Getting to Know Your Data

Jakramate Bootkrajang [1]

August 18, 2016

[1] Based on material by Jiawei Han and Micheline Kamber

Outlines

- **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

	colour	height	weight	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 vector (point) in Euclidean space. $x_i = \{x_i^1, \ldots, x_i^M\}$ where M is dimensionality of the data
- So, a dataset of N data points is a $N \times M$ 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 equally important
    - ⋆ e.g., gender
  - Asymmetric binary:outcomes not equally important.
    - ⋆ e.g., medical test (positive vs. negative)

Feature Types

- 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)

Outlines

- 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 variation, quantile.

Measuring the Central Tendency

- Mean
  - Arithmetic mean: $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
  - Weighted arithmetic mean $\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}$
  - Trimmed mean: chopping extreme values
- 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

- Mode
  - Value that occurs most frequently in the data
  - A dataset can be unimodal (1 mode), bimodal (2 modes), etc.

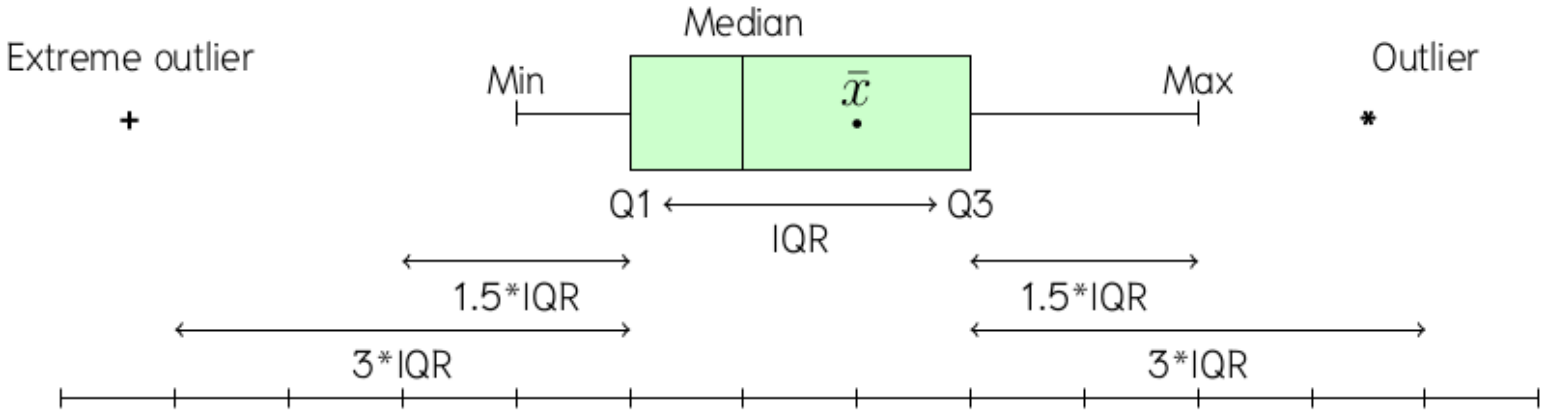
Measuring dispersion of data

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

Measuring dispersion of data

- Inter-quartile range: $\text{IQR} = \text{Q3} - \text{Q1}$
- 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)

