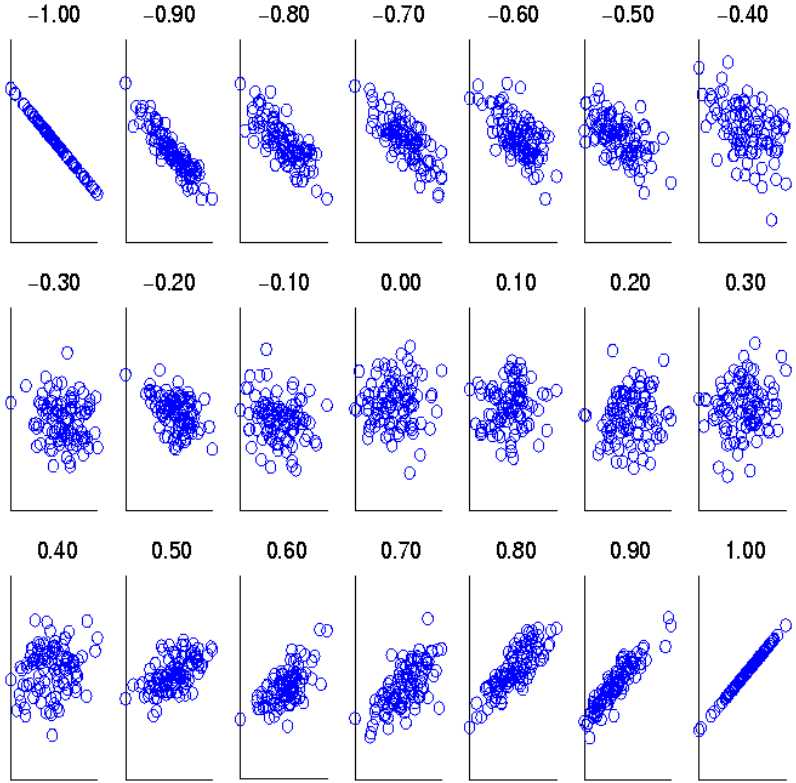
- Correlation measures the linear relationship between objects
- Standardize data objects (p and q) and then take their dot product

$$p'_k = (p_k - mean(p)) / std(p)$$

$$q'_k = (q_k - mean(q)) / std(q)$$

$$correlation(p, q) = p' \bullet q'$$

Visually Evaluating Correlation



Scatter plots showing the similarity from -1 to 1

Topics

- Data and Data Types
- Data Quality
- Data Preprocessing
- Similarity and Dissimilarity
- **Data Exploration and Visualization**

What is data exploration?

- Visualization and calculation to better understand characteristics of data
- Key motivations of data exploration:
  - Helping to select the right tool for preprocessing or analysis
  - Making use of humans' abilities to recognize patterns
    - People can recognize patterns not captured by data analysis tools
- Related to the field of Exploratory Data Analysis (EDA)
  - Created by statistician John Tukey
  - Seminal book is Exploratory Data Analysis by Tukey
  - Nice online introduction in Chapter 1 of the NIST Engineering Statistics Handbook (http://www.itl.nist.gov/div898/handbook/index.htm)

Techniques Used In Data Exploration

Original EDA definition

Focus on visualization

Clustering and anomaly detection were viewed as exploratory techniques

Now

Clustering and anomaly detection are major areas of interest, not just exploratory

Our focus

Summary statistics

Visualization

Summary Statistics

Numbers that summarize properties of the data

- Examples:
  - Frequency - counts
  - Center – mean
  - Spread – standard deviation
- Most can be calculated in a single pass through the data

Frequency and Mode

- The **frequency** of an attribute value is a percentage measuring how often the value occurs in the data set
  - Example: 'gender'
    - In a representative population of people, 'female' occurs about 50% of the time
- The **mode** of an attribute is the most frequent attribute value
- Typically used with categorical data

Percentiles

For continuous data, the notion of a percentile is more useful

Given an ordinal or continuous attribute *x* and a number *p* between 0 and 100, the *p*th percentile is the value x_p such that *p*% of the observed values of x are less than x_p

The 50th percentile is the value $x_{50\%}$ such that 50% of all values of x are less than $x_{50\%}$

Percentiles

Example: You are the fourth tallest person in a group of 20

80% of people are shorter than you:

That means you are at the **80th percentile**.

