LECTURE 2:
GETTING TO KNOW YOUR DATA

By: Jiawei Han (Additions and modifications: Siamak Sarmady)

Outline

- →Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary

Data Objects and Attribute Types

Introduction

- **Real-world data:** are typically noisy, enormous in volume, and originate from heterogeneous sources.
- **Questions:** when we get some data, we are likely to have some of these questions about it
  - What are the types of attributes or fields that makeup your data?
  - What kind of values the attributes have?
  - Are they continuous or discrete?
  - How are the values distributed?
  - How we can visualize them?
  - How we can spot outliers?
  - Can we measure the similarity of some data objects?

Types of Data Sets

- Record
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - Video data: sequence of images
  - Temporal data: time-series (e.g. rain, stock price)
  - Sequential Data: transaction sequences
  - Genetic sequence data
- Spatial, image and multimedia:
  - Spatial data: maps
  - Image data
  - Video data

Term-frequency 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

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

Transaction Data

Introduction

Main Tools:

- **Basic statistics:** helps in
  - Filling in missing values
  - Smoothing noisy values
  - Spotting outliers during preprocessing
  - Fixing inconsistencies incurred during data integration
- **Visualization:** helps in identifying relations, trends and biases by graphical means
- **Measuring similarity/dissimilarity:** Assume we have a database of patient data, describing them by their symptoms. Finding similarity or dissimilarity factor helps in:
  - Finding clusters of similar patients
  - Perform nearest neighbor classification (guess the disease for a new patient)
  - To detect outliers
  - To find possibly wrong diagnosis

Important Characteristics of Structured Data

- Dimensionality
  - **Curse of dimensionality:** If dimensions are too big, it will be difficult or sometimes impossible to perform specific types of operations (e.g. some classifiers won't work in proper time)
- Sparsity
  - **Only presence counts:** how much of the attribute space is used
- Resolution
  - **Patterns depend on the scale:** if enough resolution is not provided some patterns won't be found
- Distribution
  - **Centrality and dispersion:** how much the data has been dispersed, where is the middle

Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.
    - sales database: customers, store items, sales
    - medical database: patients, treatments
    - university database: students, professors, courses
  - Also called *samples , examples, instances, data points, objects, tuples (database rows)*.
  - Data objects are described by **attributes**.
- Database rows → data objects; columns → attributes.

Attributes

- **Attribute (data mining and database), dimension (data warehousing), feature (machine learning) or variables (statistics):** a data field, representing a characteristic or feature of a data object.
  - **Attribute vector:** a set of attributes used to describe a given object.
    - *E.g., customer _ID, name, address*
  - **Observations:** values observed for an attribute
  - **Distribution:** distribution variable is called univariate for one variable, bivariate for two etc.
- **Attribute Types:**
  - Nominal
  - Binary
  - Numeric

Attribute Types

- **Nominal:** means "relating to names" including categories, states, or “names of things”. In computer science they are called "enumerations".
  - *Hair_color = {black, blond, brown, grey, red, white}*
  - marital status, occupation, ID numbers, zip codes
  - Do not have meaningful order, mean, median etc.
  - Can be shown with numbers, but these numbers are not intended for quantitative use.
  - Qualitative
- **Binary:**
  - Nominal attribute with only 2 states (0 and 1). Called Boolean if the values are true and false.
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - **Convention:** assign 1 to most important outcome (e.g., HIV positive), 0 to least important (HIV Negative)
  - Qualitative

Attribute Types

- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size* = {*small*, *medium*, *large*}, grades, army rankings, lecturer ranks, satisfaction level
  - Central tendency can be represented by its mode and its median (middle value in ordered sequence) but mean cannot be defined.
  - Again if numbers are used they only represent codes, not values and should not be used for calculations.
  - Qualitative

Numeric Attribute Types

- Numerical (integer or real-valued): is quantitative, i.e. it is a measurable quantity
  - **Interval-scaled attributes**
    - Measured on a scale of **equal-sized units**
    - Values have order
      - E.g., *temperature in C°or F°, calendar dates*
    - Allows comparing and quantifying the difference between values
    - No true zero-point
    - But we cannot talk of a temperature being a multiple of another (e.g. 10 C° being two times warmer than 5 C°)
  - **Ratio-scaled attributes**
    - Numeric value with inherent **zero-point**
    - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K° is twice as high as 5 K°).
      - e.g., *temperature in Kelvin, monetary quantities, years of experience, number of words, weight, height, length, counts*

Discrete vs. Continuous Attributes

- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
    - E.g., zip codes, profession, or the set of words in a collection of documents, hair color, smoker
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
- **Continuous Attribute**
  - Has real numbers as attribute values
    - E.g., 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

Outline

- Data Objects and Attribute Types
- → Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary

Basic Statistical Descriptions of Data

Basic Statistical Descriptions of Data

- **Motivation:**
  - To better understand the data and identify its properties: central tendency, variation and spread
  - To identify which data values should be treated as noise or outliers
- **Data dispersion characteristics:**
  - Range, median, max, min, quantiles, outliers, variance, five number summary, etc.
- **Graphic Display of Basic Statistical Characteristics:**
  - Quantile plot, Quantile-quantile plot, Histogram, Scatter Plots and Correlation

Measuring the Central Tendency

- Mean is an algebraic measure (sample vs. population):
  - For Sample:

    $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (n is sample size)
  - For population:

    $\mu = \frac{\sum x}{N}$ (N is population size)
  - Weighted arithmetic mean:

    $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$ (for samples)
  - **Trimmed mean:** calculating mean after chopping extreme values

Measuring the Central Tendency

- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Median is expensive to compute when we have a large number of observations
  - **For grouped data:** estimated by interpolation

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

age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

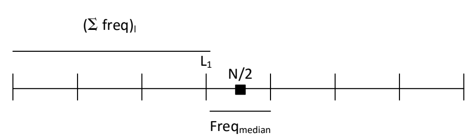
Median interval: 21–50

L1: lower boundary of the median interval

n: number of data in all of the entire dataset

(Σfreq)$_l$: sum of the frequencies of all of the intervals that are lower than the median interval

Width: width of the median interval



Measuring the Central Tendency

- **Mode:**
  - Value that occurs most frequently in the data
  - Can be determined for both qualitative and quantitative data
  - Unimodal, bimodal, trimodal (i.e. dataset with one, two r three modes)
  - If each data has a frequency of one, there is no mode
  - The mode for unimodal frequency curves that are moderately skewed can easily be approximated if mean and median are known:

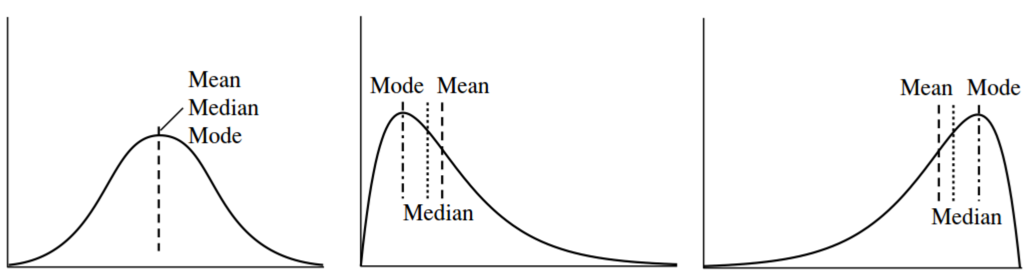
Empirical formula: **mean – mode ≈ 3 * (mean – median)**

- **Midrange:**
  - Is the average of the smallest and largest data:

(max()-mean())/2

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data



(a) Symmetric data

(b) Positively skewed data

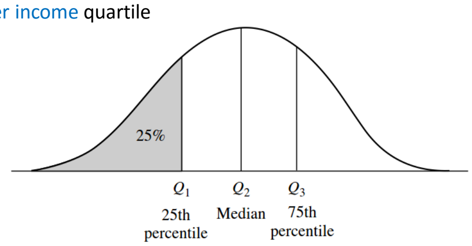
mode occurs at value smaller than the median

(c) Negatively skewed data

mode occurs at value larger than the median

Measuring the Dispersion of Data

- **Range:** max – min
- **Quantiles**: if data is sorted in increasing order and we divide the data into equal sections (based on the number of items in each section), the **points** are called quantiles.
- **2-Quantile:** the point which divides the data into two halves, is in fact the **median**
- **Quartiles**: 4-quantile are the three points that divide the data into four equal sections.
  - Give an indication of a distribution's center, spread and shape.
    - If there are 12 income observations (sorted in increasing order) , then the 3rd, 6th and 9th values are the quantiles.
    - There are 3 of people in lower income quartile



Measuring the Dispersion of Data

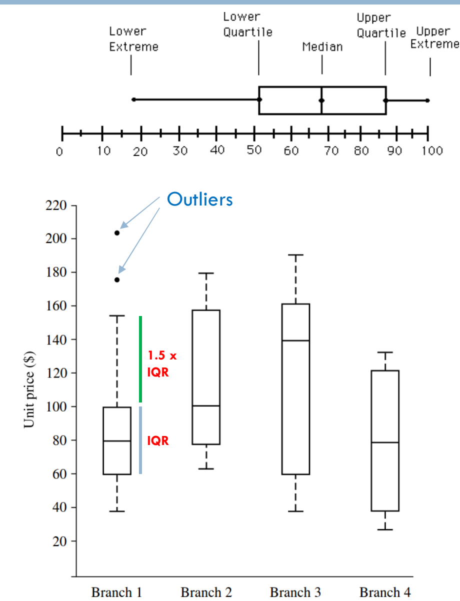
- **Percentiles:** the 99 points that divide the data into 100 parts.
  - Quartiles can be mentioned as Q_1 (25th percentile), Q_2 (50th percentile), Q_3 (75th percentile)
- **Inter-quartile range**: the distance between first and third quartiles i.e. IQR = $Q_3 - Q_1$
  - In the income example above, if Q1=$47000 and Q3=$63000 then the IQR = $63000-$47000 = $16000

Note: When we talk about quartiles, quantiles and percentiles, we are normally talking about data points, not a subset of data.

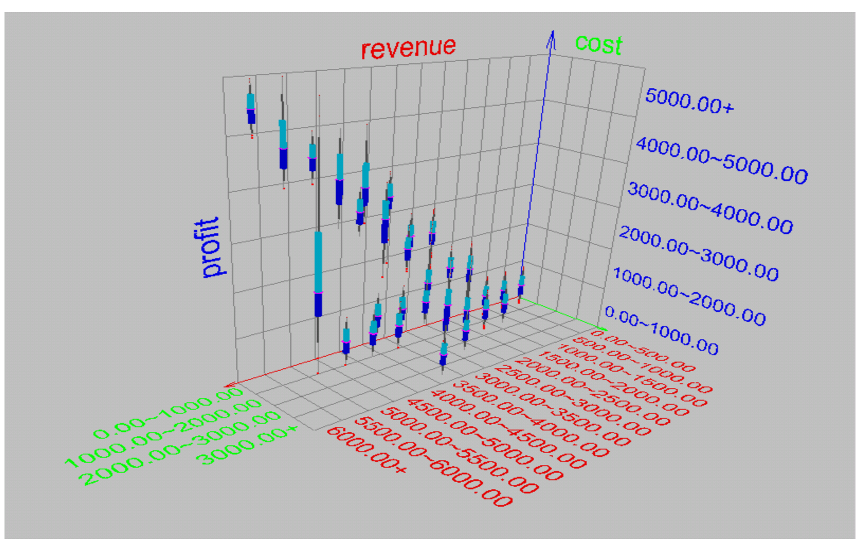
- **Five number summary**: Single metrics of spread are not enough for describing skewed distributions. "min, Q_1, median, Q_3, max" provide a fuller summary of the shape of distribution.

Boxplot Analysis

- Data is represented with a box
- The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
- The median is marked by a line within the box
- **Whiskers:** two lines outside the box extended to Minimum and Maximum
  - **Note:** whiskers are extended to the extreme low and high observations, only if these values are less than 1.5 x IQR beyond the quartiles.
- **Outliers:** points beyond a specified outlier threshold, plotted individually (values beyond 1.5 x IQR from quartiles)
- Boxplot can be computed in O(n log n) time (linear or sub linear time for estimates depending on the requirements).



Visualization of Data Dispersion: 3-D Boxplots



Measuring the Dispersion of Data

- Variance and standard deviation (*sample: s, population: σ*) are measures of data dispersion, in relation to mean.
  - **Variance**: (algebraic, scalable computation)

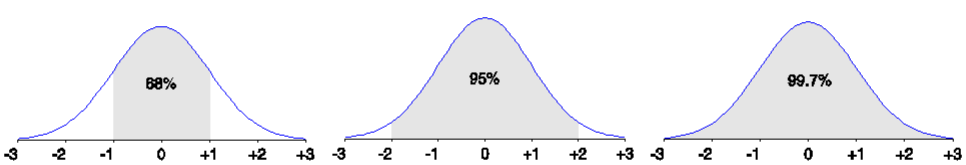
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2\right]$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = (\frac{1}{N}\sum_{i=1}^{n} x_i^2) - \mu^2$$

  - **Standard deviation:** *s (or σ)* is the square root of variance s^2 *(or σ^2)*
  - **Why a good dispersion measure:** An observation is unlikely to be more than several standard deviations away from the mean (proved using Chebyshev's inequality). Therefore the SD is a good indicator of the spread of a data set (See next page).