Outlines

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

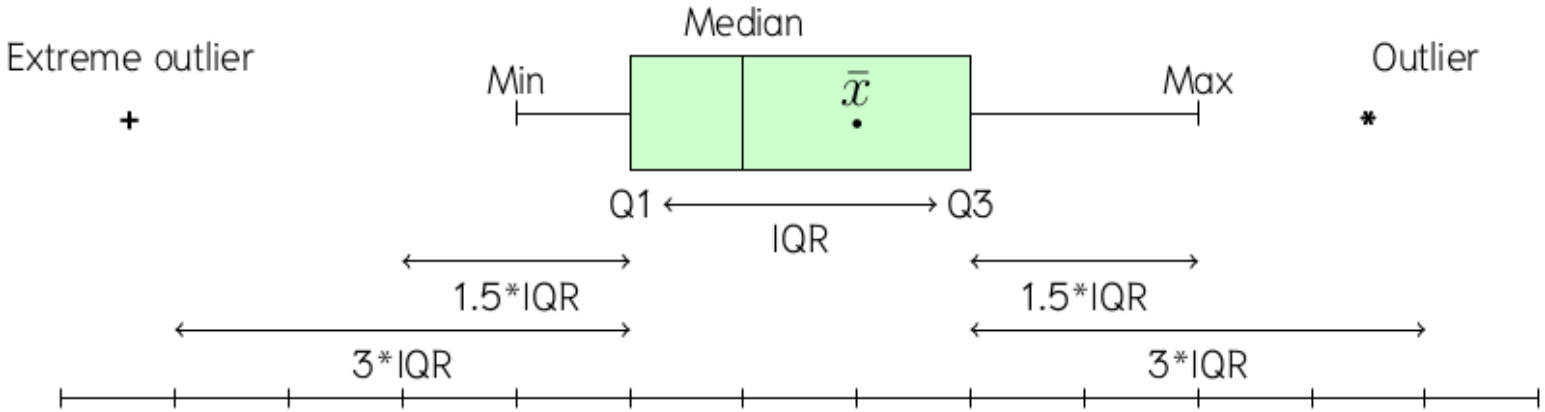
Data Visualisation

- 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

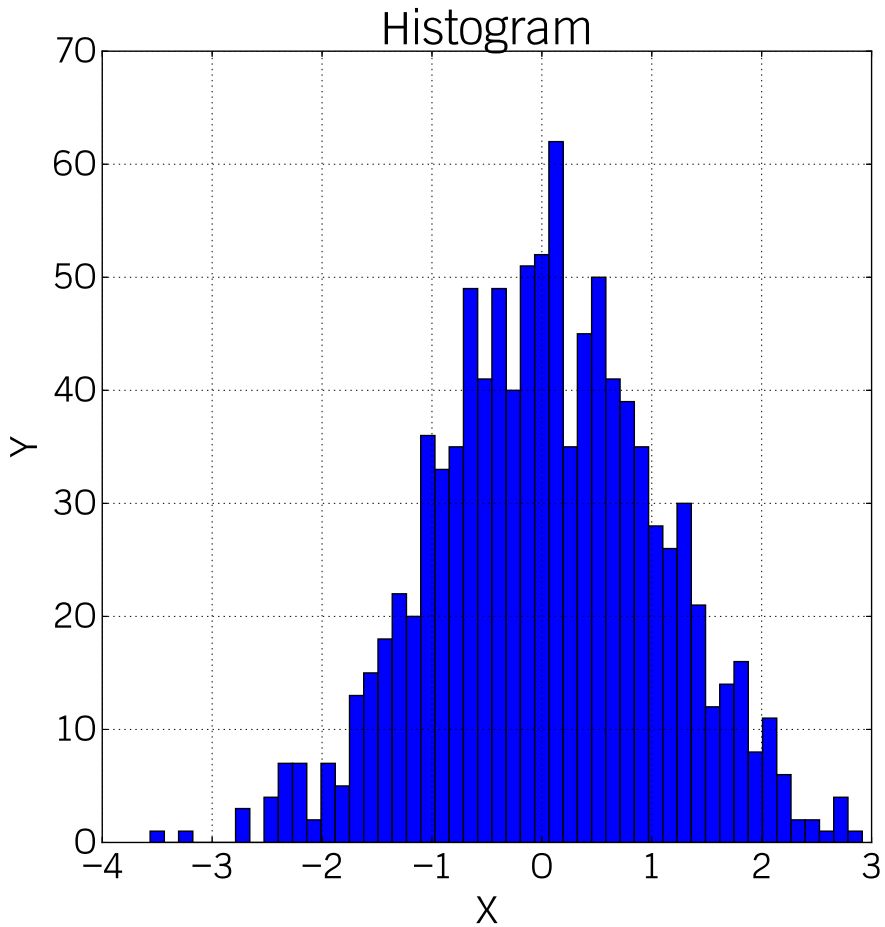


รูปภาพ 2.1: แผนภาพแบบกล่องและตัวเลขทางสถิติ 5 ตัว รวมถึงค่าผิดปกติต่างๆ (อัตราส่วนอาจไม่ตรง)

