CS423: Data Mining

Basic concepts

Jakramate Bootkrajang

Department of Computer Science
Chiang Mai University

Outline

- Data point and Feature
- Basic Statistical Descriptions of Data
- Data Visualisation
- Measuring Data Similarity and Dissimilarity
- Summary

Data point

- A 'data point' represents an entity.
  - Data points are described by feature/attributes.
  - Also called, example, input instance, or simply point.
- Feature (or dimensions or attributes)
  - A field representing state of nature or characteristic of a data point.
- Example: a dog data point with 4 features namely colour, height, weight and temperament
  - $x_i = \{black, 56, 34.2, gentle\}$

Dataset

- A 'dataset' is made up of data points.
- Examples
  - Patient dataset containing information of 100 patients.
  - Dog dataset containing information of 40 dogs.

Data matrix

- Mathematically, a data is a M dimensional vector (point) in some space (e.g., Euclidean)
  - $\mathbf{x}_i = \{x_i^1, x_i^2, \ldots, x_i^M\}$. Note bold font means **x** is a vector.

- So, a dataset of N data points is an $N \times M$ matrix.

$$\begin{bmatrix} x_1^1 & \ldots & x_1^m & \ldots & x_1^M \\ \vdots & \ddots & \vdots & & \vdots \\ x_n^1 & \ldots & x_n^m & \ldots & x_n^M \\ \vdots & & \vdots & \ddots & \vdots \\ x_N^1 & \ldots & x_N^m & \ldots & x_N^M \end{bmatrix}$$

Feature Types

- Nominal (categorical): categories, states, or names of things
  - HairColor = black, blond, brown, grey, red, white
  - marital status, occupation
- Binary: nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes are equally important (e.g., gender)
  - Asymmetric binary: outcomes are not equally important (e.g., medical test)

Feature Types [2]

- Ordinal
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - Size = small, medium, large, grades, army rankings
- Numeric
  - Quantity (integer or real-valued)

Outline

- Data point and Feature
- Basic Statistical Descriptions of Data
- Data Visualisation
- Measuring Data Similarity and Dissimilarity
- Summary

Basic Statistical Descriptions of Data

- Motivation
  - To better understand the data: central tendency, variation and spread.
- Central tendency
  - Mean, median, mode
- Variation and spread
  - Variance, standard deviation, quantile.

Measuring the Central Tendency

- Mean
  - Arithmetic mean: $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$
  - Weighted arithmetic mean: $\bar{x} = \frac{1}{N} \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}$
  - Trimmed mean: chopping extreme values before calculating the mean
- Median
  - Sort the dataset in an increasing order
  - Take the middle value if there is odd number of data points
  - Take an average of the middle two values if there is even number of data points

Measuring the Central Tendency [2]

- Mode
  - Value that occurs most frequently in the data
  - A dataset can be unimodal (1 mode), bimodal (2modes) or more.
- When working with vectorial data, calculate the central tendency on each of the dimension.
  - $\bar{\mathbf{x}} = \{\bar{x}^1, \bar{x}^2, \ldots, \bar{x}^M\}$

Measuring dispersion of data

- Variance and standard deviation
- Variance $= \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$
  - Standard deviation σ is the square root of variance
- Quantile: set of P-1 points that divide the range of probability distribution into P intervals (with equal probability density)
  - When P=2: The point is called a Median
  - When P=4: each of the points is a Quartile
  - When P=100: each of the points is a Percentile

Measuring dispersion of data [2]

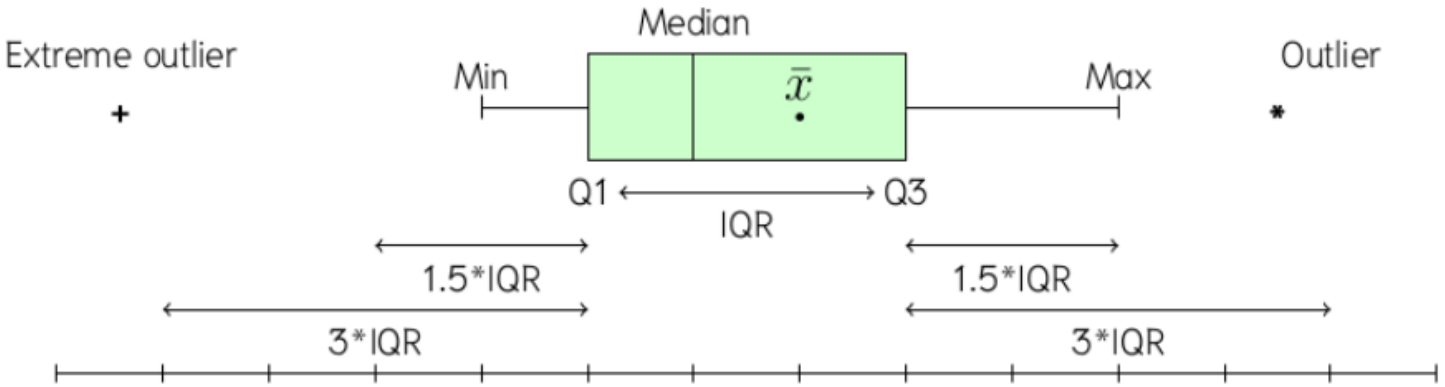
- When working with vectorial data we can calculate covariance
- Covariance indicates the joint variability of two features.
- Recall variance $= \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$
- Covariance $= cov(x^j, x^k) = \frac{1}{N}\sum_{i=1}^{N}(x_i^j - \bar{x}^j)(x_i^k - \bar{x}^k)$
- Covariance matrix is a matrix summarising covariances of between pairs of features