Properties of Normal Distribution Curve

- The normal (distribution) curve
  - From μ–σ to μ+σ: contains about 68% of the measurements (μ: mean, σ: standard deviation)
  - From μ–2σ to μ+2σ: contains about 95% of it
  - From μ–3σ to μ+3σ: contains about 99.7% of it

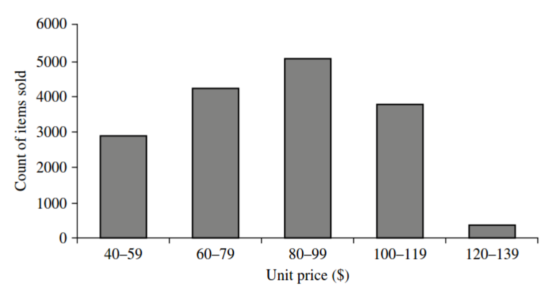


Graphic Displays of Basic Statistical Descriptions

- **Boxplot:** graphic display of five-number summary
- **Histogram:** x-axis are values, y-axis represents frequencies
- **Quantile plot:** each value x_i is paired with f_i indicating that approximately 100 f_i % of data are $\leq x_i$
- **Quantile-quantile (q-q) plot:** graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- **Scatter plot:** each pair of values is a pair of coordinates and plotted as points in the plane

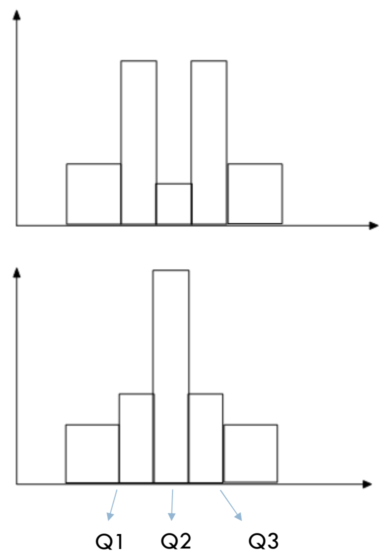
Histogram Analysis

- **Histogram:** shows the distribution of a given attribute (i.e. frequency in different values of an attribute)
  - It shows what proportion of cases fall into each of several categories (e.g. sales value boundaries)
  - The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent
- **Bar chart:** The height of each bar indicates the frequency of the attribute X for that value
- **General Histogram:** the *area* of the bar denotes the value, not the height (if the categories have different width)



Histograms Often Tell More than Boxplots

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

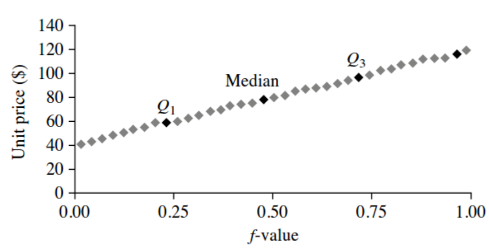


Quantile Plot

- **Quantile plot:** a simple and effective way to have a first look at univariate data distribution.
  - Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
  - Plots **quantile** information:
    - *1) x_i data items are* sorted in increasing order 2) Quantiles are specified
    - f_i indicates that approximately 100%*f_i of the data are below or equal to the value x_i

A Set of Unit Price Data for Items Sold at a Branch of *AllElectronics*

Unit price ($)	*Count of items sold*
40	275
43	300
47	250
–	–
74	360
75	515
78	540
–	–
115	320
117	270
120	350

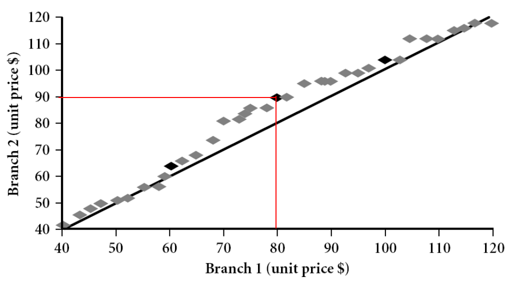


0% under $40
35% under $60
50% under $90
75% under $100
100% under $120

$$f_i = \frac{i-0.5}{N} \approx \frac{i}{N}$$

Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
  - View: Is there is a shift in going from one distribution to another?
  - Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.
  - Q1,Q2 (median),Q3 are plotted darker to provide comparison at specific points
  - If number of observations in one branch was less M<N, we could only draw M points on the q-q plot (interpolation might be needed for calculations)



50% of products sold at branch 2 are less than $90

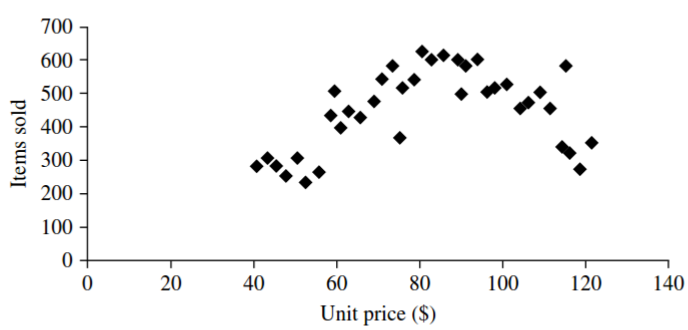
While

50% of products sold at branch 1 are less than $80

Above the line, branch 2 has better values, below it, branch 1 has better…

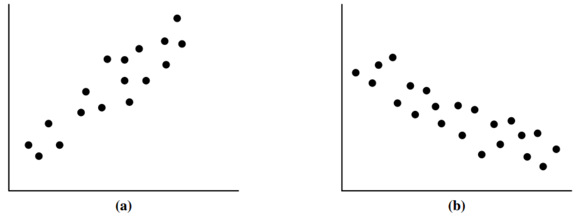
Scatter plot

- Effective visual method to determine whether a relationship, pattern or trend between two numeric attributes exists. It may also show the outliers.
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane

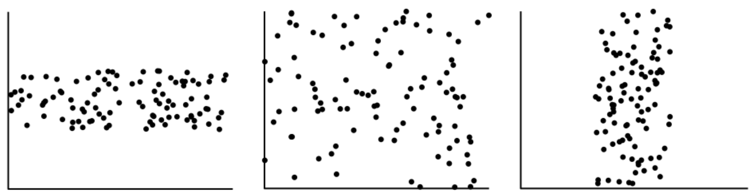


Positively and Negatively Correlated Data

- Scatter plot can also help to see whether the two attributes have correlations
  - The two attributes in (a) have positive correlation while the attributes in (b) have negative correlation



