CEU
Universidad San Pablo

Clustering (Data mining)

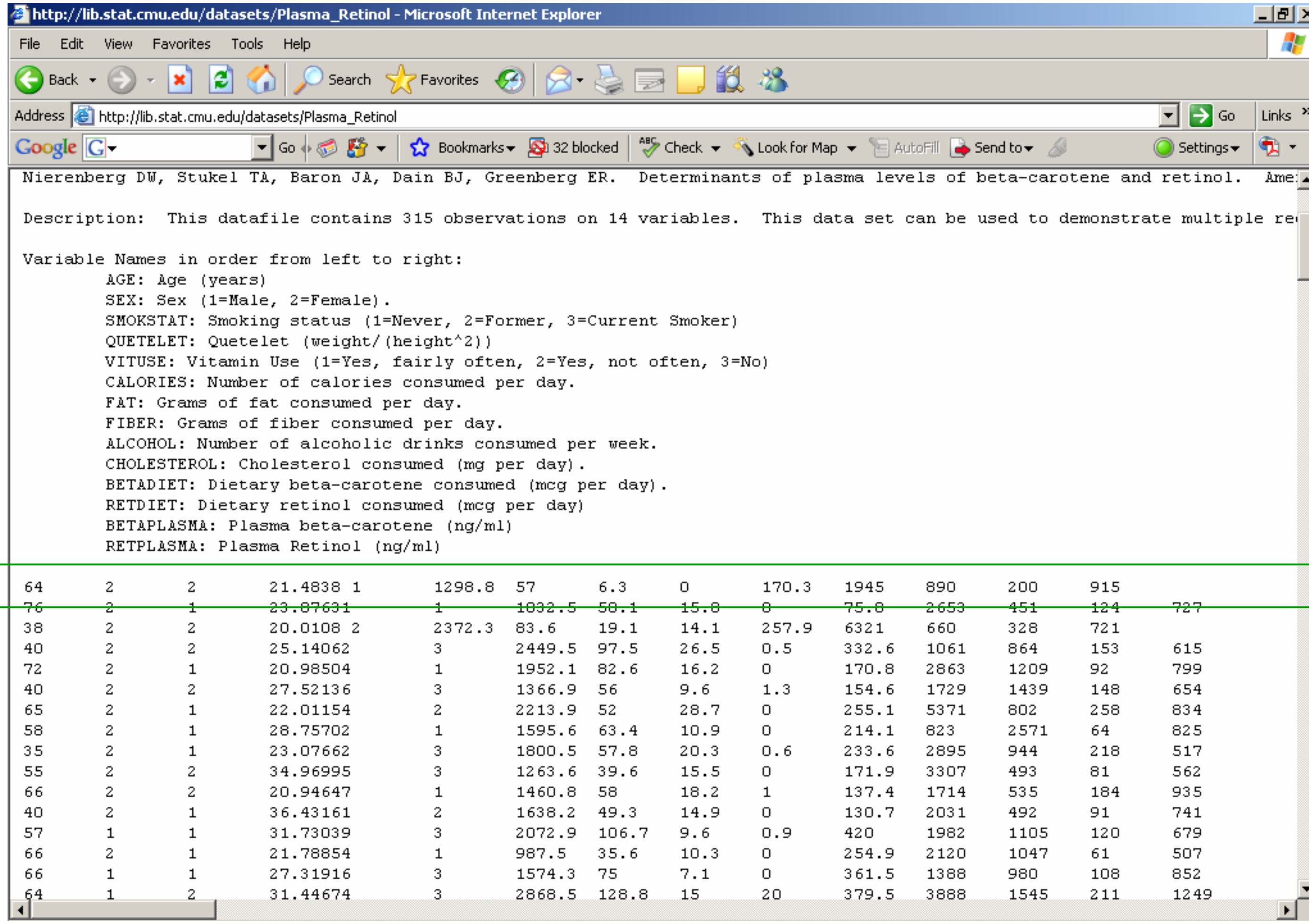
Session 1: Introduction

Carlos Óscar Sánchez Sorzano, Ph.D.
Madrid, July 23rd 2007

Course outline: Session 1

1. Introduction
    1.1 Problem formulation
    1.2 Types of features
    1.3 Feature extraction
    1.4 Graphical examination
    1.5 Data quality
    1.6 Distance measures
    1.7 Preprocessing
    1.8 Data reduction
    1.9 Types of clustering: partitional, hierarchical, probabilistic