$$\begin{bmatrix} cov(x^1, x^1) & \dots & cov(x^1, x^m) & \dots & cov(x^1, x^M) \\ \vdots & \ddots & \vdots & & \vdots \\ cov(x^M, x^1) & \dots & cov(x^m, x^1) & \dots & cov(x^M, x^M) \end{bmatrix}$$

Measuring dispersion of data [3]

- Inter-quartile range: IQR = Q3 - Q1
- Five numbers summary
  - Five numbers are: min, Q1, median,Q3, max
- Outlier:
  - usually, a value greater than Q3 + 1.5 x IQR
  - or less than Q1 - 1.5 x IQR

Five Numbers Summary



รูปภาพ 2.1: แผนภาพแบบกล่องและตัวเลขทางสถิติ 5 ตัว รวมถึงค่าผิดปกติต่างๆ (อัตราส่วนอาจไม่ตรง)

Outline

- Data point and Feature
- Basic Statistical Descriptions of Data
- Data Visualisation
- Measuring Data Similarity and Dissimilarity
- Summary

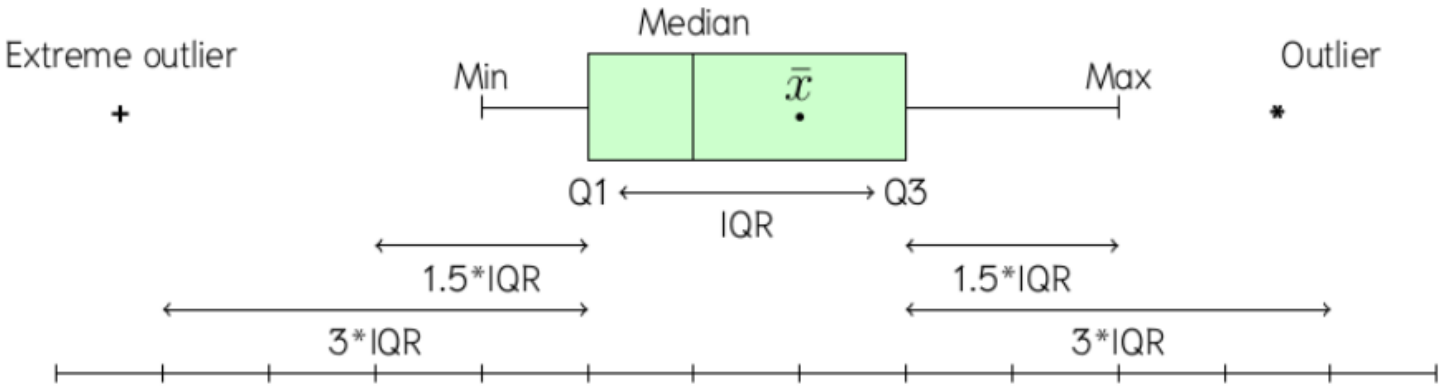
Data Visualisation

- Why data visualisation?
  - Gain insight into an information space by mapping data onto graphical primitives
  - Provide qualitative overview of large data sets
  - Help find interesting regions and suitable parameters for further quantitative analysis
- There are numerous methods:
  - Basic statistical visualisation techniques
  - Icon-based visualisation techniques
  - Hierarchical visualisation techniques

