

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, temperature, daily revenue
 - Variable, field, characteristic, feature, predictor, etc.
- A collection of attributes describe an **object**
 - Record, point, case, sample, entity, entry, instance, etc.

Objects

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

Attribute Values

Each attribute has a set of values which objects draw from.

- The same attribute can be mapped to different attribute value sets
 - Example: Movie ratings can be represented as strings ("one", "two") or integers (1, 2)
- Different attributes can be mapped to the same set of values
 - Example: Age and ID number are both usually represented as integers

Attribute Classification

- **Discrete**

- Has a finite or countably infinite set of values
- **Examples:** zip codes, click counts, colors, gender
- Often represented as integer variables
- Binary attributes are a special case of discrete attributes

- **Continuous**

- Has real numbers as attribute values
- **Examples:** temperature, height, or weight
- Often represented as floating-point variables

Attribute Classification

Categorical

- Always discrete
- Nominal
 - Has no natural order
 - Ex: eye color, zip codes, gender
- Ordinal
 - Has a natural order
 - Ex: sibling number, clothing size

Numeric

- Interval
 - Degree of difference is meaningful
 - Ex: Temperature in °C or °F
- Ratio
 - Interval, with a unique and non-arbitrary 0
 - Ex: Length, weight, duration

Types of Data Sets

■ Record

- Data Matrix
- Document Data
- Transaction Data

■ Graph

- World Wide Web
- Molecular Structures

■ Ordered

- Spatial Data
- Temporal Data
- Sequential Data

Record Data

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

<i>Pid</i>	<i>Sex</i>	<i>Age</i>	<i>Pclass</i>	<i>Survived</i>
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 $m \times n$ 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"

	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

Each component (attribute) of the vector represents a term

The value of each component is the number of times the corresponding term occurs in the document

Record: Transaction Data

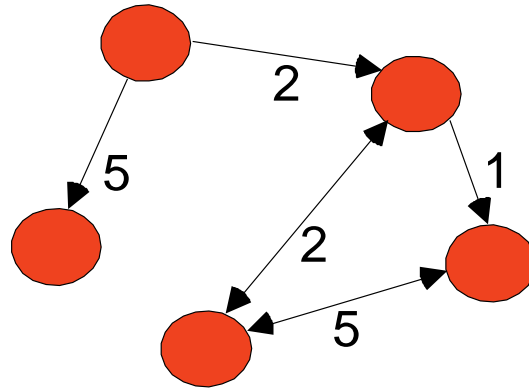
Each record is a "transaction" and has an associated set of "items"

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.

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Graph Data

Data which consists of a list of "vertices" and "edges" which connect two vertices.



Graph: HTML Data

```
<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
```

Vertices: Websites

Edges: Directed from a page with a link to the linked page

Ordered Data

Data with an ordering among objects which needs to be preserved

- Whether data is ordered or not will depend on the question being asked

Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure.

Ordered: Medical Data

Genomic sequence data

- Ordered if we are trying to predict the next triplet in a sequence.

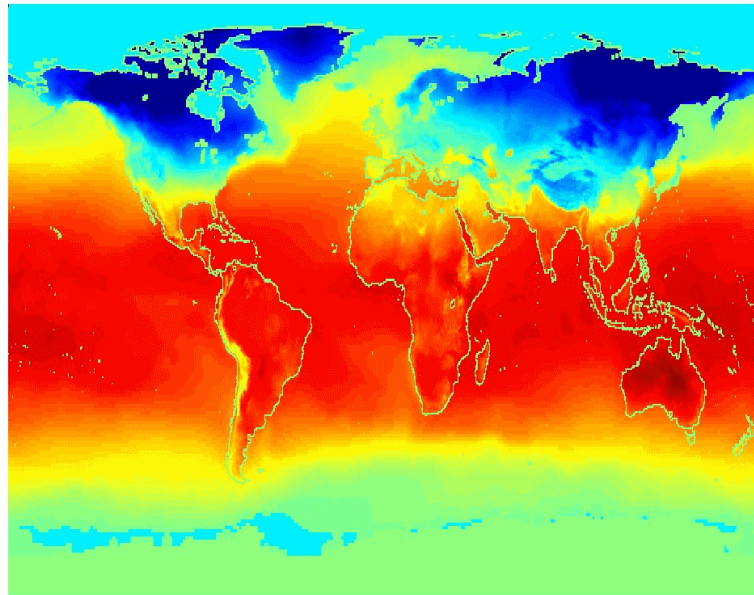
```
GGTTCCGCCTTCAGCCCCGCGCC  
CGCAGGGCCCGCCCCGCGCCGTC  
GAGAAGGGCCCGCCTGGCGGGCG  
GGGGGAGGCGGGGCGCCCGAGC  
CCAACCGAGTCCGACCAGGTGCC  
CCCTCTGCTCGGCCTAGACCTGA  
GCTCATTAGGCGGCAGCGGACAG  
GCCAAGTAGAACACGCGAAGCGC  
TGGGCTGCCTGCTGCGACCAGGG
```