- Attributes with no correlation:



Outline

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- → Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary

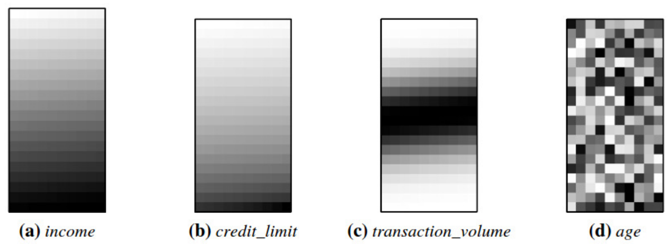
Data Visualization

Data Visualization

- Why data visualization?
  - Gain better insight of an information space by mapping data onto graphical primitives
  - Provide qualitative overview of large data sets
  - Search for patterns, trends, structure, irregularities, relationships among data
  - Help find interesting regions and suitable parameters for further quantitative analysis
  - Provide a visual proof of computer representations derived
- Categorization of visualization methods:
  - Pixel-oriented visualization techniques
  - Geometric projection visualization techniques
  - Icon-based visualization techniques
  - Hierarchical visualization techniques
  - Visualizing complex data and relations

Pixel-Oriented Visualization Techniques

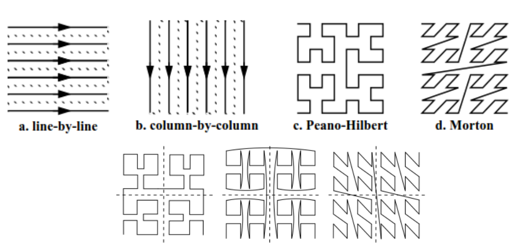
- For a data set of m dimensions, create m windows on the screen, one for each dimension
- Sort the data based on one of the attributes (income in below example)
- The m dimension values of a record are mapped to m pixels at the corresponding positions in the windows
- The colors of the pixels reflect the corresponding values

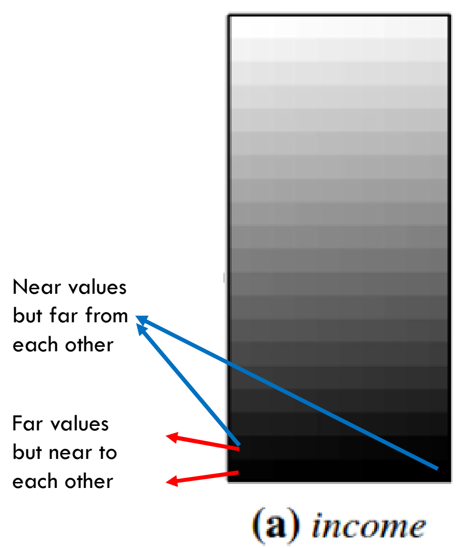


- Credit level values are distributed almost similar to income (correlation)
- Those with middle income have higher transactions.
- The age income does not show meaningful relation to income (at least visually)

Pixel-Oriented Visualization Techniques

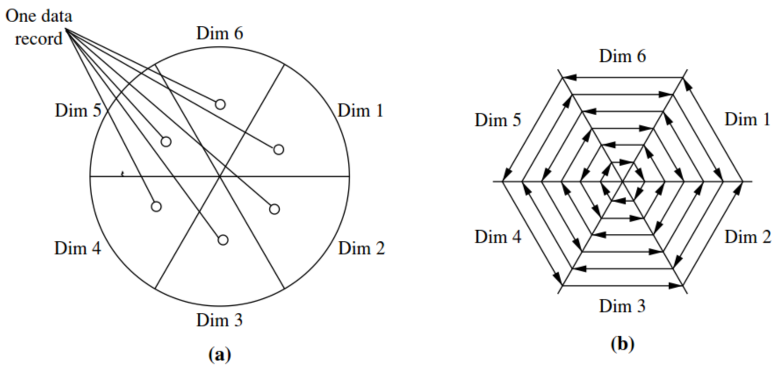
- Filing the windows by layering out the data records in linear way may not work well (specially for wide window)
- The first pixel in a row is far away from the last pixel in the previous row, though they are next to each other in the global order.
- We can lay out the data records using a space-filling curve (some may have less problem)



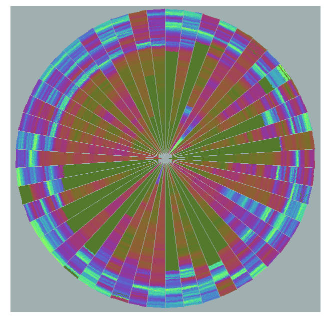


Laying Out Pixels in Circle Segments

- To save space and show the connections among multiple dimensions, space filling is often done in a circle segments (instead of separate windows).



The circle segment technique. (a) Representing a data record in circle segments. (b) Laying out pixels in circle segments.



Representing about 265,000 50-dimensional Data Items with the "Circle Segments" Technique

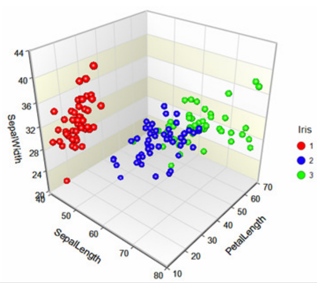
Geometric Projection Visualization Techniques

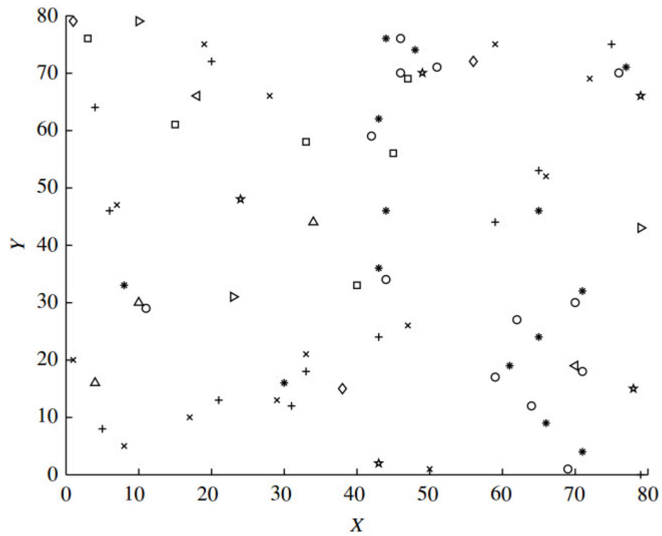
- **Limitation of pixel-oriented methods:** They cannot help in understanding the distribution of data in a multidimensional space (i.e. distribution of attributes in relation to others).
  - For example they do not show whether there is a dense area in a multidimensional subspace.
  - Distribution may not be quite recognizable from pixel based methods (even for a single attribute)

- **nD Scatter plots:** display the data points using n dimensions out of their total dimensions
  - **2D Scatter plot:** A scatter plot displays 2-D data points using Cartesian coordinates
  - **3D Scatter Plots:** can be built to show the relation of 3 attributes of data items
  - **4D Scatter Plot:** colors and shapes can be added to 3D scatter plot to show 4 dimensions
- **Scatter Plot Matrix:** for an N dimensional data scatter plots are usually ineffective. The scatter plot matrix is a n * n grid of 2-D scatter plots that provide visualization of the relation between each pair of attributes.