Histogram

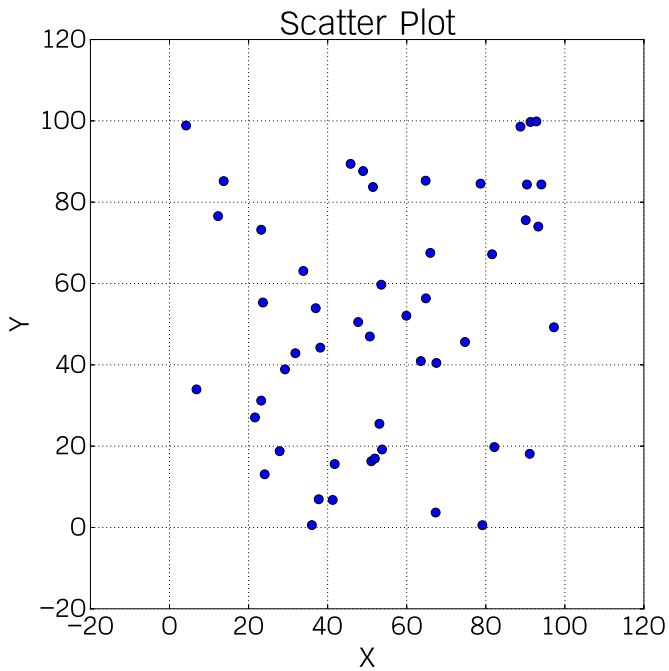
- Histogram: Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of bars
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent

Histogram

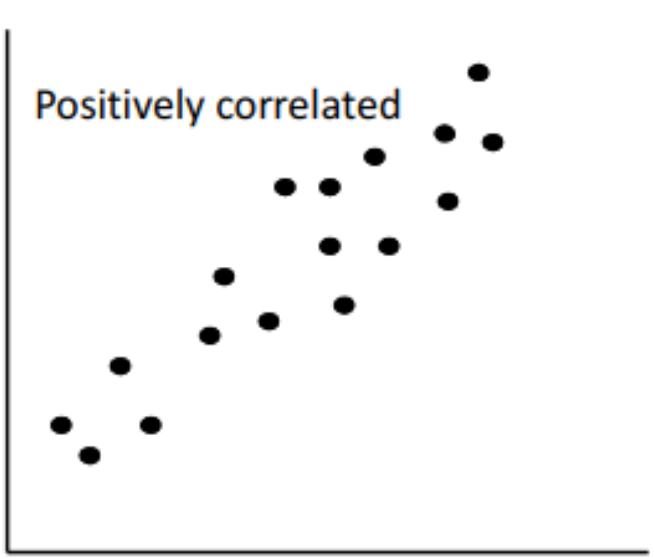


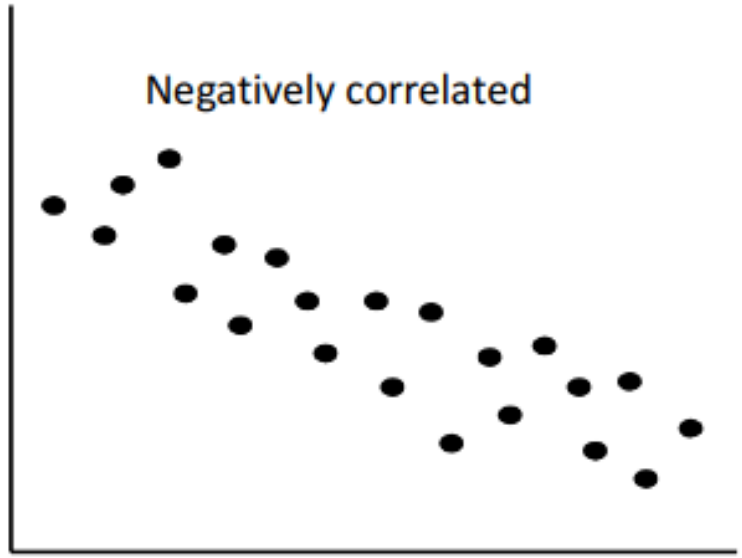
Scatter plot

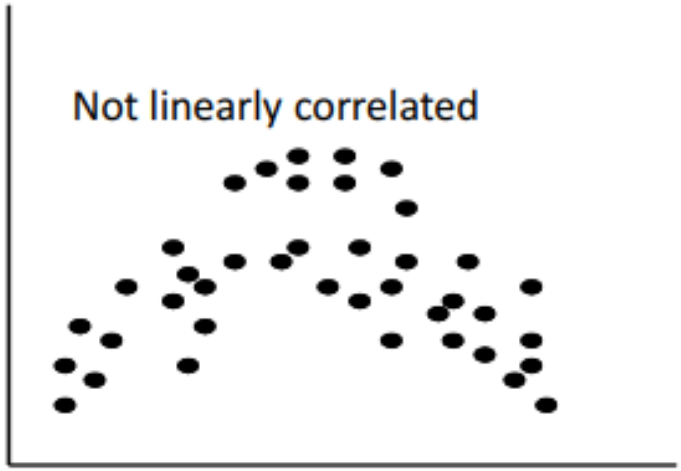
- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



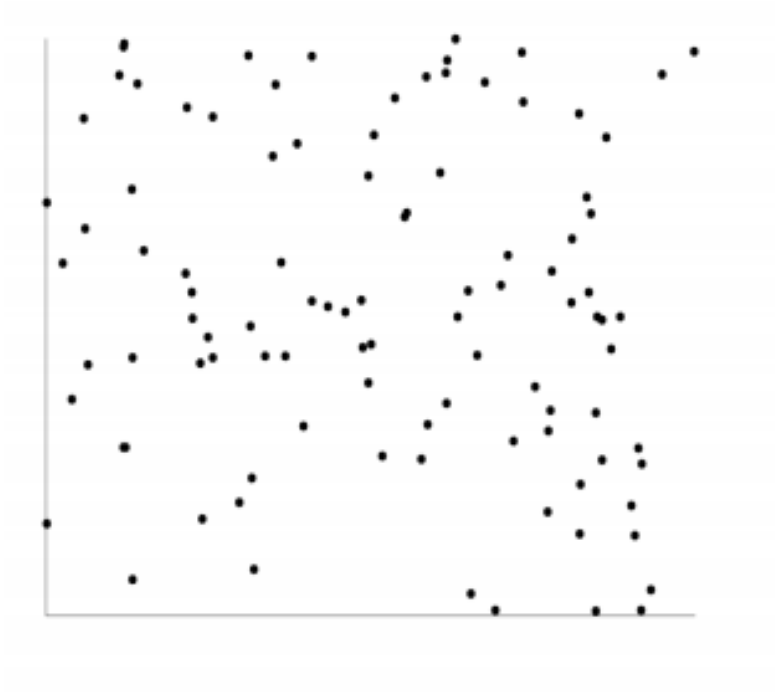
Scatter plot examples





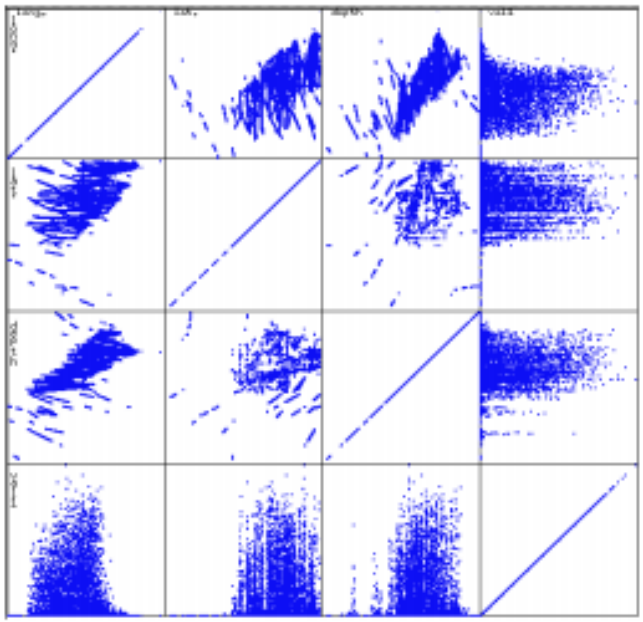


Scatter plot examples (uncorrelated data)