Basic Statistical Visualisation

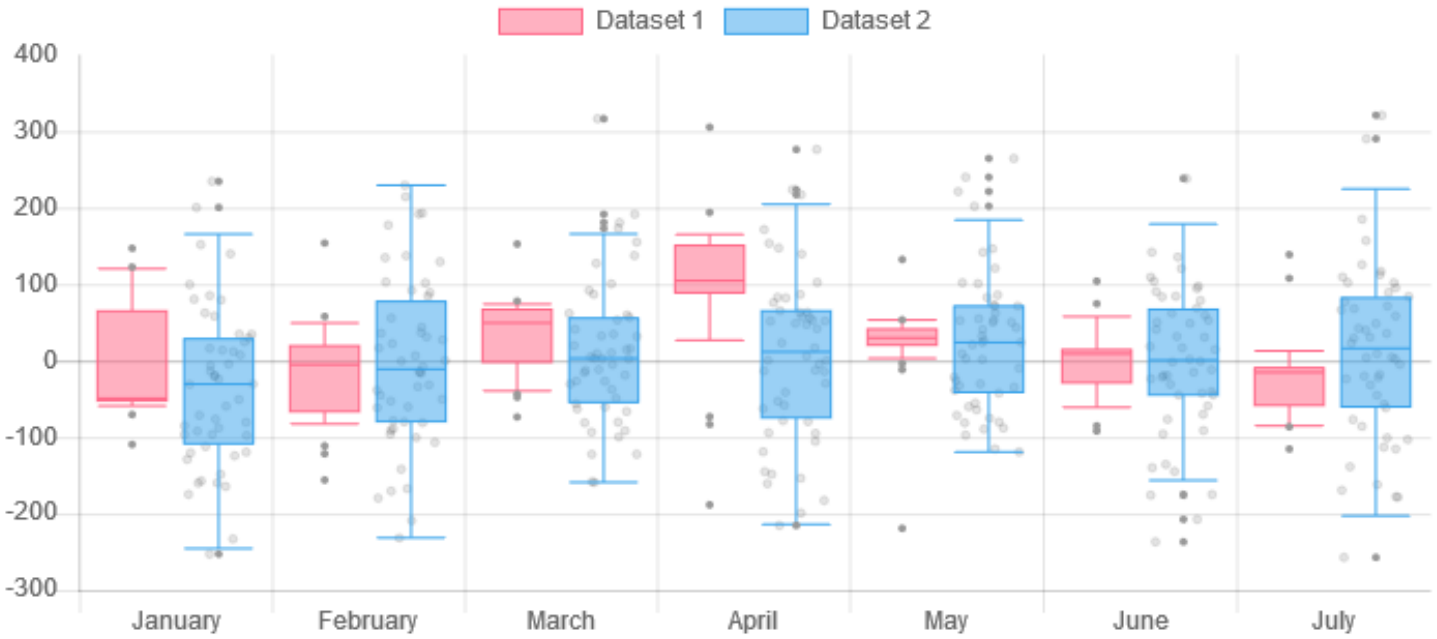
- Boxplot: graphic display of five-number summary
  - Histogram: x-axis are values, y-axis represents frequencies
  - Scatter plot: each pair of values is a pair of coordinates and plotted as points in the plane

Boxplot



รูปภาพ 2.1: แผนภาพแบบกล่องและตัวเลขทางสถิติ 5 ตัว รวมถึงค่าผิดปกติต่างๆ (อัตราส่วนอาจไม่ตรง)

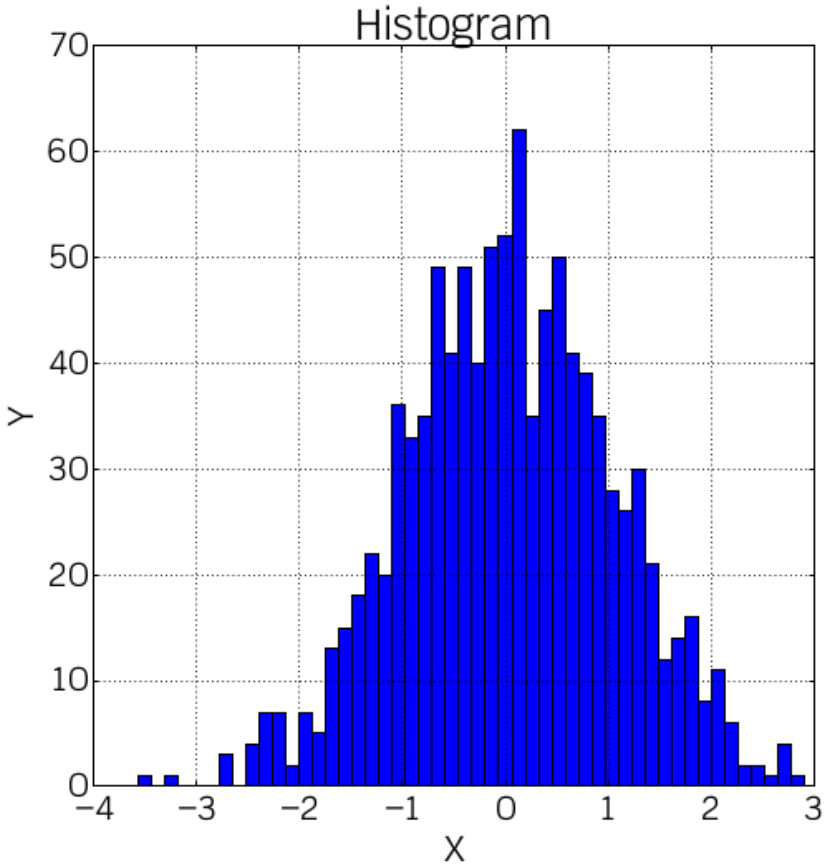
Boxplot [2]