If your height is 1.85m then "1.85m" is the 80th percentile height in that group.

https://www.mathsisfun.com/data/percentiles.html

Measures of Center: Mean and Median

- The mean is the most common measure of the center of a set of points
  - Very sensitive to outliers
- The median or a trimmed mean is also commonly used

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

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

Measures of Spread: Range and Variance

- **Range** is the difference between the max and min
- **Variance** and **standard deviation** are the most common measure of the spread of a set of points

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

- These are sensitive to outliers, other measures used include

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

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

$$\text{interquartile range}(x) = x_{75\%} - x_{25\%}$$

Visualization

Represent data in a visual or tabular format

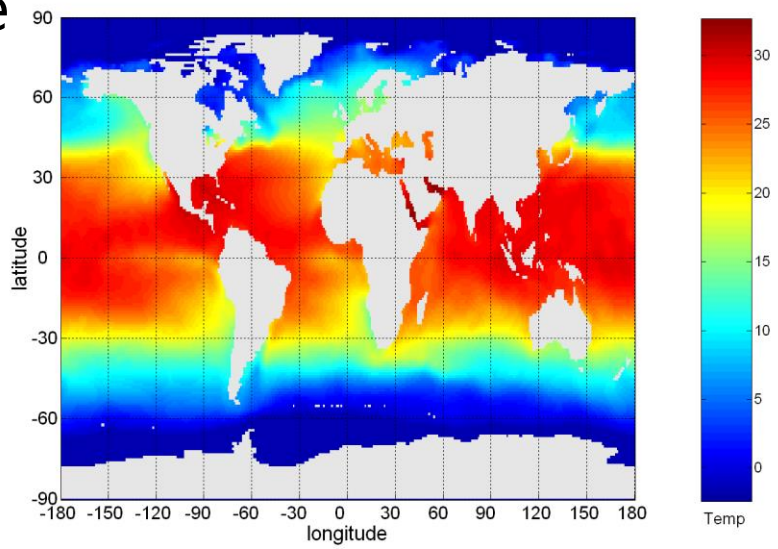
- Characteristics of the data and relationships among data items or attributes can be analyzed and/or reported.

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
- Detect general patterns and trends
- Detect outliers and unusual patterns

Example: Sea Surface Temperature

- The following shows the Sea Surface Temperature (SST) for July 1982
  - Tens of thousands of data points are summarized in a single figure

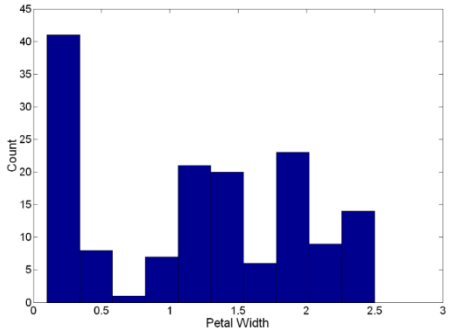


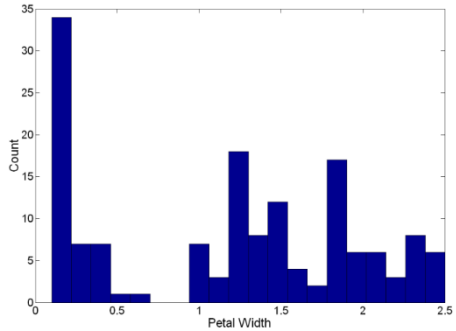
Representation

- The mapping of information to a visual format
- Data objects, their attributes, and the relationships among data objects are translated into graphical elements such as points, lines, shapes, and colors
- Example:
  - Objects are often represented as points
  - Their attribute values can be represented as the position of the points or the characteristics of the points, e.g., color, size, and shape
  - If position is used, then the relationships of points, i.e., whether they form groups or a point is an outlier, is easily perceived.