Scatter plot for high dimensional 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



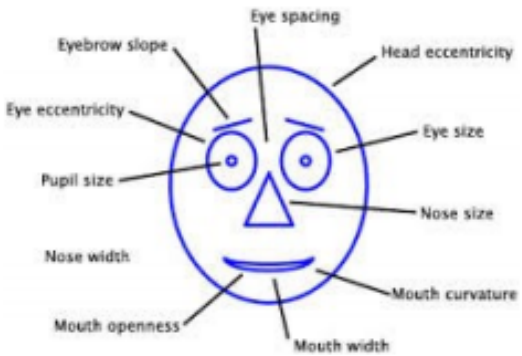
Icon-based Visualisation Techniques

- 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

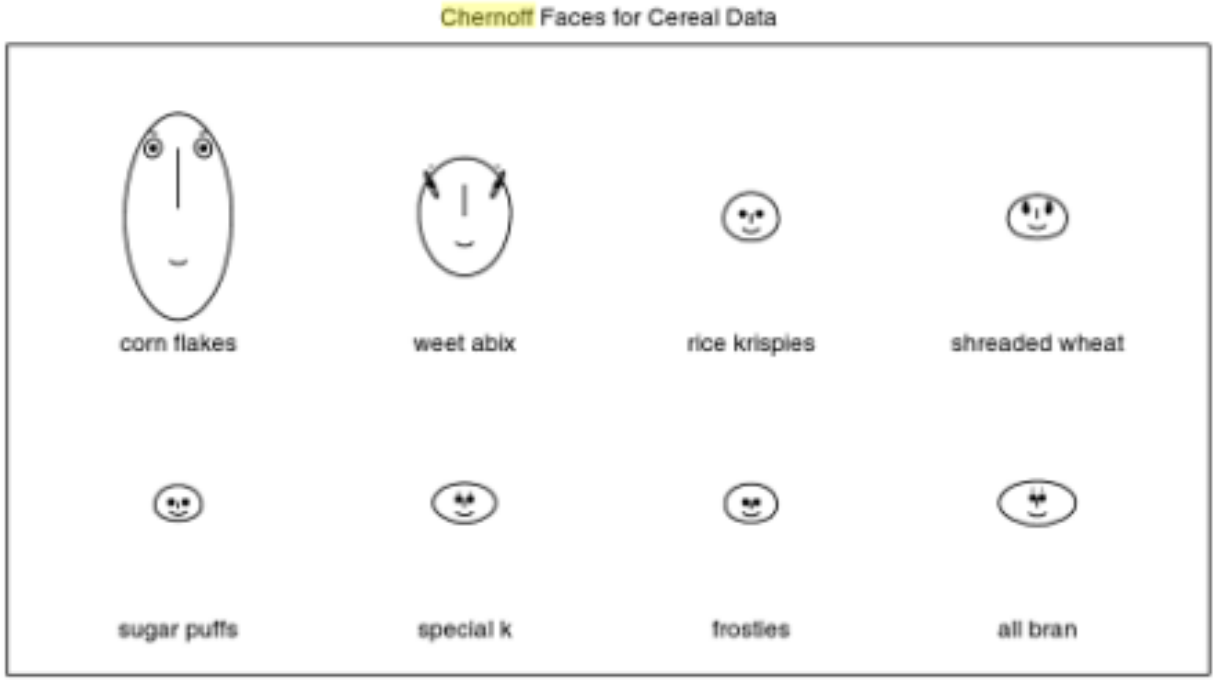


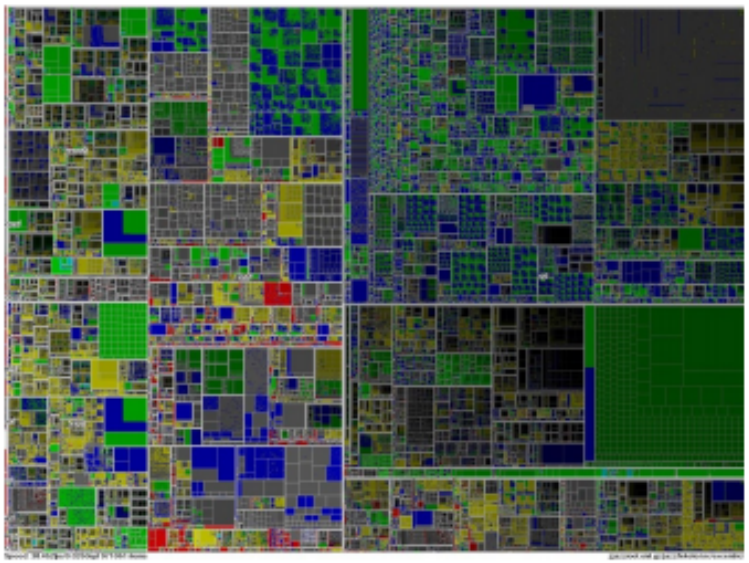
FIGURE 10.1
This shows the Chernoff faces for the **`cereal`** data, where we have 8 observations and 11 variables. The shape and size of various facial features (head, eyes, brows, mouth, etc.) correspond to the values of the variables. The variables represent the percent agreement to statements about 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

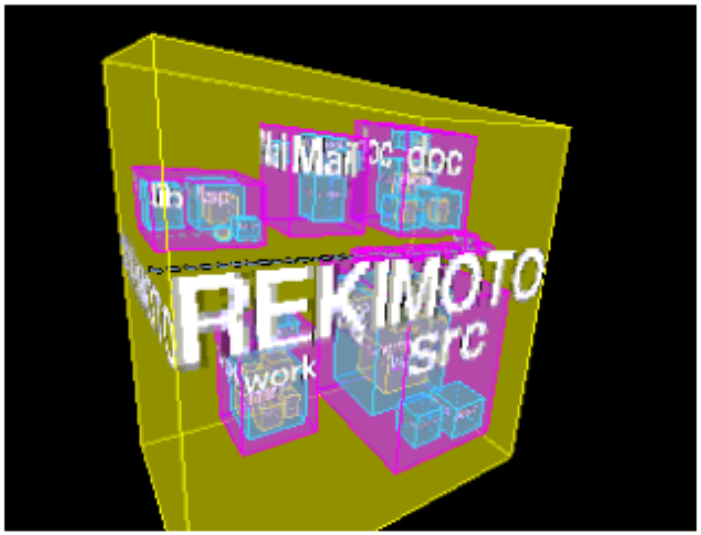
Tree-map

- 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



Outlines

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

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

Disimilarity matrix

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

$$\begin{bmatrix} 0 & & & & \\ d(x_2, x_1) & 0 & & & \\ d(x_3, x_1) & d(x_3, x_2) & 0 & & \\ \vdots & \vdots & \vdots & 0 & \\ d(x_N, x_1) & d(x_N, x_2) & \cdots & d(x_N, x_{N-1}) & 0 \end{bmatrix}$$

Dissimilarity Measure for Binary Features

- Can be summarised in a contigency table

	1	0
1	q	r
0	s	t

- Distance measure for symmetric binary variables:

$$d(x_i, x_j) = \frac{r+s}{q+r+s+t}$$

- Distance measure for asymmetric binary variables:

$$d(x_i, x_j) = \frac{r+s}{q+r+s}$$

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 a symmetric attribute
- The remaining attributes are asymmetric binary
- Let the values Y and P be 1, and the value N be 0
  - $d(\text{Jack,Mary}) = \frac{0+1}{2+0+1} = 0.33$
  - $d(\text{Jack,Jim}) = \frac{1+1}{1+1+1} = 0.67$
  - $d(\text{Jim,Mary}) = ???$