Ordered: Climate Data

Spatial-Temporal Data

- Ordered if distance in space or time is important to our question

Jan



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

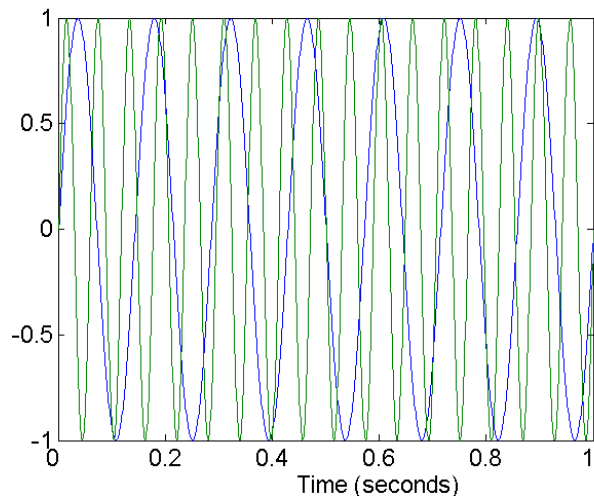
■ Questions to Ask

- What problems should we worry about?
 - Noise
 - Outliers
 - Missing values
 - Duplicates
 - Domain specific problems
- How can we detect these problems?
- What can we do about these problems?

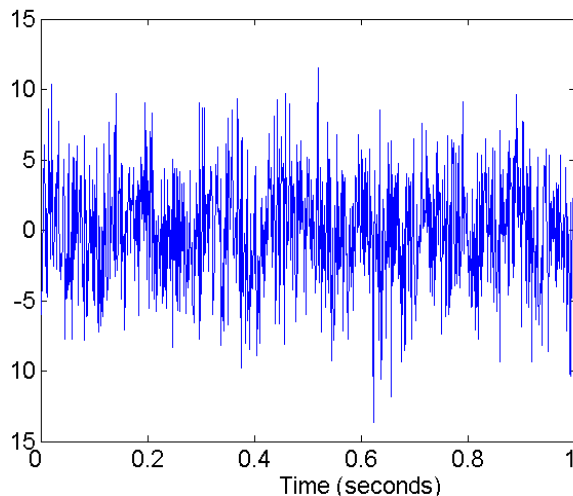
Noise

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



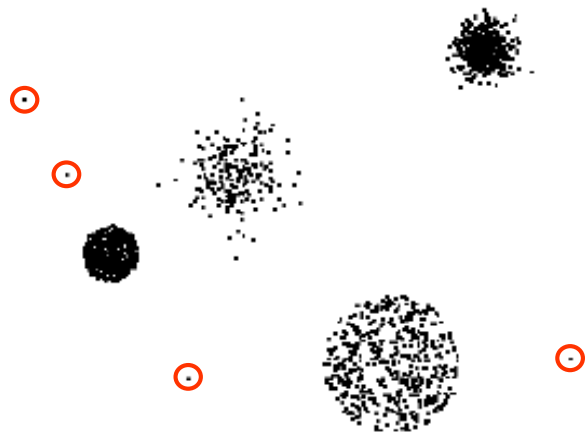
Two Sine Waves



Two Sine Waves + Noise

Outliers

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 or lost
 - People decline to give their age and weight
- Attributes may not be applicable to all cases
 - Annual income is not applicable to children

Handling missing values

- Remove object from dataset
- Ignore the missing value (not always possible)
- Replace with static value (mean, median, mode, etc)
- Replace with random value (weighted by frequency of value)

Duplicate Data

Data objects that represent an identical instance

- Example: Same person with multiple email addresses
- Major issue when merging data from multiple sources
- Carefully filter your data and remove/merge duplicates

Topics

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

Data Preprocessing Techniques

- Sampling
- Object Transformation
- Attribute Transformation
- Attribute Reduction

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
- Widely used in traditional statistical studies

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: Key Principle

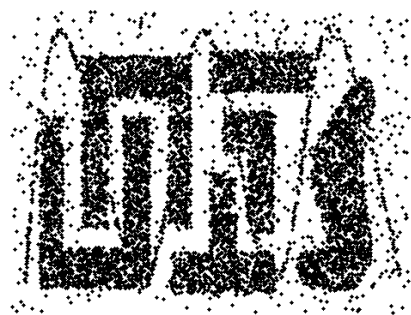
A sample will work almost as well as using the entire data set **if the sample is representative.**