Dataset 1
Dataset 2
400
300
200
100
0
-100
-200
-300
January
February
March
April
May
June
July

Histogram

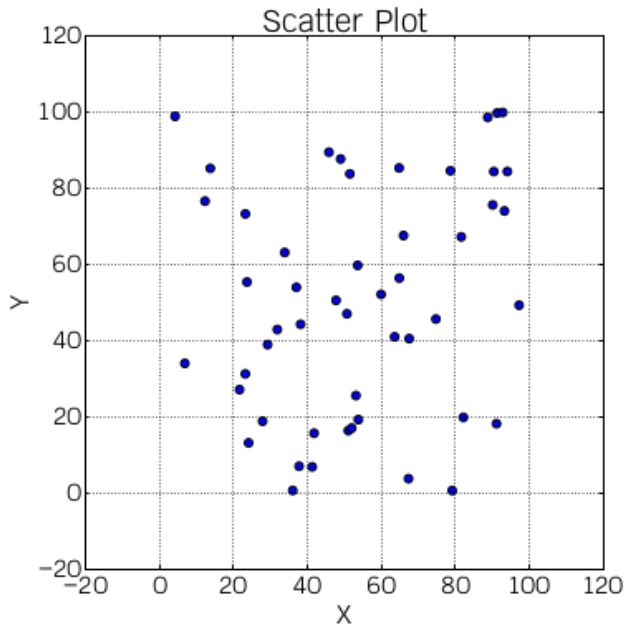
- Histogram: Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of bars
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent

Histogram [2]