Visualization Techniques: Histograms

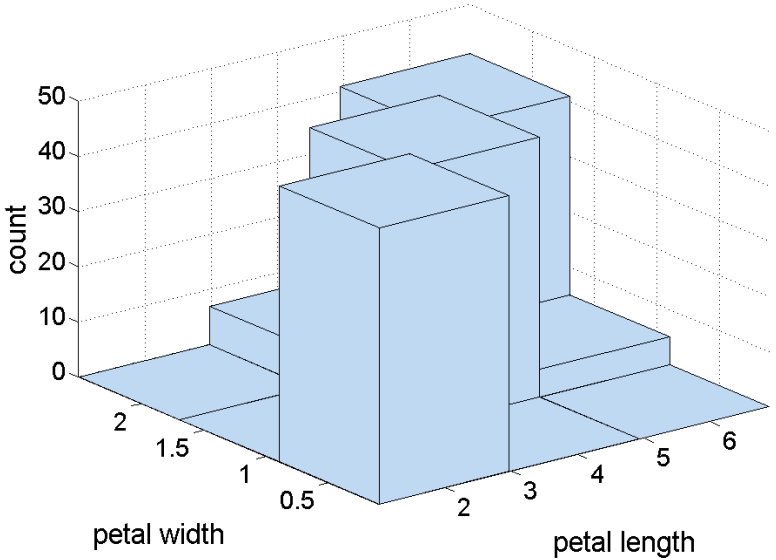
- Histogram
  - Shows the distribution of values of a single variable
  - Divide the values into bins and show a bar plot of the number of objects in each bin
  - The height of each bar indicates the number of objects
  - Shape of histogram depends on the number of bins - experiment
- Example: Petal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





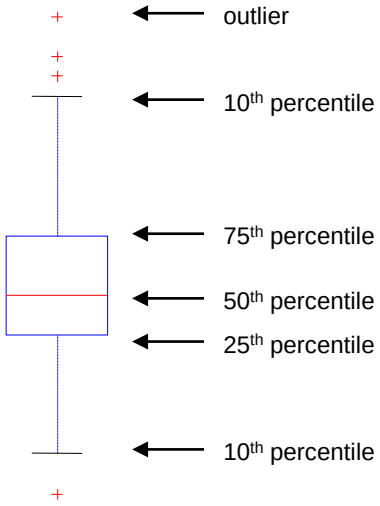
Two-Dimensional Histograms

- Show the joint distribution of the values of two attributes
- Example: petal width and petal length



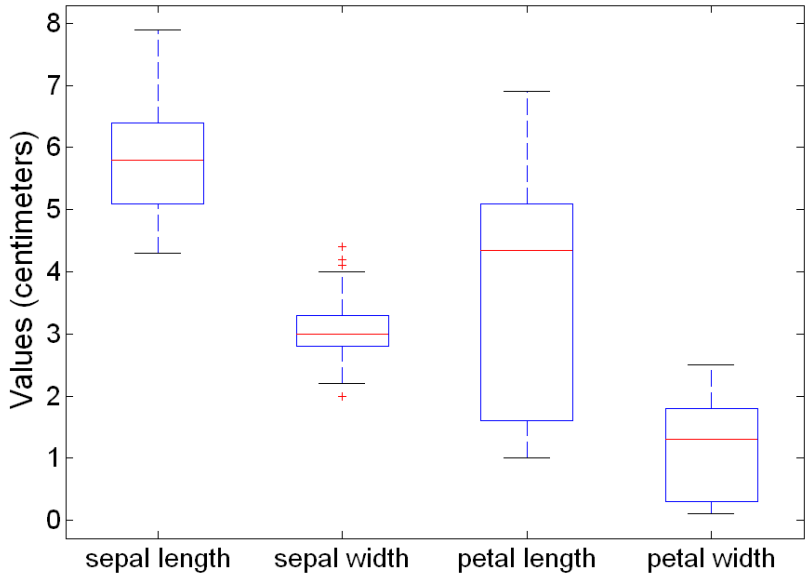
Visualization Techniques: Box Plots

- Box Plots
  - Invented by J. Tukey
  - Displays distribution of data
  - Format:



Example of Box Plots

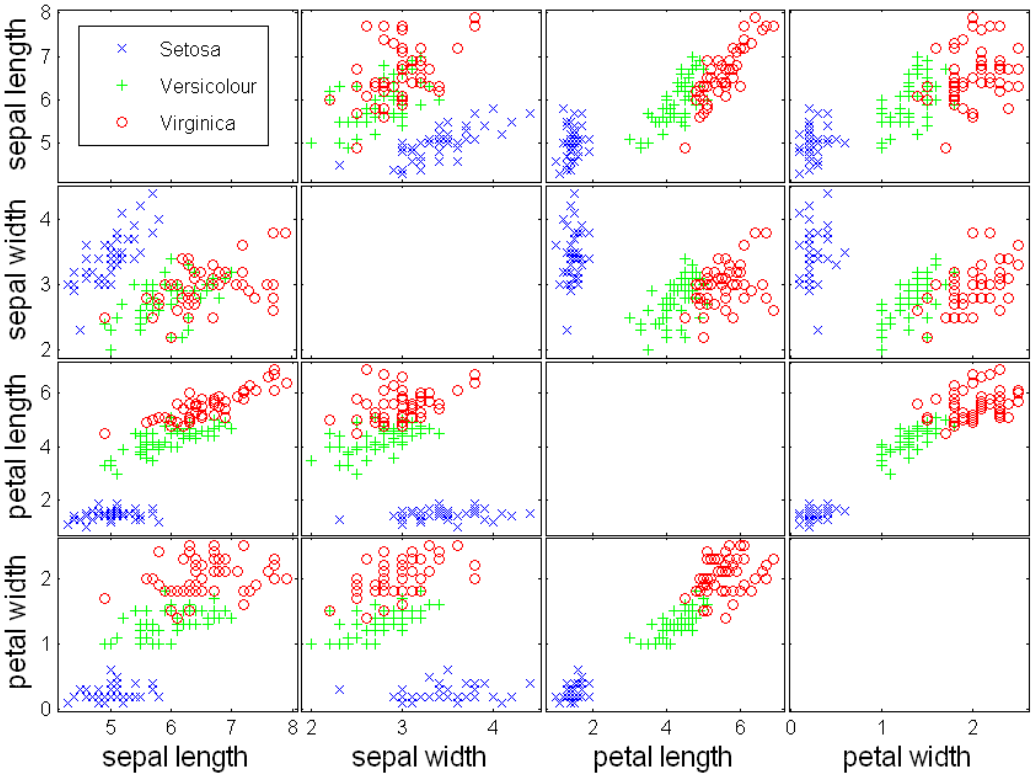
- Box plots can be used to compare attributes



Visualization Techniques: Scatter Plots

- Scatter plots
  - Attribute values determine the position
  - Two-dimensional scatter plots most common
    - Three-dimensional scatter plots also used
  - Use the size, shape, and color of markers to display supplementary attributes
  - Arrays of scatter plots compactly summarize factor relationships

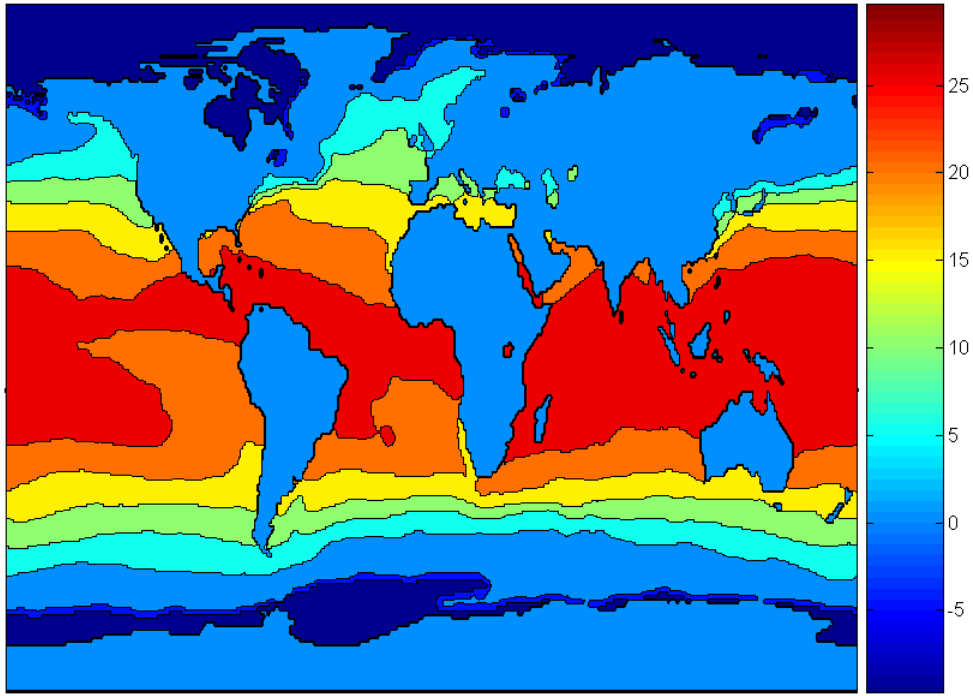
Scatter Plot Array of Iris Attributes



Visualization Techniques: Contour Plots

- Contour plots
  - Used for continuous attributes on a spatial grid
  - Partition the plane into regions of similar values
  - “Contour lines” that form the boundaries of these regions connect points with equal values
  - Examples:
    - Elevation
    - Climate Data

Contour Plot Example: SST Dec, 1998



Questions?