Types of Sampling

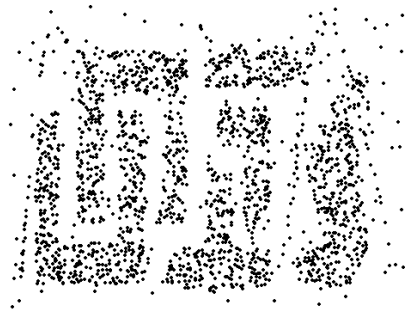
- Simple
 - There is an equal probability of selecting any particular item
- Stratified
 - Split the data into several partitions
 - Select fixed number of random samples from each partition
- Without replacement
 - As each item is selected, it is removed from the population
- With replacement
 - Objects are not removed from the population as they are selected for the sample
 - The same object can be selected more than once

Sample Size

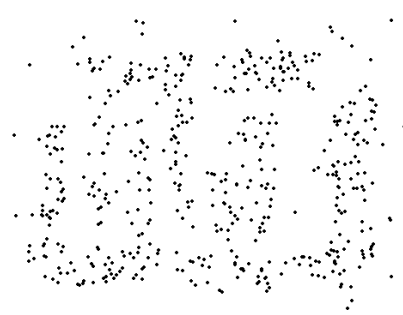
How large a sample should we use?



8000 points



2000 Points



500 Points

Data Preprocessing Techniques

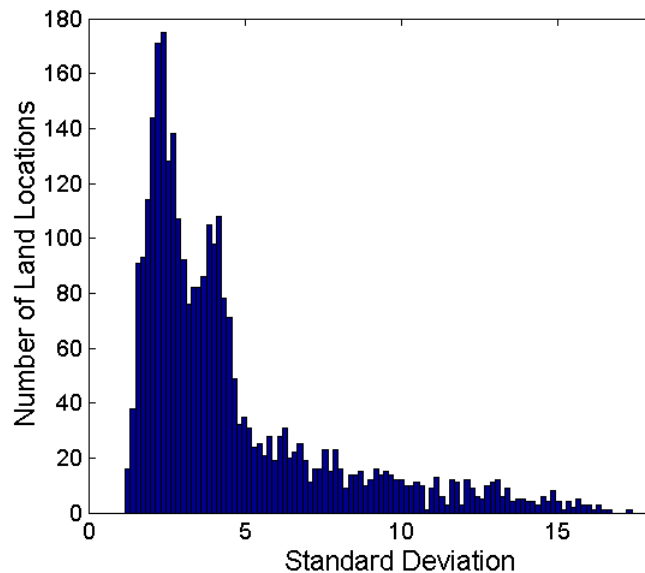
- Sampling
- **Object Transformation**
- Attribute Transformation
- Attribute Reduction

Object Aggregation

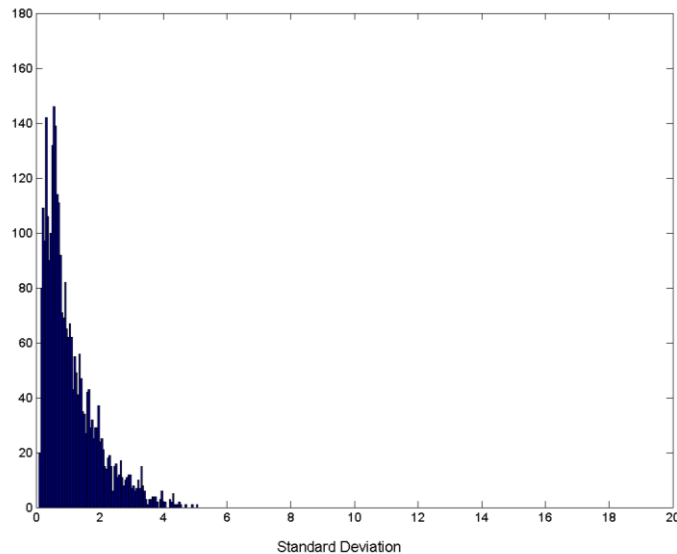
- Combine two or more objects into a single object
 - Examples: Average, sum, difference, product
- Why do this?
 - Change of scale (cities -> states -> nations)
 - Stability of data (reduces variance)