Visualizing Three Dimensions in a Window using Scatter Plot

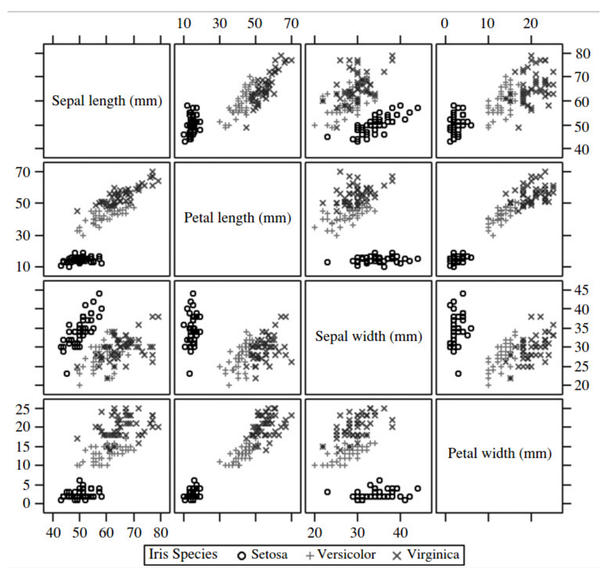
- **Two dimensions:** can be visualized in a scatterplot (as seen before).
- **Third dimension:** can be added using different shapes or colors
- **3D scatter plot:** can display points in the space
- **Fourth dimension:** can be added with colors





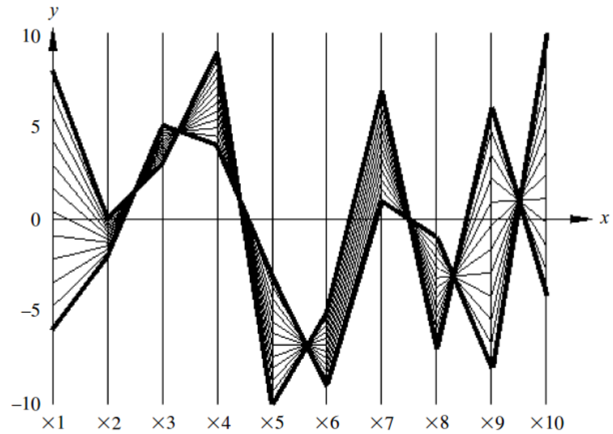
Scatter Plot Matrix

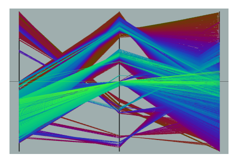
- For an N dimensional data scatter plots are usually ineffective. The scatter plot matrix is a n * n grid of 2-D scatter plots that provide visualization of the relation between each pair of attributes.
- **Data:** The Iris Flower data
- **Five dimensions:** length and width of sepal, length and width of petal, and species.
  - 4 dimensions are covered with scatter plot matrix
  - The relation with the 5th attribute (species type) is shown using shapes.



Parallel Coordinates

- Scatter plot matrix becomes less effective as the dimensionality increases.
  - Parallel coordinates can handle higher dimensionality
- **Axes:** Draws n equally spaced axes, one for each dimension, parallel to one of the display axes.
- **Items:** A data record is represented by a polygonal line that intersects each axis at the point that corresponds to the dimension value.
- **Limitation:** cannot show too many data points and becomes cluttered when number of items increase



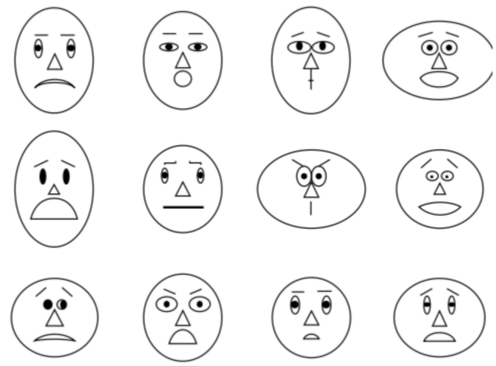


Icon-Based Visualization Techniques

- Visualization of the data values as features of icons
- **Typical visualization methods:**
  - **Chernoff Faces**
  - **Stick Figures**
- **General techniques:**
  - **Shape coding:** Use shape to represent certain information encoding
  - **Color icons:** Use color icons to encode more information
  - **Tile bars:** Use small icons to represent the relevant feature vectors in document retrieval

Chernoff Faces

- A way to display several variables on a two-dimensional surface.
- The figure shows faces produced using 10 characteristics head eccentricity, eye size, eye spacing, eye eccentricity, pupil size, eyebrow slant, nose size, mouth shape, mouth size, and mouth opening): Each assigned one of 10 possible values, generated using *Mathematica* (S. Dickson)



Stick Figure

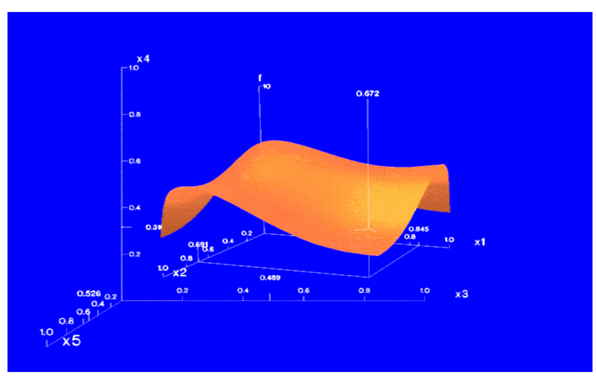
- A census data figure showing age, income, gender, education, etc.
- A 5-piece stick figure (1 body and 4 limbs w. different angle/length)

Hierarchical Visualization Techniques

- Visualization of the data using a hierarchical partitioning into subspaces
- Methods
  - Worlds-within-Worlds
  - Tree-Map
  - Cone Trees
  - InfoCube
  - Dimensional Stacking

Worlds-within-Worlds

- For a large data set of high dimensionality it is difficullt to visualize all dimensions at the same time.
- Assign the function and two most important parameters to innermost world
- Fix all other parameters at constant values - draw other (1 or 2 or 3 dimensional worlds choosing these as the axes)
- Software that uses this paradigm
  - N–vision: Dynamic interaction through data glove and stereo displays, including rotation, scaling (inner) and translation (inner/outer)
  - Auto Visual: Static interaction by means of queries

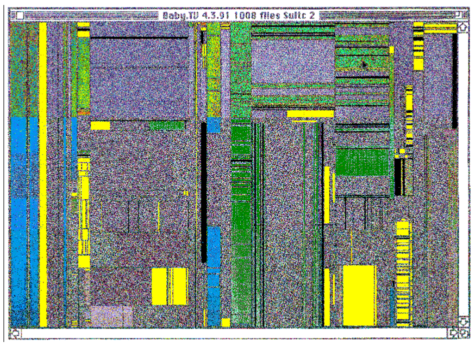


We interactively move the inner world, and see the change of the inner world.