1.1 Problem formulation



Nierenberg DW, Stukel TA, Baron JA, Dain BJ, Greenberg ER. Determinants of plasma levels of beta-carotene and retinol. Ame

Description: This datafile contains 315 observations on 14 variables. This data set can be used to demonstrate multiple re

Variable Names in order from left to right:

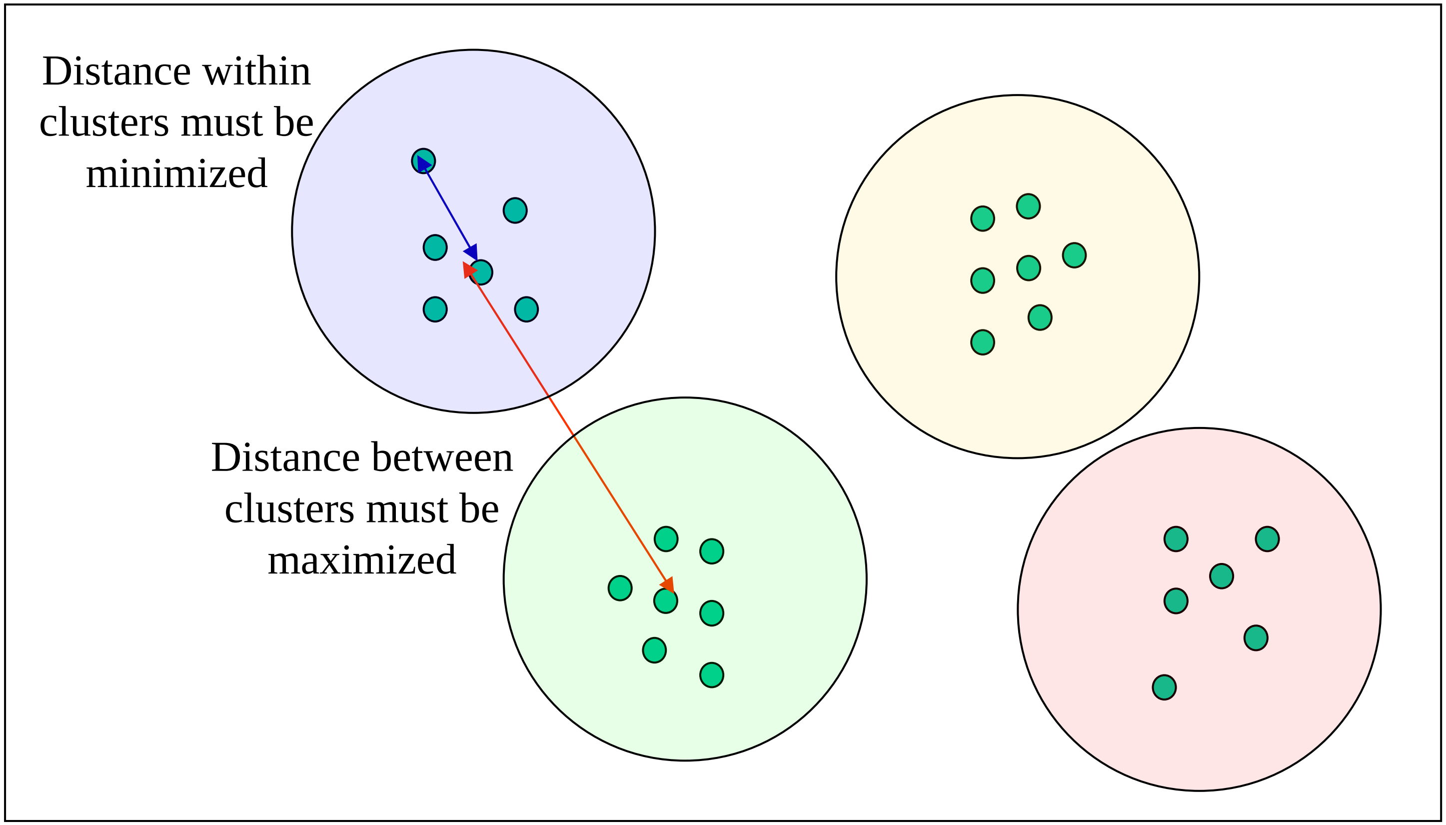
- AGE: Age (years)
- SEX: Sex (1=Male, 2=Female).
- SMOKSTAT: Smoking status (1=Never, 2=Former, 3=Current Smoker)
- QUETELET: Quetelet (weight/(height^2))
- VITUSE: Vitamin Use (1=Yes, fairly often, 2=Yes, not often, 3=No)
- CALORIES: Number of calories consumed per day.
- FAT: Grams of fat consumed per day.
- FIBER: Grams of fiber consumed per day.
- ALCOHOL: Number of alcoholic drinks consumed per week.
- CHOLESTEROL: Cholesterol consumed (mg per day).
- BETADIET: Dietary beta-carotene consumed (mcg per day).
- RETDIET: Dietary retinol consumed (mcg per day)
- BETAPLASMA: Plasma beta-carotene (ng/ml)
- RETPLASMA: Plasma Retinol (ng/ml)