Distance on Numeric Data: Minkowski Distance

- Minkowski distance

$$d(x_i, x_j) = (|x_i^1 - x_j^1|^h + |x_i^2 - x_j^2|^h + \cdots + |x_i^M - x_j^M|^h)^{1/h}$$

where x_i and x_j are two M-dimensional data points,

- and h is the order. (the distance defined is called L-h norm)
- Properties
  - $d(x_i, x_j) \geq 0$, and $d(x_i, x_i) = 0$ (Positive definiteness)
  - $d(x_i, x_j) = d(x_j, x_i)$ (Symmetry)
  - $d(x_i, x_k) = d(x_i, x_j) + d(x_j, x_k)$ (Triangle Inequality)
  - A distance that satisfies these properties is called metric

Special Cases of Minkowski Distance

- h=1 : Manhattan (city block, $L1$-norm) distance

$$d(x_i, x_j) = (|x_i^1 - x_j^1| + |x_i^2 - x_j^2| + \cdots + |x_i^M - x_j^M|)$$

- h=2 : ($L2$ norm) Euclidean distance

$$d(x_i, x_j) = \sqrt{(|x_i^1 - x_j^1|^2 + |x_i^2 - x_j^2|^2 + \cdots + |x_i^M - x_j^M|^2)}$$

- h = ∞: supremum (Lmax norm, L_∞ norm)
  - This is the maximum difference between any component (attribute) of the vectors

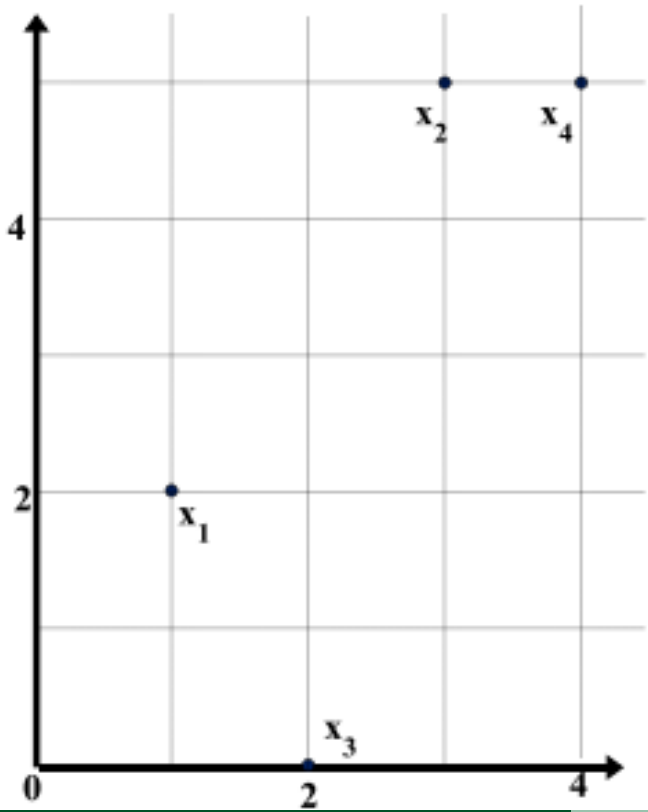
$$d(x_i, x_j) = \max_m |x_i^m - x_j^m|$$

Dissimilarity Measure for Categorical Features

- Categoricala feature can take 2 or more states, e.g., red, yellow, blue, green (generalisation of a binary feature)
- Method 1: Simple matching
  - $d(x_i, x_j) = \frac{M-P}{M}$
  - M = # of features , P = # 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

Example

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

Distance measure for Ordinal Features

- Method 1: Simply use the rank
- 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

Similarity for Textual data - Cosine Similarity

- 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 (cont.)

- Other possible vector objects: gene features in micro-arrays,
- Cosine measure:

$$cosine(x_i, x_j) = \frac{x_i^T x_j}{||x_i|| \cdot ||x_j||}$$

where numerator is the dot product and $||x_i||$ is the length of vector x_i.

Example: Cosine Similarity

- Find the similarity (angle) between documents 1 and 2.

$$d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$$

$$d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$$

$$d_1^T d_2 = ??,\ ||d_1|| = ??,\ ||d_2|| = ??$$

$$cosine(d_1, d_2) = ??$$

- 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

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