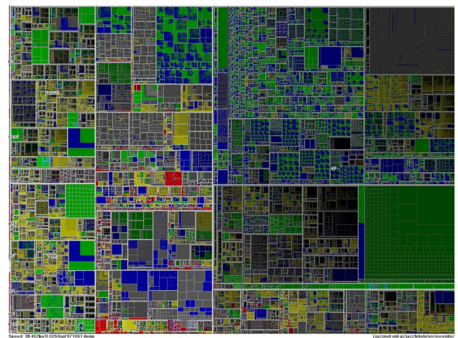
3 vars for the outer, 3 vars for the inner

Tree-Map

- Displays hierarchical data as a set of nested rectangles
- Uses a hierarchical partitioning of the screen into regions depending on the attribute values
- The x- and y-dimension of the screen are partitioned alternately according to the attribute values (classes)



Schneiderman@UMD: Tree-Map of a File System



Schneiderman@UMD: Tree-Map to support large data sets of a million items

Tree-Map for Visualizing Complex Data

- Use of Tree-maps to visualize Google news headline stories
- Stories organized into seven categories (each shown with a unique color).
- Within each major category, stories are further categorized into sub-categories.

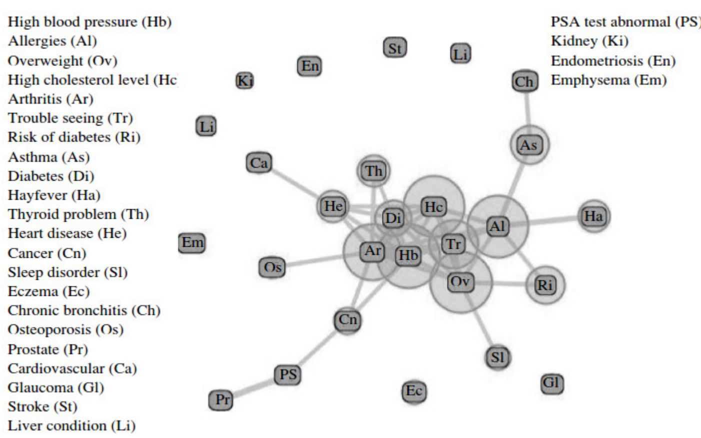
Newsmap: Google News Stories in 2005

Tags with Size

- Tag cloud: visualizing user-generated tags
  - Often listed alphabetically or a user-preferred order
  - The importance of tag is represented by font size/color
- Two ways to use:
  - When different tags refer to an item, the size of the tags may show the number of times it was applied to the item
  - When the tags are applied to different things, the size could show number of times the tag has been used (i.e. the popularity)

Distance Influence Graph

- The nodes in the graph are diseases and the size of each node is proportional to the prevalence of the corresponding disease.
- The nodes are linked if there is a strong correlation.
- The width shows the strength of the correlation.



Disease influence graph of people at least 20 years old in the NHANES data set.

Outline

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- → Measuring Data Similarity and Dissimilarity
- Summary

Measuring Data Similarity and Dissimilarity

Similarity and Dissimilarity

- **Why:** In data mining applications like clustering, outlier analysis, nearest neighbor classification (kNN), we need a way to assess how alike or unlike objects are in comparison to each other.
- **Cluster:** a collection of data objects such that the objects within a cluster are similar to one another and dissimilar to the objects in other clusters.
- **Outlier analysis:** potential outliers are highly dissimilar to other objects.
- **kNN classification:** in nearest neighbor classification, a given object (e.g. patient) is assigned a class label (e.g. a diagnosis) based on its similarity toward other objects in the model.
- **Measures of proximity:** a measure that represents how near or far two objects are to each other. These typically include similarity and dissimilarity.

Similarity and Dissimilarity

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

Data Matrix and Dissimilarity Matrix

- **Single Attribute Objects:** previously we calculated statistical parameters for single attributes, or objects with one dimension.
- **Multiple Attribute Objects:** for multi-attribute objects, we need to show the object with a vector: $x_1 = (x_{11}, x_{12}, x_{13}, \ldots, x_{1p})$
- Data matrix
  - n data points with p dimensions
  - Two modes
- Dissimilarity matrix
  - n data points, but registers only the distance
  - A triangular matrix (symmetric i.e. upper half has similar values) d(i,j) = d(j,i) and d(i,i)=0
  - Single mode (does not contain two entities or things i.e. rows and columns, just one kind of thing)

Attributes of One Object

$$\text{N objects} \quad \begin{bmatrix} \boxed{x_{11} & \cdots & x_{1f} & \cdots & x_{1p}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

$$\begin{bmatrix} 0 & d(1.2) & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ : & : & : & & \\ d(n,1) & d(n,2) & \ldots & \ldots & 0 \end{bmatrix}$$

Calculating Dissimilarity for Nominal Attributes

- **Nominal attribute:** can take M states, e.g., red, yellow, blue, green (generalization of a binary attribute)
- **Method 1:** dissimilarity is computed based on the ratio of mismatches.
  - *m:* number of attributes that match, *p:* total number of attributes

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

  - **Note:** if an attribute has larger number of states, it has higher potential for introducing dissimilarity.
  - Weights can be assigned to increase the effect of m or to assign greater weight the matches in attributes that have larger number of states

Calculating Dissimilarity for Nominal Attributes

- **Method 2:** a binary attribute is created to represent each of the M states of a nominal attribute. Now the binary new attributes are used to calculate dissimilarity.
  - One of the new attributes that matches the specific state of the object is set to 1, the remaining to 0.
    - What we do is to compare whether both objects are green? Both are blue? Both are red?...
- **Calculating similarity:**

$$sim(i,j) = 1 - d(i,j) = \frac{m}{p}$$

- **Example:** calculate dissimilarity matrix for the objects in the table

Object Identifier	test-1 (nominal)
1	code A
2	code B
3	code C
4	code A

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix} = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Calculating Dissimilarity for Binary Attributes