$\mathbf{x}^t$

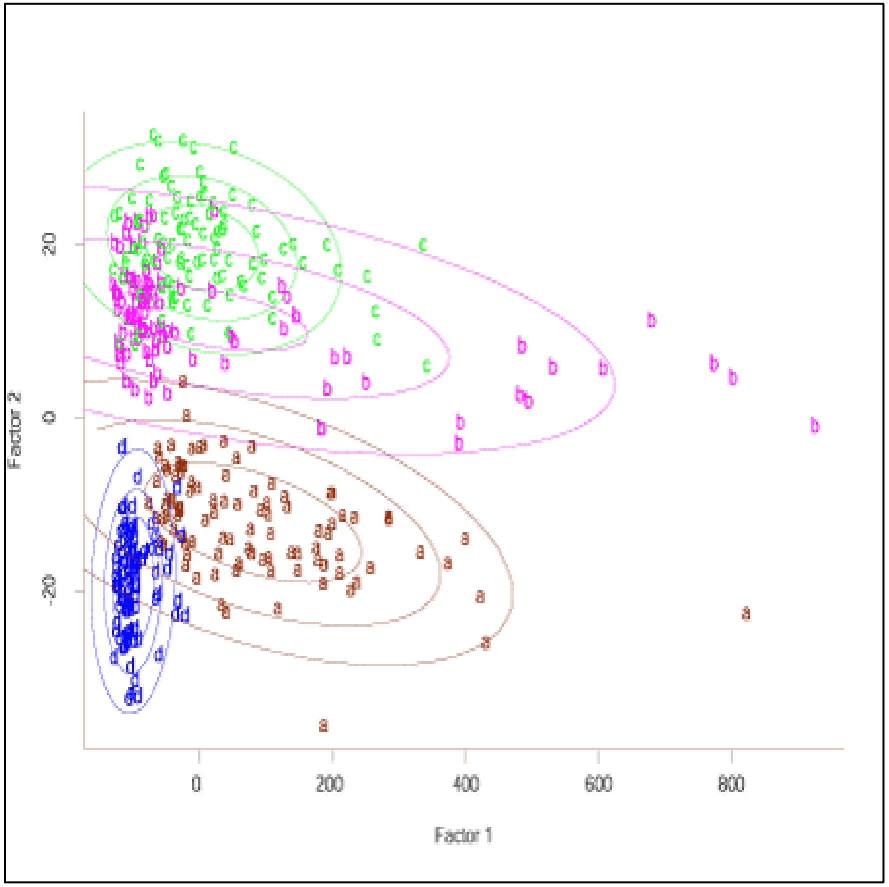
64	2	2	21.4838 1	1298.8	57	6.3	0	170.3	1945	890	200	915	
76	2	1	23.87631	1	1032.5	50.1	15.8	0	75.8	2653	451	124	727
38	2	2	20.0108 2	2372.3	83.6	19.1	14.1	257.9	6321	660	328	721	
40	2	2	25.14062	3	2449.5	97.5	26.5	0.5	332.6	1061	864	153	615
72	2	1	20.98504	1	1952.1	82.6	16.2	0	170.8	2863	1209	92	799
40	2	2	27.52136	3	1366.9	56	9.6	1.3	154.6	1729	1439	148	654
65	2	1	22.01154	2	2213.9	52	28.7	0	255.1	5371	802	258	834
58	2	1	28.75702	1	1595.6	63.4	10.9	0	214.1	823	2571	64	825
35	2	1	23.07662	3	1800.5	57.8	20.3	0.6	233.6	2895	944	218	517
55	2	2	34.96995	3	1263.6	39.6	15.5	0	171.9	3307	493	81	562
66	2	2	20.94647	1	1460.8	58	18.2	1	137.4	1714	535	184	935
40	2	1	36.43161	2	1638.2	49.3	14.9	0	130.7	2031	492	91	741
57	1	1	31.73039	3	2072.9	106.7	9.6	0.9	420	1982	1105	120	679
66	2	1	21.78854	1	987.5	35.6	10.3	0	254.9	2120	1047	61	507
66	1	1	27.31916	3	1574.3	75	7.1	0	361.5	1388	980	108	852
64	1	2	31.44674	3	2868.5	128.8	15	20	379.5	3888	1545	211	1249