Aggregation

Variation of Precipitation in Australia



Standard Deviation of Average
Monthly Precipitation

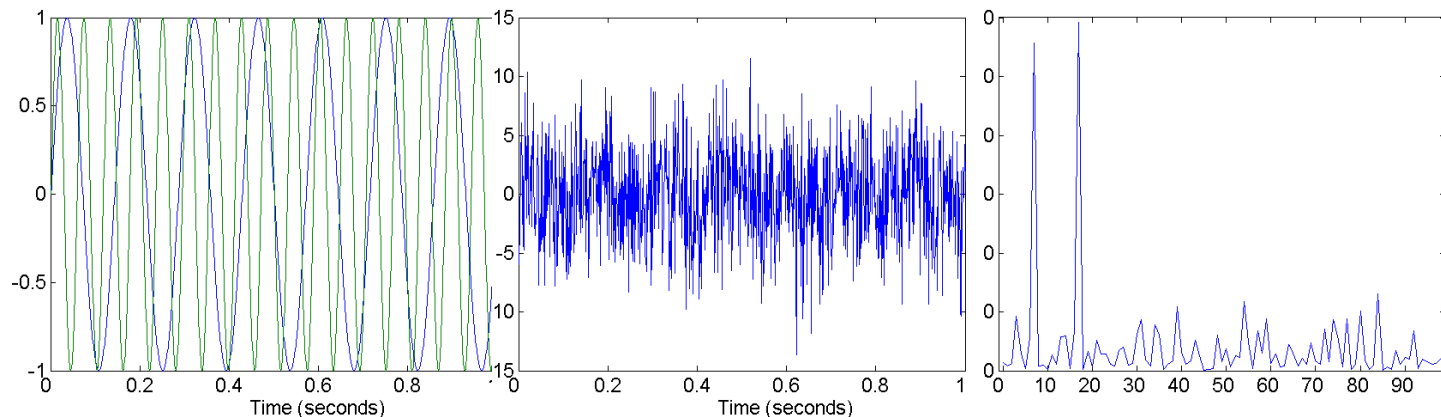


Standard Deviation of Average Yearly
Precipitation

Mapping Objects to a New Space

Find patterns by transforming representation

- Example: Fourier transform, Wavelet transform
- Common in signal processing applications



Two Sine Waves

Two Sine Waves + Noise

Frequency

Data Preprocessing Techniques

- Sampling
- Object Transformation
- **Attribute Transformation**
- Attribute Reduction

Attribute Transformations

- Standardization
 - Force numeric column to have mean 0 and standard deviation 1
 - Subtract mean and divide by standard deviation
- Normalization
 - Set minimum value to 0 and maximum value to 1
 - Subtract minimum and divide by new maximum

Attribute Transformations

- Functional Transformation

- Map the set of values of a given attribute to a new set of values such that each old value can be identified with one of the new values
- Example functions: x^k , $\log(x)$, e^x , $|x|$

- Feature Extraction

- Create new column from old
- Domain Specific
- Example: Extract profit earned from total transaction price

Data Preprocessing Techniques

- Sampling
- Object Transformation
- Attribute Transformation
- **Attribute Reduction**

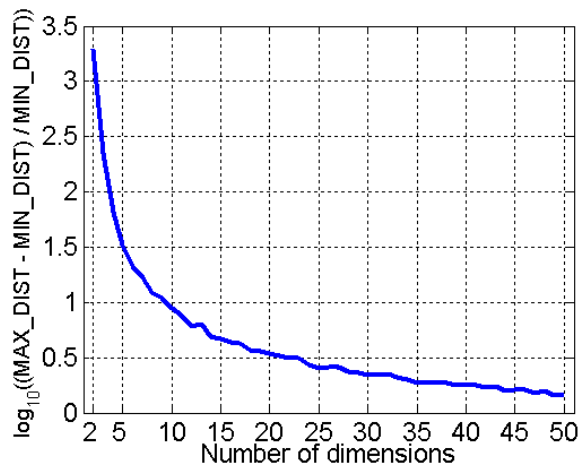
Curse of Dimensionality

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

Definitions of density and distance between points become less meaningful

Exercise

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



Attribute Aggregation

- Combine two or more attributes into one attribute
 - Examples: Average, sum, difference, product
- Why do this?
 - Reduce redundancy
 - Reduce variance

Feature Subset Selection

- Remove features from dataset
 - 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: student ID is often irrelevant when predicting GPA

Feature Subset Selection

■ Techniques

- Brute-force
 - Try all possible feature subsets as input to data mining algorithm
- Embed
 - Feature selection occurs naturally as part of the data mining algorithm
- Manual Filter
 - Features are selected before data mining algorithm is run
- Wrapper
 - Use a data mining algorithm as a black box to find best subset of attributes

Dimensionality Reduction

- Most columns have small correlations with target
- Combine columns to create fewer, but more correlated, columns
- Useful for visualization of high-dimensionality datasets
- Techniques:
 - Principle Component Analysis
 - Singular Value Decomposition
 - Others: various supervised and non-linear techniques

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
- **Larger** when objects are more alike
- Often falls in the range $[0,1]$