- Treating binary values as numeric can be misleading. A method specific to them is required.
- **Symmetric binaries:** if all binary attributes are considered with same weight and they are a symmetric binaries too, the following contingency table can be used.
  - q: number of attributes that are 1 for both objects
  - s, r: number of dissimilarities
  - t: number of attributes that are 0 for both objects
  - The dissimilarity of the two objects can then be calculated using:

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

Contingency Table for Binary Attributes

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

Calculating Dissimilarity for Binary Attributes

- **Asymmetric binaries:** for asymmetric binary attributes (two states are not equally important) like the outcome of a disease test (e.g. HIV positive and negative), the negative matches t, are considered unimportant and ignored:

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

t ignored

- **Similarity (Jaccard coefficient):** alternatively we can calculate asymmetric binary similarity between two objects:

$$sim(i,j) = \frac{q}{q+r+s} = 1 - d(i,j)$$

- **Coherence:** Jaccard coefficient is the same as coherence

$$coherence(i,j) = \frac{sup(i,j)}{sup(i) + sup(j) - sup(i,j)} = \frac{q}{(q+r)+(q+s)-q}$$

Calculating Dissimilarity 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

  - Gender is a symmetric attribute, the remaining attributes are asymmetric binary
  - For asymmetric values, let the values Y and P be 1, and the value N be 0
  - Suppose that the distance between two objects (patients) is computed based on only asymmetric values.

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

  - The measure suggests that Jim and Mary are unlikely to have a similar disease because they have the highest dissimilarity. Jack and Marry are more likely to have similar disease.

Distance on Numeric Data – Normalizing (Standardizing) Numeric Data

- In some cases the numeric data are normalized before applying distance calculations.
  - The data is transformed to fall within a smaller distance such as [-1,1] or [0.0, 1.0].
  - That is because if multiple attributes are involved, the attributes with larger size could have larger and unfair effect on the dissimilarity calculation.
- **Z-score:** every value is transformed using $z = \frac{x - \mu}{\sigma}$
  - X: raw score to be standardized, μ: mean of the population, σ: standard deviation
  - the distance between the raw score and the population mean in units of the standard deviation
  - negative when the raw score is below the mean, "+" when above

Distance on Numeric Data – Normalizing (Standardizing) Numeric Data

- An alternative way: Calculate the mean absolute deviation

$$S_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|)$$

where

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \cdots + x_{1f})$$

and standardized measure (*z-score*):

$$Z_{if} = \frac{x_{if} - m_f}{S_f}$$

- Using mean absolute deviation is more robust than using standard deviation

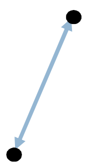
Distance on Numeric Data – Euclidian Distance

- Distance between an object with several numerical attributes can be measured using Euclidian distance

point	attribute1	attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

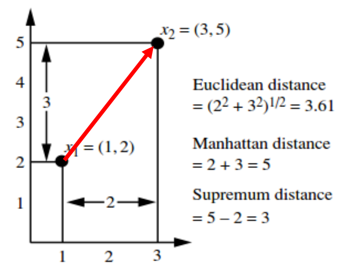
$$d(i,j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{ip} - x_{jp}|^2}$$

- Example: calculate dissimilarity matrix of the provided data



Dissimilarity Matrix (with Euclidean Distance)

	x1	*x2*	*x3*	*x4*
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0



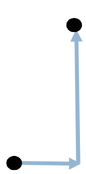
Distance on Numeric Data – Manhattan Distance

- Distance between an object with several numerical attributes can be measured using Manhattan (city block) distance too. It resembles the walking distance between to places in a city.

point	attribute1	attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

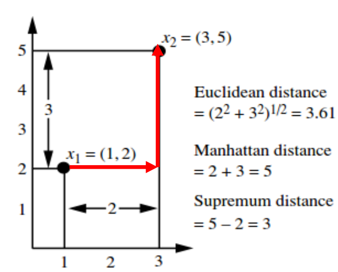
$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|$$

- Example: calculate dissimilarity matrix of the provided data



Dissimilarity Matrix (with Manhattan Distance)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0



Distance on Numeric Data – Euclidian and Manhattan Distance Properties

- Both Euclidian and Manhattan distance satisfy the following mathematical properties
  - **Non-negativity:** d(i,j) ≥ 0
  - **Identity of indiscernible:** d(i,i) = 0 (distance of an object to itself is 0)
  - **Symmetry:** d(i,j) = d(j,i) (distance is a symmetric function)
  - **Triangle inequality:** d(i,j) ≤ d(i,k) + d(k,j) (going direct to another object is not larger than any detour over any other object k)
- **Metric:** A measure that satisfies these conditions is known as metric.

Distance on Numeric Data – Minkowski Distance

- Minkowski distance is generalization of the Euclidian and Manhattan distances:

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

  - h: real and h≥1
- Such distance is called L_p norm in which the symbol p refers to the h in above formula.
  - L_1 norm is therefore Manhattan distance and L_2 refers to Euclidian distance

Distance on Numeric Data – Supremum or Chebyshev Distance

- Supremum distance (L_{max}, L_∞ norm, Chebyshev distance) is a generalization of the Minkowski distance for h→∞.

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

- To find it we find the attribute that gives the maximum difference in values between the objects. The attribute f has the largest distance among the attributes of the objects i and j.

Distance on Numeric Data – Weighted Euclidian Distance

- **Weighted Euclidian Distance:** If each attribute is assigned a weight according to its perceived importance, the weighted Euclidian distance can be computed as

$$d(i,j) = \sqrt{w_1\left|x_{i1} - x_{j1}\right|^2 + w_2\left|x_{i2} - x_{j2}\right|^2 + \cdots + w_m\left|x_{ip} - x_{jp}\right|^2}$$

  - Weighting can be applied to other distance measure as well.