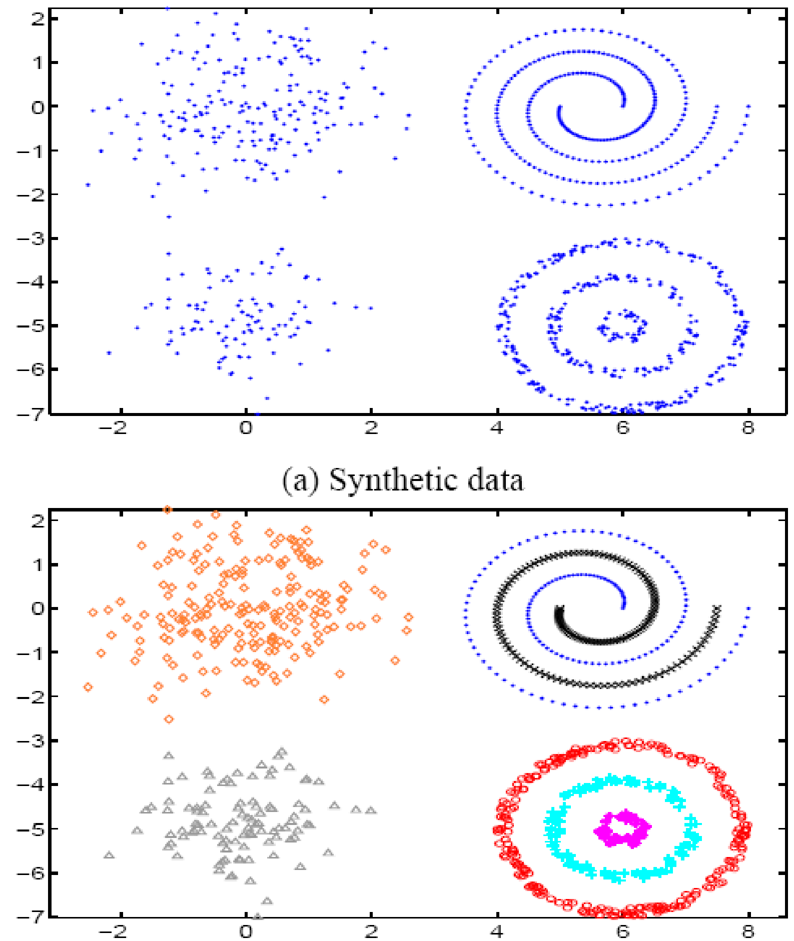
1.1 Problem formulation

Find groups of points that are close to each other within the cluster and far from the rest of clusters



1.1 Problem formulation





(a) Synthetic data

(b) Clusters discovered by multiobjective clustering

1.1 Problem formulation

Application> Marketing segmentation

- Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
- Approach:

Feature extraction

  – Collect different attributes of customers based on their geographical and lifestyle related information.

Distance definition

  – Define an apropriate distance measure between a pair of customers.

Cluster algorithm

  – Find clusters of similar customers.