■ Dissimilarity

- Numerical measure of how **different** two data objects are
- **Smaller** when objects are more alike
- Minimum dissimilarity is often 0; upper limit varies

■ Proximity

- Refers to both similarity/dissimilarity

Similarity/Dissimilarity for Single Attributes

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	$d = \frac{ p-q }{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{ p-q }{n-1}$
Interval or Ratio	$d = p - q $	$s = -d, s = \frac{1}{1+d} \text{ or } s = 1 - \frac{d - \min_d}{\max_d - \min_d}$

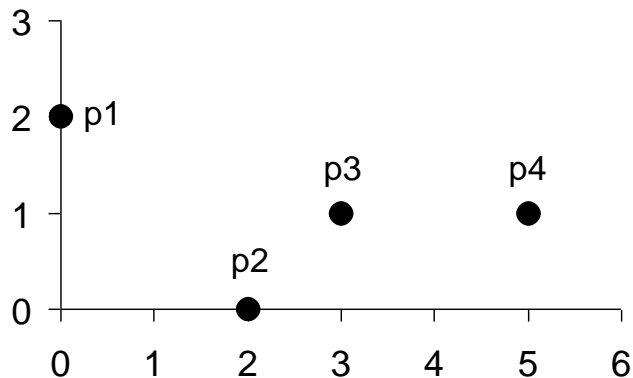
p and q are the attribute values for two data objects

Object Dissimilarity

- Euclidean Distance

- $d(p, q) = \sqrt{\sum_{k=1}^n (p_k - q_k)^2}$
 - n is the number of dimensions (attributes)
 - p_k and q_k are the k^{th} attribute values (components) of data objects p and q , respectively.
- Most common distance metric

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

Object Similarity

- Cosine Similarity

$$\cos(p, q) = \frac{p \cdot q}{|p||q|} = \frac{\sum_{k=1}^n p_k * q_k}{\sqrt{\sum_{k=1}^n p_k^2} \sqrt{\sum_{k=1}^n q_k^2}}$$

Example:

$$p = [3, 2, 0, 5, 0, 0, 0, 2, 0, 0]$$

$$q = [1, 0, 0, 0, 0, 0, 0, 1, 0, 2]$$

$$p \cdot q = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

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

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

$$\cos(p, q) = .3150$$

General Similarity

■ Correlation

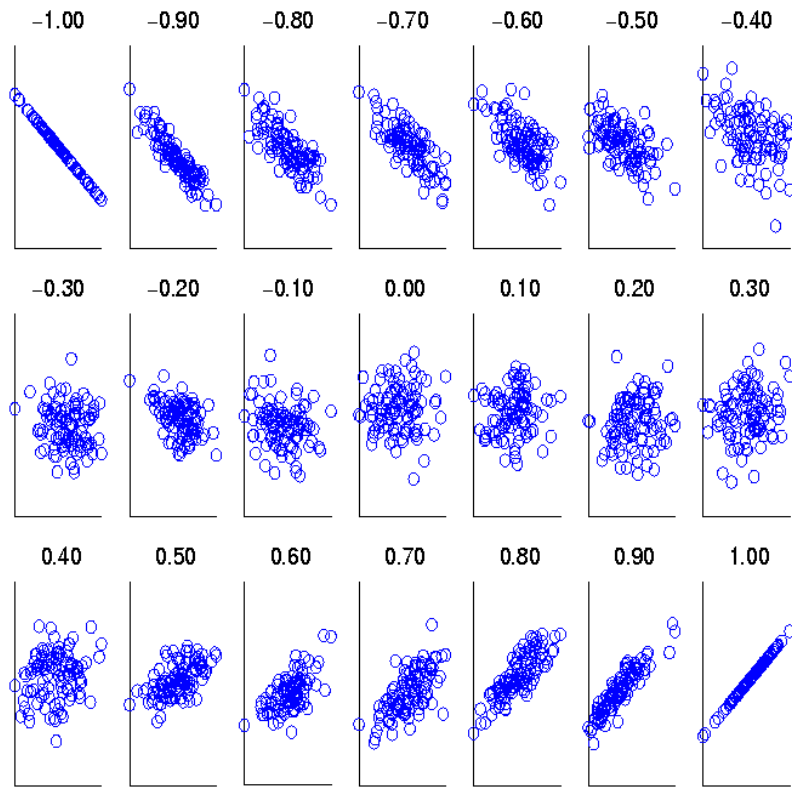
- Measures the linearity of the relationship between attributes or objects
- Standardize two objects or attributes (p and q) and then take their dot product

$$p'_k = (p_k - \text{mean}(p)) / \text{std}(p)$$

$$q'_k = (q_k - \text{mean}(q)) / \text{std}(q)$$

$$\text{correlation}(p, q) = p' \bullet q'$$

Visually Evaluating Correlation



Scatter plots between two attributes, with correlations 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
 - 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
- Exploratory Data Analysis (EDA)
 - Subfield of statistics 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>)