Distance on Ordinal Data

- **Ordinal Data:** values have meaningful order but the difference between successive values is unknown (e.g. small, medium, large, for a size attribute)
- **Distance:** ordinal attributes are treated similar to the numeric attributes
  - The attribute f for the i-th object is x_{if} , we first replace the value of with its rank $r_{if} \in \{1,\ldots,M_f\}$
  - We normally map the range of each attribute to [0.0,1.0] so that attributes have equal weight

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

Value of attribute for this object

Maximum value of the attribute

  - Dissimilarity can then be computed using any of the distance measures described for numeric values

Distance on Ordinal Data

- **Example:** assume we have performed a test for different objects and we have obtained the results of the table

Object Identifier	*test-2 (ordinal)*
1	excellent
2	fair
3	good
4	excellent

  - There are three states for test-2: fair, good, excellent. We replace the results with rank numbers (3,1,2,3 respectively)
  - We normalize the values: 1→0.0, 2→0.5, 3→1.0
  - Now we can use any distance type, e.g. Euclidian distance to measure the distance of two objects

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

Distance of Objects with Mixed Attributes

- In many database, objects are described by a mixture of attribute types. There are different approaches for computing dissimilarity:
- **Approach 1:** Group attributes with similar types together, perform separate data mining (e.g. clustering) analysis for each type
  - This is feasible if the analysis results produces compatible results for all groups, but this is unlikely in most scenarios.
- **Approach 2:** Process all attribute types together and perform a single analysis. Different techniques are used. One technique is to bring all of the meaningful attributes onto a common scale of the interval [0.0,1.0]. (next page)

Distance of Objects with Mixed Attributes

- The dissimilarity d(i,j) between objects i and j is defined as:

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

  - $\sum_{f=1}^{P} \ldots$: for all attributes of i and j
  - $\delta_{ij}^{(f)}$: is 0 if
    - x_{if} or x_{jf} is missing. So if either is missing, the attribute is not taken into account
    - $x_{if}=x_{jf}=0$ and attribute f is asymmetric binary (i.e. the negative value of 0 is not important)
  - **Numerator:** calculates differences
  - **Denominator:** counts the number of attributes
  - Contribution of attribute f to the dissimilarity $d_{ij}^{(f)}$ depends on its type (next page)

Distance of Objects with Mixed Attributes

- The dissimilarity d(i,j) …
  - Contribution of attribute f to the dissimilarity $d_{ij}^{(f)}$ depends on its type
    - **Numeric:** use normalized distance i.e. $\frac{|x_{if}-x_{jf}|}{max_h x_{hf} - min_h x_{hf}}$ where h runs over all non-missing objects, for attribute f
    - **Nominal or binary:** 0 if $x_{if}=x_{jf}$, otherwise 1
    - **Ordinal:** compute the ranks and normalize the ranks to [0,1] and treat as numeric
- Example: calculate the difference for objects in the table
  - For each type calculate the dissimilarity matrix i.e. $d_{ij}^{(1)}$, $d_{ij}^{(2)}$, $d_{ij}^{(3)}$ (note that two of them was done before)

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

Distance of Objects with Mixed Attributes - Example

- Example: continued…
  - For each type calculate the dissimilarity matrix i.e. $d_{ij}^{(1)}$, $d_{ij}^{(2)}$, $d_{ij}^{(3)}$ (note that two of them was done before, we just need to compute for test-3)
  - In order to calculate $d_{ij}^{(3)}$ we need the following: $max_h x_h$ = 64 and $min_h x_h$ = 22 (used for normalizing the distance between test-3 attributes)
  - Now we calculate elements of the dissimilarity matrix for test-3

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

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

  - Now we can calculate elements of final dissimilarity matrix e.g. $d(3,1) = \frac{1(1)+1(0.50)+1(0.45)}{3} = 0.65$

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

Cosine Similarity

- **Term-frequency vector:** A document can be represented by thousands of attributes (a word and its frequency).

Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

  - Very long and sparse (out of 30,000 possible words, only a few hundred are used)
  - Traditional distance measure that we studied don't work well for sparse numeric data because two documents might not share many words, but that does not mean they are similar!!
  - We need a measure that focuses on the words that the two documents have in common
- **Applications:** information retrieval, text document clustering, biological taxonomy, gene feature mapping

Cosine Similarity

- **Cosine Similarity:** gives a ranking of documents with respect to a given "vector of query words"
  - Let x and y be the two vectors being compared.

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

    - x.y is dot multiplication of the two vectors
    - $||x||$ is the Euclidian norm of vector x = (x_1,x_2,…,x_p) defined as $\sqrt{x_1^2 + x_2^2 + \ldots + x_p^2}$
    - The similarity factor measures the cosine of the angle between the two vectors
    - Similarity of 0 means they are orthogonal to each other (90 degrees). Greater Cosine (near to 1) means higher similarity.
    - Cosine similarity is not a metric measure because it des not satisfy all the properties of such measure.

Cosine Similarity

- **Example:** Suppose x and y are the two first term-frequency vectors in below table. How similar are x and y?

Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

$$\begin{aligned} \boldsymbol{x}^t \cdot \boldsymbol{y} &= 5\times3+0\times0+3\times2+0\times0+2\times1+0\times1+0\times0+2\times1 \\ &\quad +0\times0+0\times1 = 25 \\ ||\boldsymbol{x}|| &= \sqrt{5^2+0^2+3^2+0^2+2^2+0^2+0^2+2^2+0^2+0^2} = 6.48 \\ ||\boldsymbol{y}|| &= \sqrt{3^2+0^2+2^2+0^2+1^2+1^2+0^2+1^2+0^2+1^2} = 4.12 \end{aligned}$$

$$sim(\boldsymbol{x},\boldsymbol{y}) = 0.94$$

Cosine Similarity – Binary Attributes

- For binary valued attributes, the cosine similarity function is interpreted in terms of shared features or attributes.

- Sim(x,y) is a measure of relative possession of common attributes.

$$sim(x,y) = \frac{x.y}{x.x + y.y - x.y}$$

  - x.y will be equal to the number of attributes possessed by both x and y (if either is 0 the whole term is removed in x.y)
  - |x||y| is the geometric mean of the number of attributes possessed by x and the number possessed by y.

- The function is known as Tanimoto coefficient or distance, is frequency used in information retrieval and biology taxonomy.

Outline

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- → Summary

Summary

- Data attribute types: nominal, binary, ordinal, interval-scaled, ratio-scaled
- Many types of data sets, e.g., numerical, text, graph, Web, image.
- Gain insight into the data by:
  - **Basic statistical data description:** central tendency, dispersion, graphical displays
  - **Data visualization:** map data onto graphical primitives
  - **Measure data similarity:** calculate the difference of attribute pairs and then the total difference
- Above steps are the beginning of data preprocessing
- Many methods have been developed but still an active area of research

References

- W. Cleveland, Visualizing Data, Hobart Press, 1993
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- U. Fayyad, G. Grinstein, and A. Wierse. Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- H. V. Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Tech. Committee on Data Eng., 20(4), Dec. 1997
- D. A. Keim. Information visualization and visual data mining, IEEE trans. on Visualization and Computer Graphics, 8(1), 2002
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- S. Santini and R. Jain," Similarity measures", IEEE Trans. on Pattern Analysis and Machine Intelligence, 21(9), 1999
- E. R. Tufte. The Visual Display of Quantitative Information, 2nd ed., Graphics Press, 2001
- C. Yu et al., Visual data mining of multimedia data for social and behavioral studies, Information Visualization, 8(1), 2009