Cluster validation

  – Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

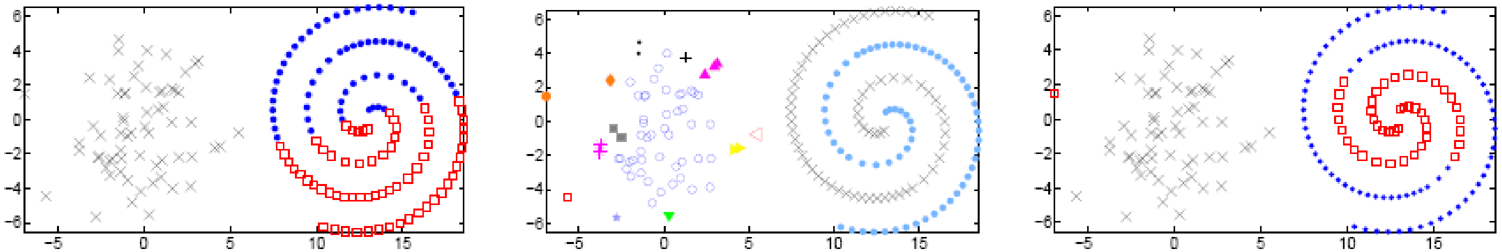
1.1 Problem formulation

Application> Document clustering

- Goal:find groups of documents that are similar to each other based on the important terms appearing in them.
- Approach:

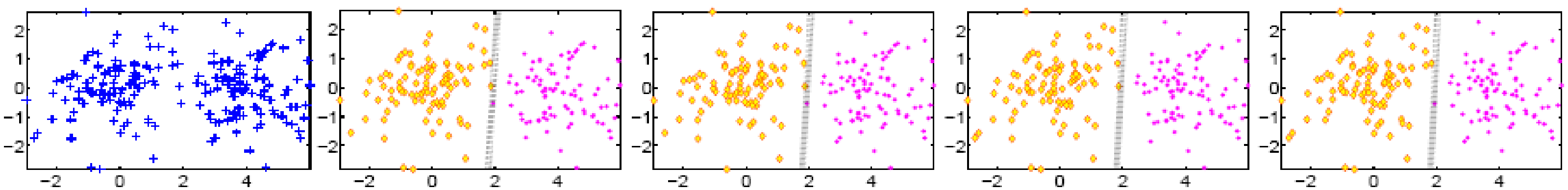
Feature extraction	– Identify frequently occurring terms in each document.
Distance definition	– Form a similarity measure based on the frequencies of different terms.
Cluster algorithm	– Use it to cluster.
Cluster validation	– Do newly arrived documents fit in the clusters?

1.1 Problem formulation

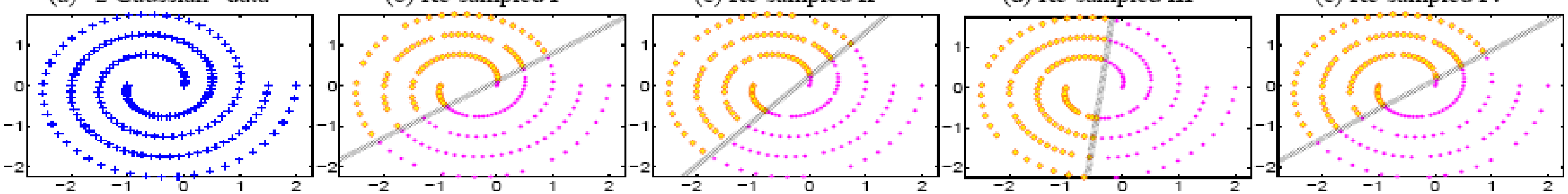


(a) k-means with $k = 3$ (b) Single-link, $k = 15$ (c) Spectral with $\sigma = 0.6$ and $k = 3$

Figure 1: The resulting partitions by (a) k-means, (b) single-link and (c) spectral clustering on this "globular-spiral" data set.