Summary Statistics

- Numbers that summarize properties of the data
 - Frequency - counts
 - Center – mean
 - Spread – standard deviation
- Most can be calculated in a single pass through the data

Frequency and Mode

- Frequency
 - Percentage measuring how often a given attribute value occurs in the data set
 - Example: 'gender' in US population
 - 'female' has a frequency of about 50%
- Mode
 - Most frequent attribute value in data set
- Typically used with categorical data

Percentiles

- Used for continuous attributes
 - The p th percentile is the value x_p such that $p\%$ of the observed values of x are less than x_p
 - Example: 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.

Measures of Center

- Mean

- $mean(X) = \bar{X} = \frac{1}{m} \sum_{i=1}^m X_i$
- The most common measure of the center of a set of points
- Very sensitive to outliers

- Median

- $median(X) = \begin{cases} X_{(\frac{m+1}{2})} & \text{if } m \text{ is odd} \\ \frac{1}{2} (X_{(\frac{m}{2})} + X_{(\frac{m}{2}+1)}) & \text{if } m \text{ is even} \end{cases}$
- Not sensitive to outliers

Measures of Spread

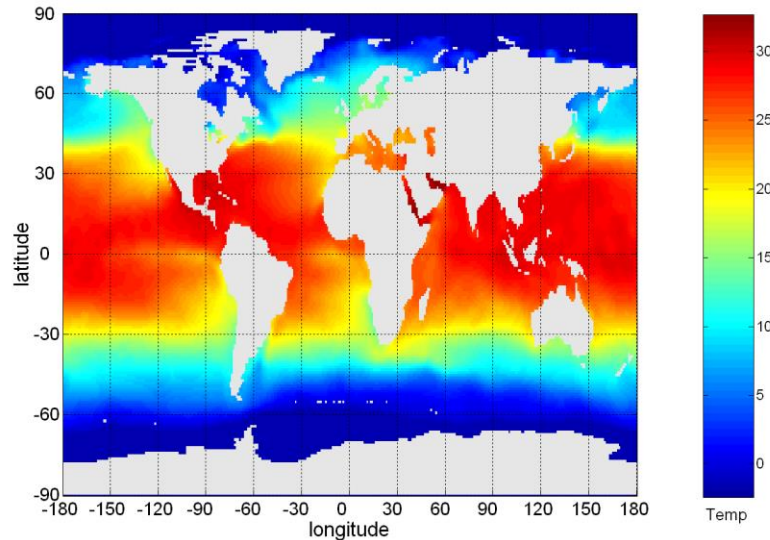
- Range
 - $range(x) = \max(x) - \min(x)$
- Variance and standard deviation
 - The most common measures of the spread of a set of points
 - $variance(x) = stdev(x)^2 = \frac{1}{m} \sum_{i=1}^m (x_i - \bar{x})^2$
 - Sensitive to outliers
- Others
 - Average absolute deviation: $AAD(x) = \frac{1}{m} \sum_{i=1}^m |x_i - \bar{x}|$
 - Median absolute deviation: $MAD(x) = median(\{|x_i - \bar{x}|\})$
 - Interquartile range: $IQR(x) = x_{75} - x_{25}$

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

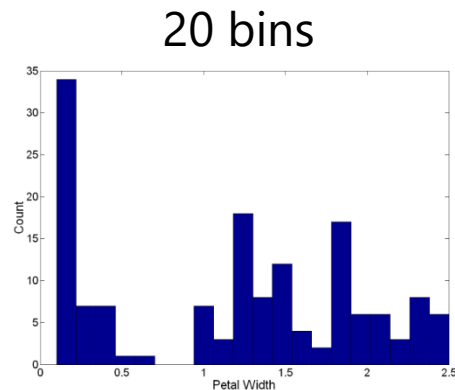
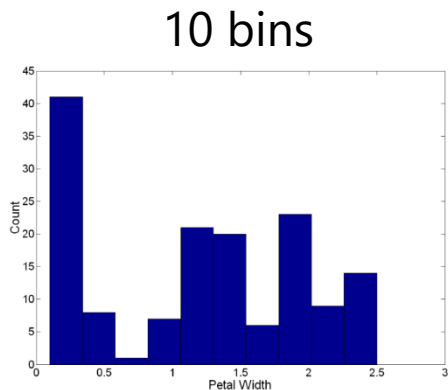


Representation

- The mapping of information to a visual format
- Translate data into graphical elements such as points, lines, shapes, and colors
- Examples
 - 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 relationships between points can often be perceived.