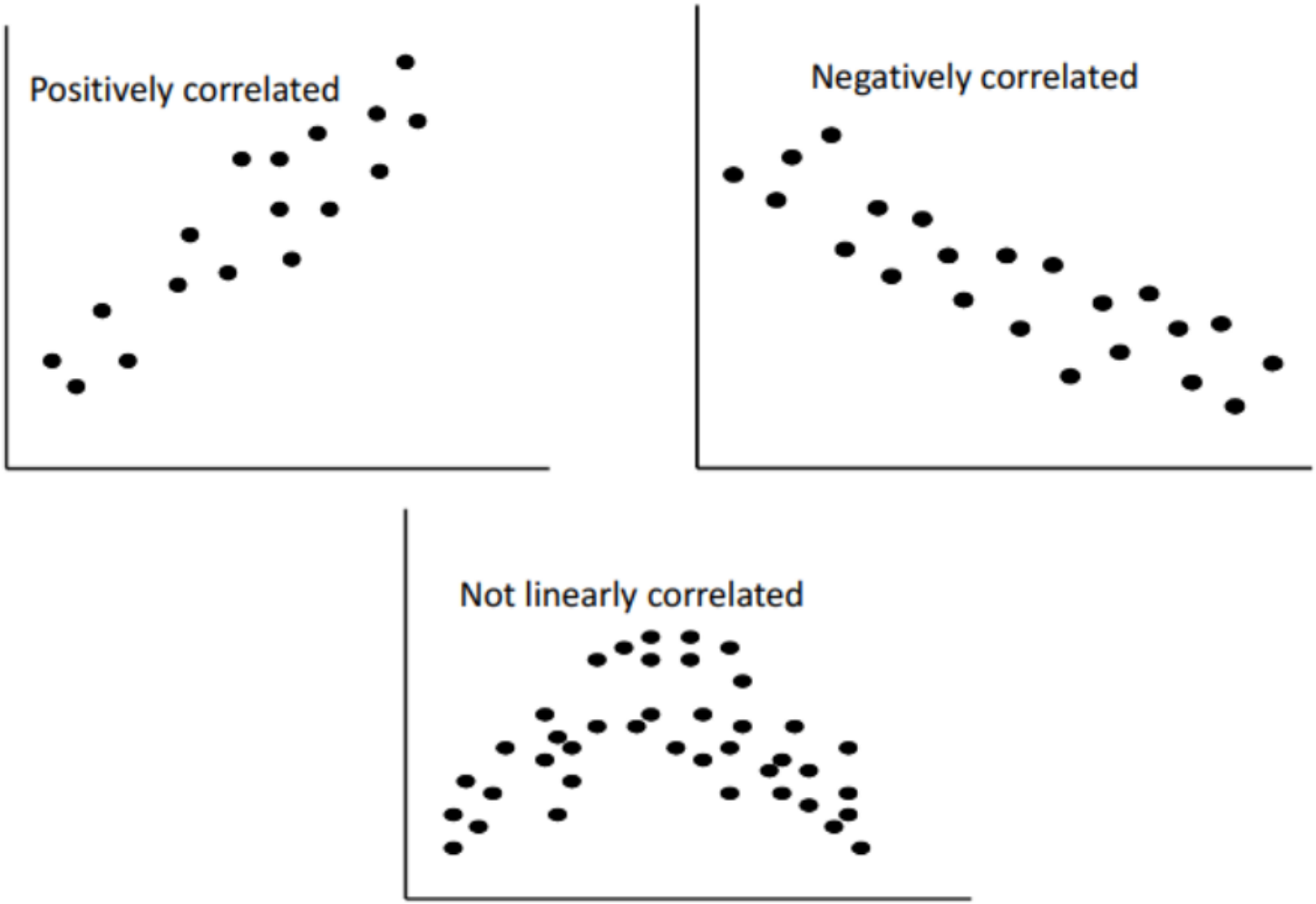
Scatter plot

- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane

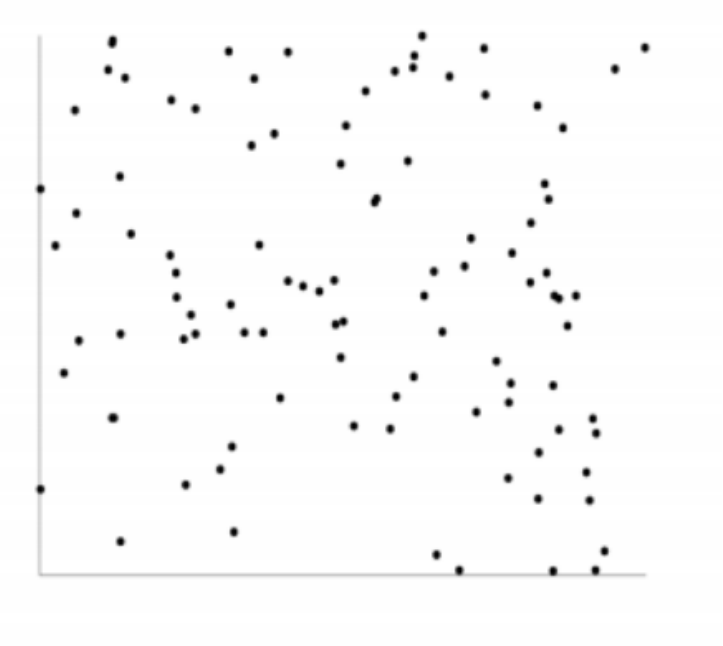


Scatter plot [2]

- Useful for observing correlation

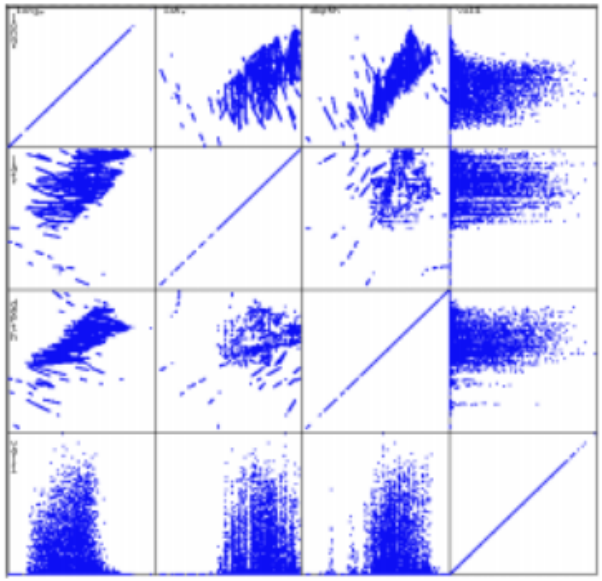


Scatter plot [3]



Scatter plot for vectorial data

- Projection data onto lower dimension, i.e. 2-D
- Do scatter plot for every pair of dimensions in turns
- This will result in a matrix of scatter plots

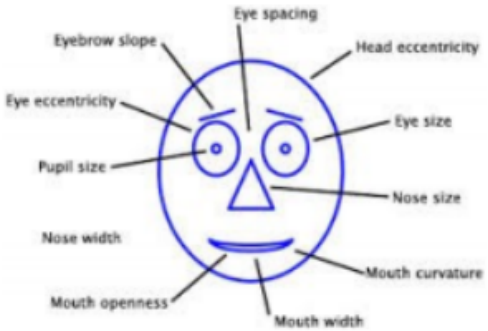


Icon-based Visualisation Techniques

- Visualisation of the data values as features of icons
- Typical visualisation methods
  - Chernoff Faces
- General techniques
  - Shape coding: Use shape to represent certain information encoding
  - Color icons: Use color icons to encode more information

Chernoff Faces

- A way to display variables on a two-dimensional surface, e.g., let x be eyebrow slant, y be eye size, z be nose length, etc.
- 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)