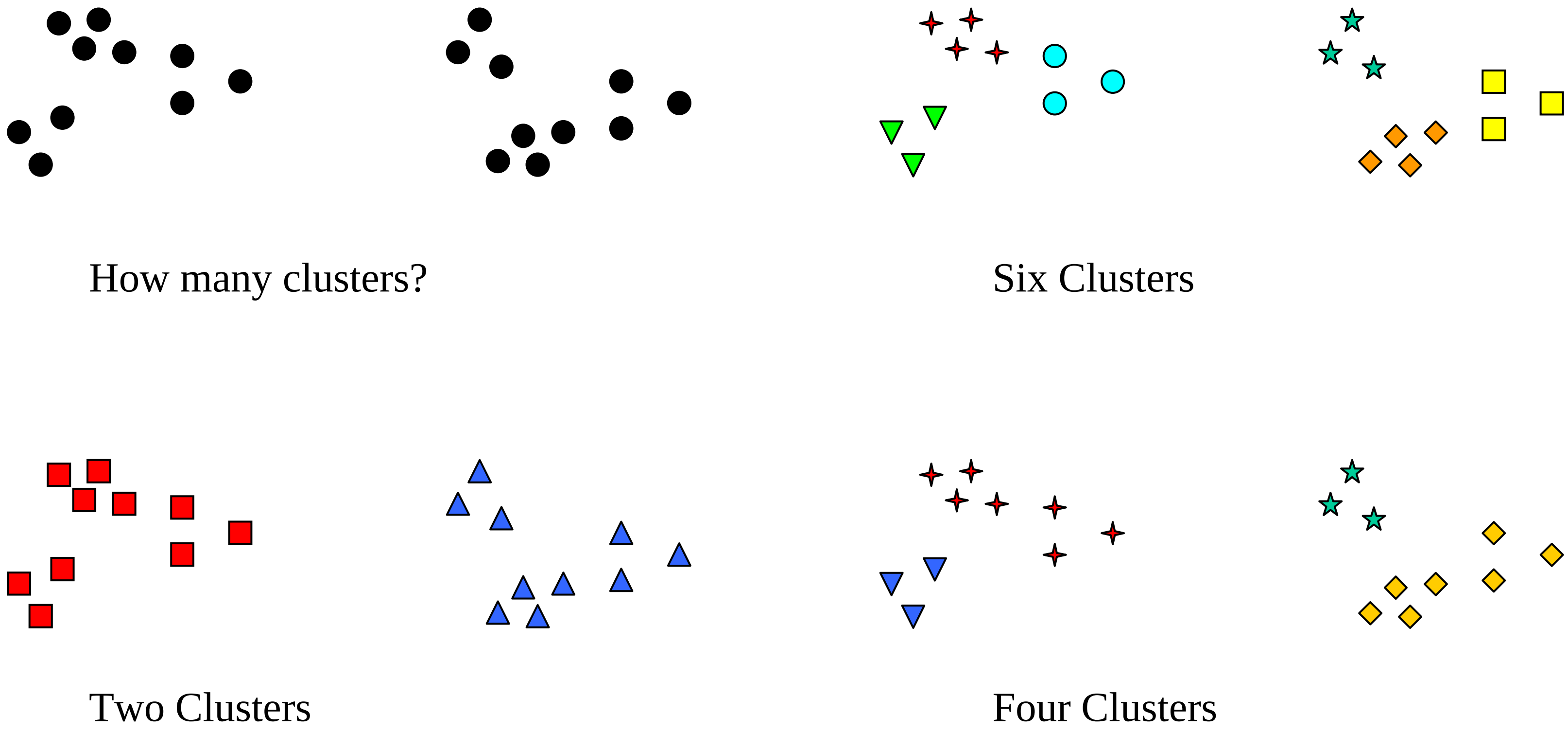
(a) "2 Gaussian" data (b) Re-sampled I (c) Re-sampled II (d) Re-sampled III (e) Re-sampled IV