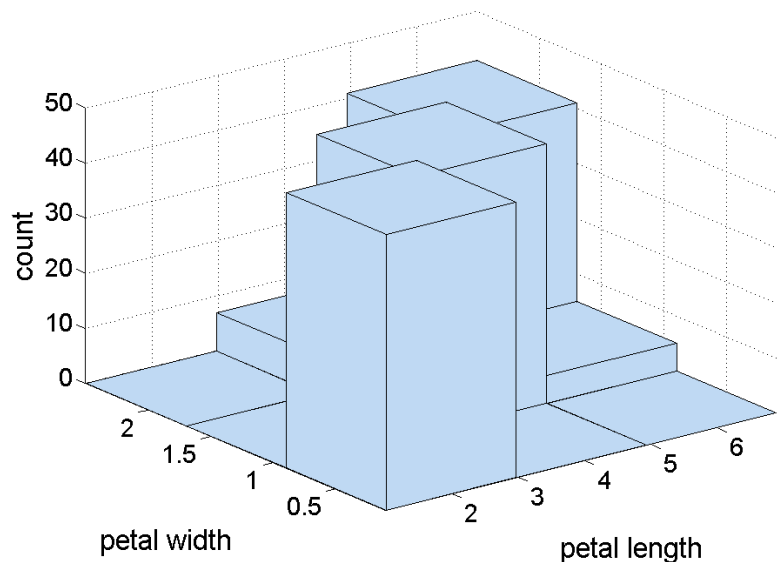
Histograms

- 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



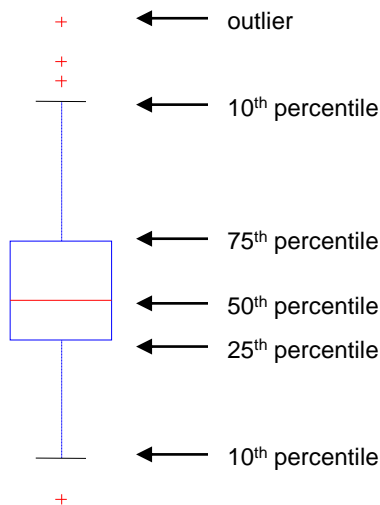
Two-Dimensional Histograms

- Show the joint distribution of the values of two attributes



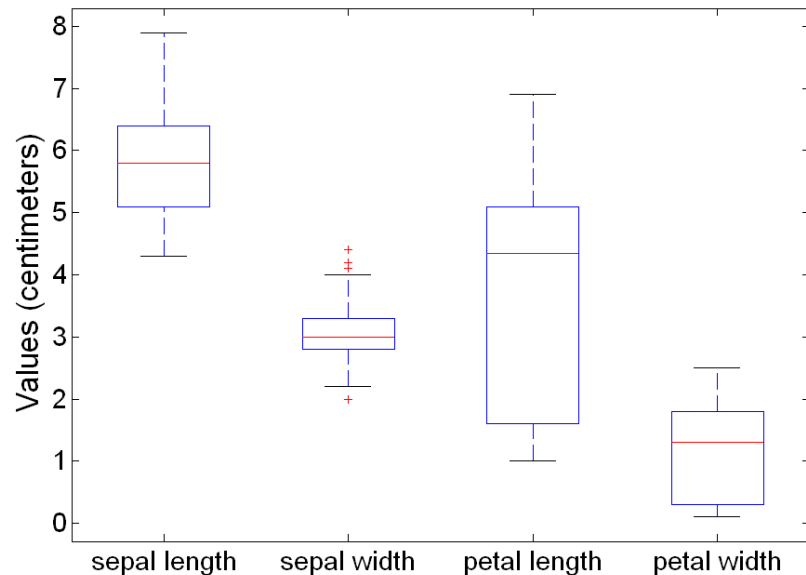
Box Plots

- Invented by J. Tukey
- Displays distribution of data



Example of Box Plots

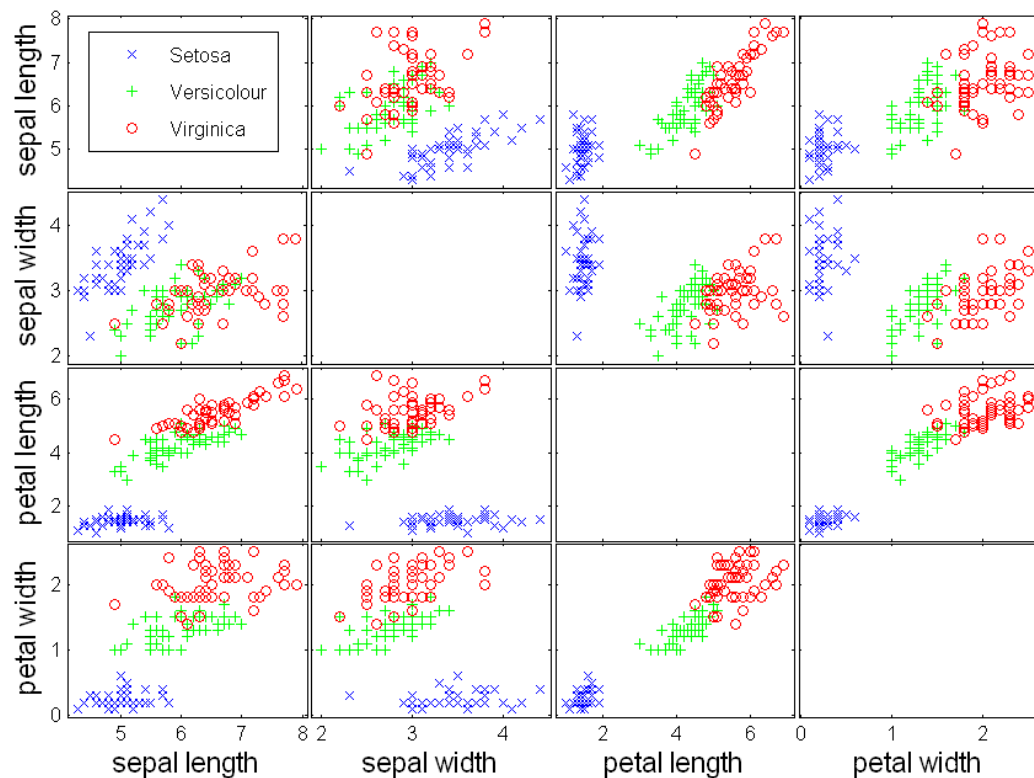
- Box plots can be used to compare attributes