Chernoff Faces for cereal data

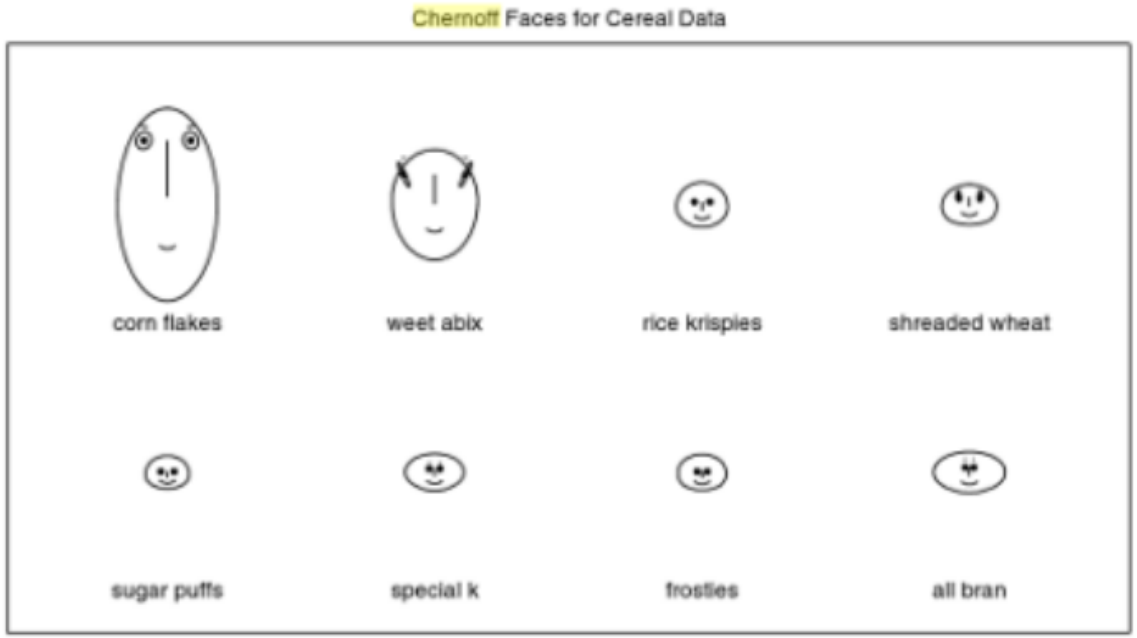


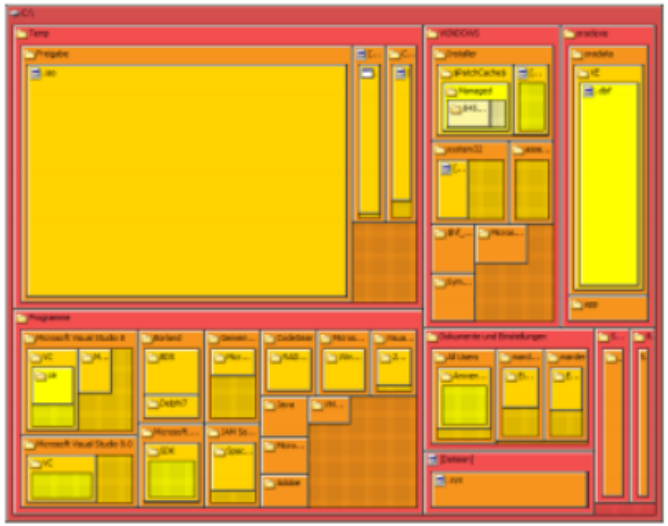
FIGURE 10.1
This shows the Chernoff faces for the **`cereal`** data, where we have 8 observations and 11 variables. The shape and size of various facial features (head, eyes, brows, mouth, etc.) correspond to the values of the variables. The variables represent the percent agreement to statements about the cereal. The statements are: comes back to, tastes nice, popular with all the family, very easy to digest, nourishing, natural flavor, reasonably priced, a lot of food value, stays crispy in milk, helps to keep you fit, fun for children to eat.

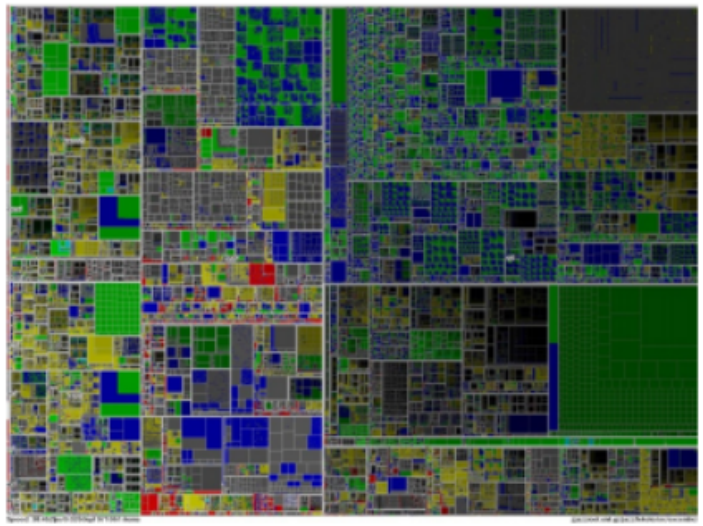
Hierarchical Visualisation Techniques

- Visualisation of the data using a hierarchical
- partitioning into subspaces
- Some of the methods
  - Tree-Map
  - InfoCube

Tree-map

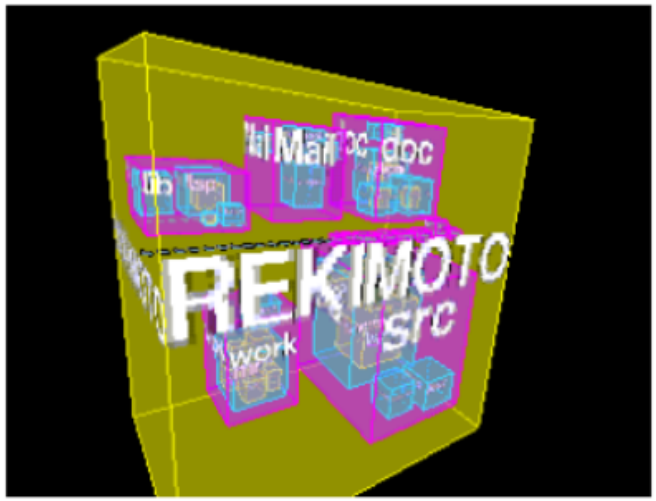
- Visualisation of hierarchical structure (think of tree).
- The method renders value of leaf node using different size and colour depending on node's value.





Info-cube

- A 3-D visualisation technique where hierarchical information is displayed as nested semi-transparent cubes
- The outermost cubes correspond to the top level data, while the sub-nodes or the lower level data are represented as smaller cubes inside the outermost cubes, and so on



Outline

- Data point and Feature
- Basic Statistical Descriptions of Data
- Data Visualisation
- Measuring Data Similarity and Dissimilarity
- Summary

Similarity and Dissimilarity

- Useful concept used in many algorithms and tasks.
- 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

Dissimilarity matrix

- Dissimilarity matrix (distance matrix)
  - Represents pair-wise distance between 2 data points.
  - A triangular matrix

$$\begin{bmatrix} 0 & & & & \\ d(\mathbf{x}_1, \mathbf{x}_2) & 0 & & & \\ d(\mathbf{x}_1, \mathbf{x}_3) & d(\mathbf{x}_2, \mathbf{x}_3) & 0 & & \\ \vdots & \vdots & \vdots & 0 & \\ d(\mathbf{x}_1, \mathbf{x}_N) & d(\mathbf{x}_2, \mathbf{x}_N) & \dots & d(\mathbf{x}_{N-1}, \mathbf{x}_N) & 0 \end{bmatrix}$$

Dissimilarity Measure for Binary Features

- For vectorial data, it can be summarised using a contigency table

	1	0
1	q	r
0	s	t

- q is the number of times when both vectors have value 1 (in the same position), and so on for r, s, t
- Distance measure for symmetric binary variables:
  - $d(\mathbf{x}_i, \mathbf{x}_j) = \frac{r+s}{q+r+s+t}$
- Distance measure for asymmetric binary variables:
  - $d(\mathbf{x}_i, \mathbf{x}_j) = \frac{r+s}{q+r+s}$

Example: Distance measure for Binary

Name	Gender	Fever	Cough	Test1	Test2	Test3	Test4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is symmetric but the rest are asymmetric binary features.
- Let the values Y and P be 1, and the value N be 0
  - 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) $=$???