Scatter Plots

- Attribute values determine the position of each point
- Two-dimensional scatter plots most common
- Use the size, shape, and color of markers to display supplementary attributes
 - Effectively create 3, 4, or higher dimensional plots
- Arrays of scatter plots compactly summarize factor relationships

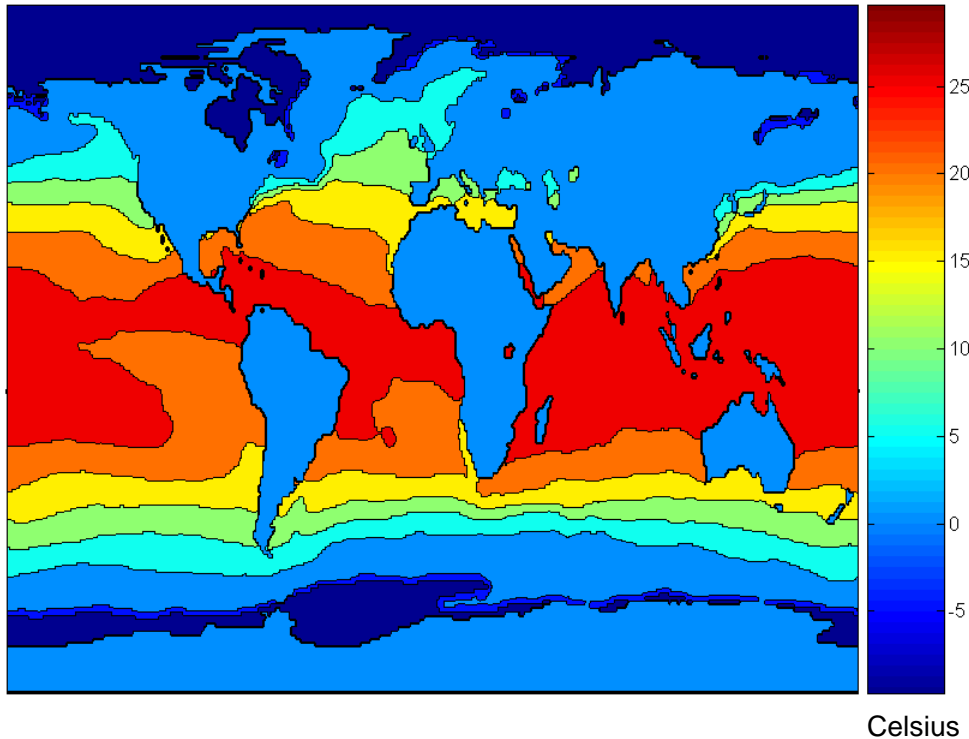
Scatter Plot Array of Iris Attributes



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: SST Dec, 1998



Questions?