Distance on Numeric Data: Minkowski Distance

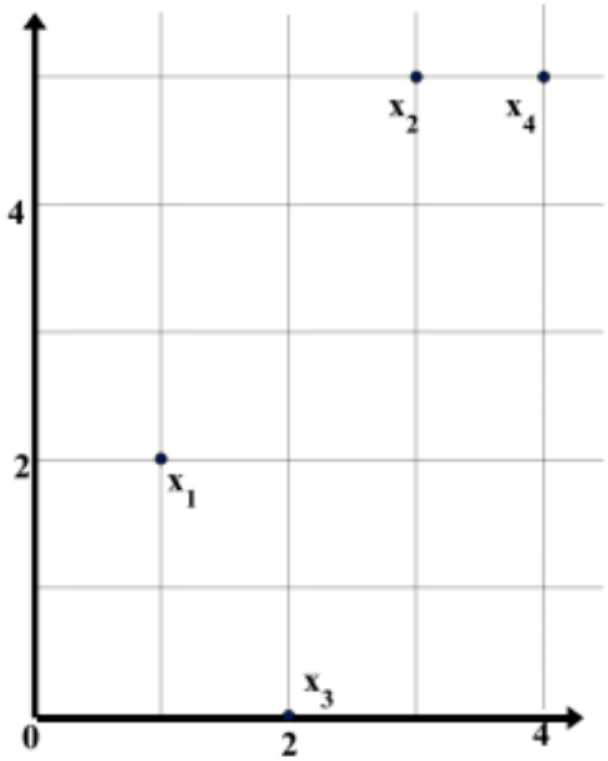
- Minkowski distance $d(\mathbf{x}_i, \mathbf{x}_j) = (|x_i^1 - x_j^1|^h + ... + |x_i^M - x_j^M|^h)^{1/h}$
- $\mathbf{x}_i, \mathbf{x}_j$ are two data points
- h is the order, we called Minkovski with order h as L-h norm.
- Properties
  - $d(\mathbf{x}_i, \mathbf{x}_j) \geq 0$ and $d(\mathbf{x}_i, \mathbf{x}_i) = 0$ (positive definiteness)
  - $d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i)$ (Symmetry)
  - $d(\mathbf{x}_i, \mathbf{x}_k) \leq d(\mathbf{x}_i, \mathbf{x}_j) + d(\mathbf{x}_j, \mathbf{x}_k)$ (Triangle Inequality)
  - A distance that satisfies these properties is called a metric

Special cases of Minkowski Distance

- h=1 : Manhattan (city block, L1-norm) distance
  - $d(\mathbf{x}_i, \mathbf{x}_j) = (|x_i^1 - x_j^1| + ... + |x_i^M - x_j^M|)$
- h=2 : (L2 norm) Euclidean distance
  - $d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(|x_i^1 - x_j^1|^2 + ... + |x_i^M - x_j^M|^2)}$
- h = ∞: supremum (Lmax norm, L_∞ norm)
  - This is the maximum difference between any component (attribute) of the vectors
  - $d(\mathbf{x}_i, \mathbf{x}_j) = \max_m |x_i^m - x_j^m|$

Example: Minkowski distance

point	Feature 1	Feature 2
X1	1	2
X2	3	5
X3	2	0
x4	4	5



L1	X1	x2	x3	X4
X1	0			
X2	5	0		
X3	3	6	0	
X4	6	1	7	0

L2	x1	x2	x3	X4
X1	0			
X2	3.61	0		
X3	2.24	5.1	0	
x4	4.24	1	5.39	0

L_inf	x1	x2	x3	X4
X1	0			
X2	3	0		
X3	2	5	0	
x4	3	1	5	0

Dissimilarity Measure for Categorical Features

- Categorical feature can take 2 or more states, e.g.,red, yellow, blue, green (generalisation of a binary feature)
- Method 1: simple matching
  - $d(\mathbf{x}_i, \mathbf{x}_j) = \frac{M-P}{M}$
  - M = number of features, P = number of matches
- Method 2: Use encoded binary features
  - creating a new binary attribute for each of the K states
  - red= 110, yellow = 010, blue = 101, green = 001
  - then measuring using method for binary attribute

Dissimilarity Measure for Ordinal Features

- Method 1: Simply use the rank as numeric feature $(1 \dots K)$
- Method 2:
  - Use Normalised Rank Transform to represent ordering information
  - Compute ranks r ($r = 1$ to K)
  - Treat Z as interval-scaled $[0, 1]$ $Z = \frac{r-1}{K-1}$
  - Z is then of type numeric

Dissimilarity Measure for textual data

- A document can be represented by thousands of attributes, each recording the frequency of a particular word (or keywords) in the document.

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

Cosine similarity

- Cosine measure:
  - $cos(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i^T \mathbf{x}_j}{||\mathbf{x}_i|| \cdot ||\mathbf{x}_j||}$
  - where numerator is the dot product and $||\mathbf{x}_i||$ is the length of vector $\mathbf{x}_i$.
- Also useful for vector objects such as gene features in micro-arrays data.

Example: Cosine Similarity

- Find the similarity (angle) between documents 1 and 2.
  - $d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$
  - $d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$
  - $cos(d_1, d_2) = ?$
- CosineDistance = 1 - CosineSimilarity
- The above is not proper distance metric (it does not satisfy triangle inequality)
- Proof ?

Summary

- Data point and Data matrix
- Data feature types: categorical, binary, ordinal, numeric
- Gain insight into the data by:
  - Basic statistical data description: central tendency, dispersion, graphical displays
  - Data visualisation: map data onto graphical primitives
  - Measure data similarity
- Above steps are the beginning of data preprocessing
- Many methods have been developed but still an active area of research

References

- U. Fayyad, G. Grinstein, and A. Wierse. Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed. , 2011
- S. Santini and R. Jain, Similarity measures, IEEE Trans. on Pattern Analysis and Machine Intelligence, 21(9), 1999