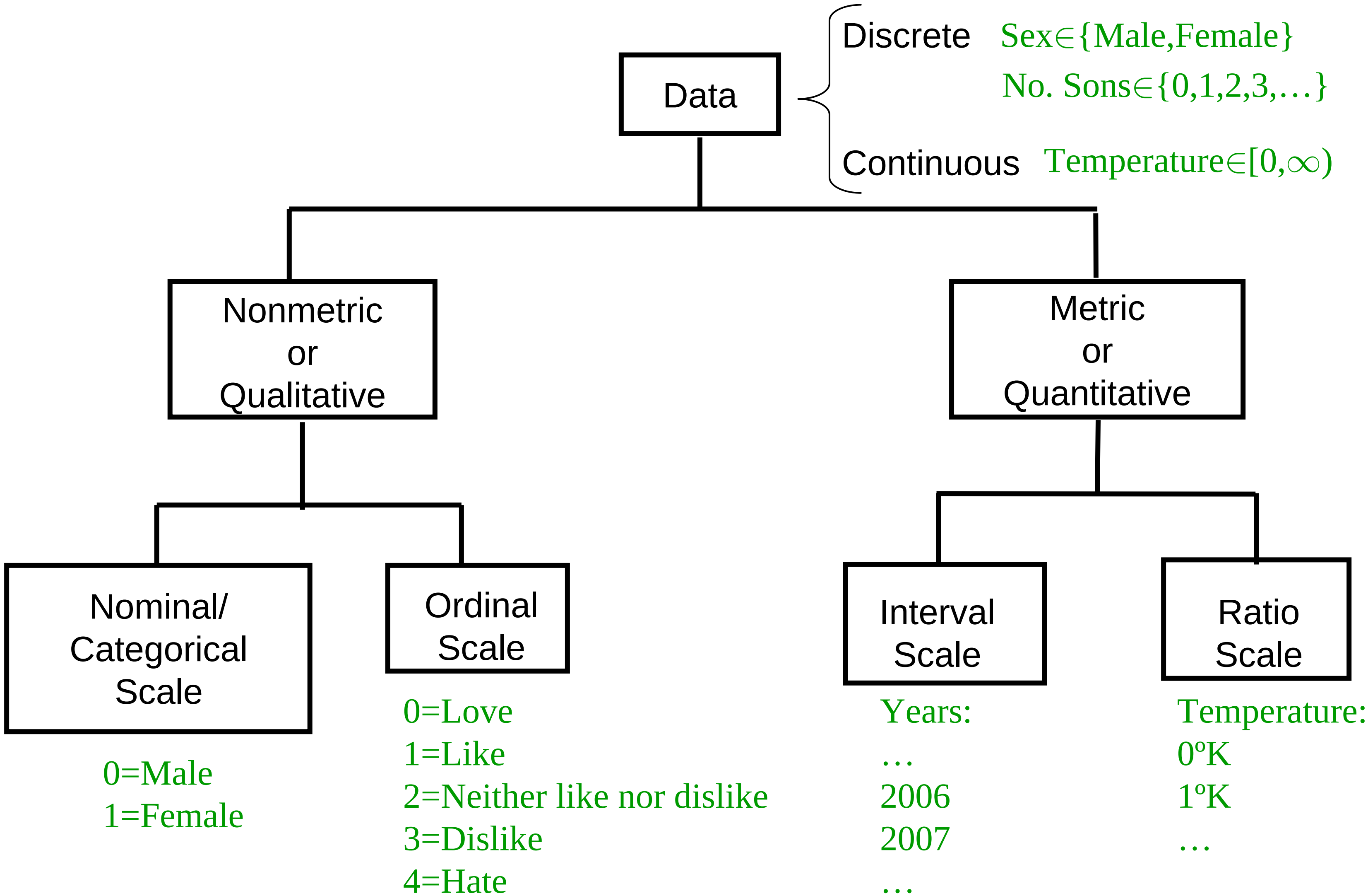
(f) "2 spiral" data (g) Re-sampled I (h) Re-sampled II (i) Re-sampled III (j) Re-sampled IV

Figure 2: Results of k-means with $k = 2$ for different re-sampled versions of two data sets. Dotted lines in the figures correspond to the cluster boundaries. The partitions of "2 Gaussian" data set are almost the same for different re-sampled versions, suggesting that k-means with $k = 2$ gives good clusters. The same cannot be said for the "2 spiral" data set.

1.1 Problem formulation



1.2 Types of features



1.2 Types of features

Coding of categorical variables

Hair Colour
{Brown, Blond, Black, Red}

$$\xrightarrow{\text{No order}} \quad \left(x_{Brown}, x_{Blond}, x_{Black}, x_{Red}\right) \in \{0,1\}^4$$

Peter: Black	Peter: $\{0,0,1,0\}$
Molly: Blond	Molly: $\{0,1,0,0\}$
Charles: Brown	Charles: $\{1,0,0,0\}$

Company size
{Small, Medium, Big}

$$\xrightarrow{\text{Implicit order}} \quad x_{size} \in \{0,1,2\}$$

Company A: Big	Company A: 2
Company B: Small	Company B: 0
Company C: Medium	Company C: 1

1.3 Feature extraction

- Most sensitive part of the process. If the right information for clustering is not present, no clustering algorithm will work.
- Specific to each field (available from Session1/Docs):
  - Web navigation: Chen2002 and Lim2005
  - Video processing: Chang1995 and Zhong1996
  - Image processing: Szepesvari
  - Character recognition: Liu2005
  - Gait recognition: Dawson2002
  - …

RETDIET: Dietary retinol consumed (mcg per day)
BETAPLASMA: Plasma beta-carotene (ng/ml)
RETPLASMA: Plasma Retinol (ng/ml)

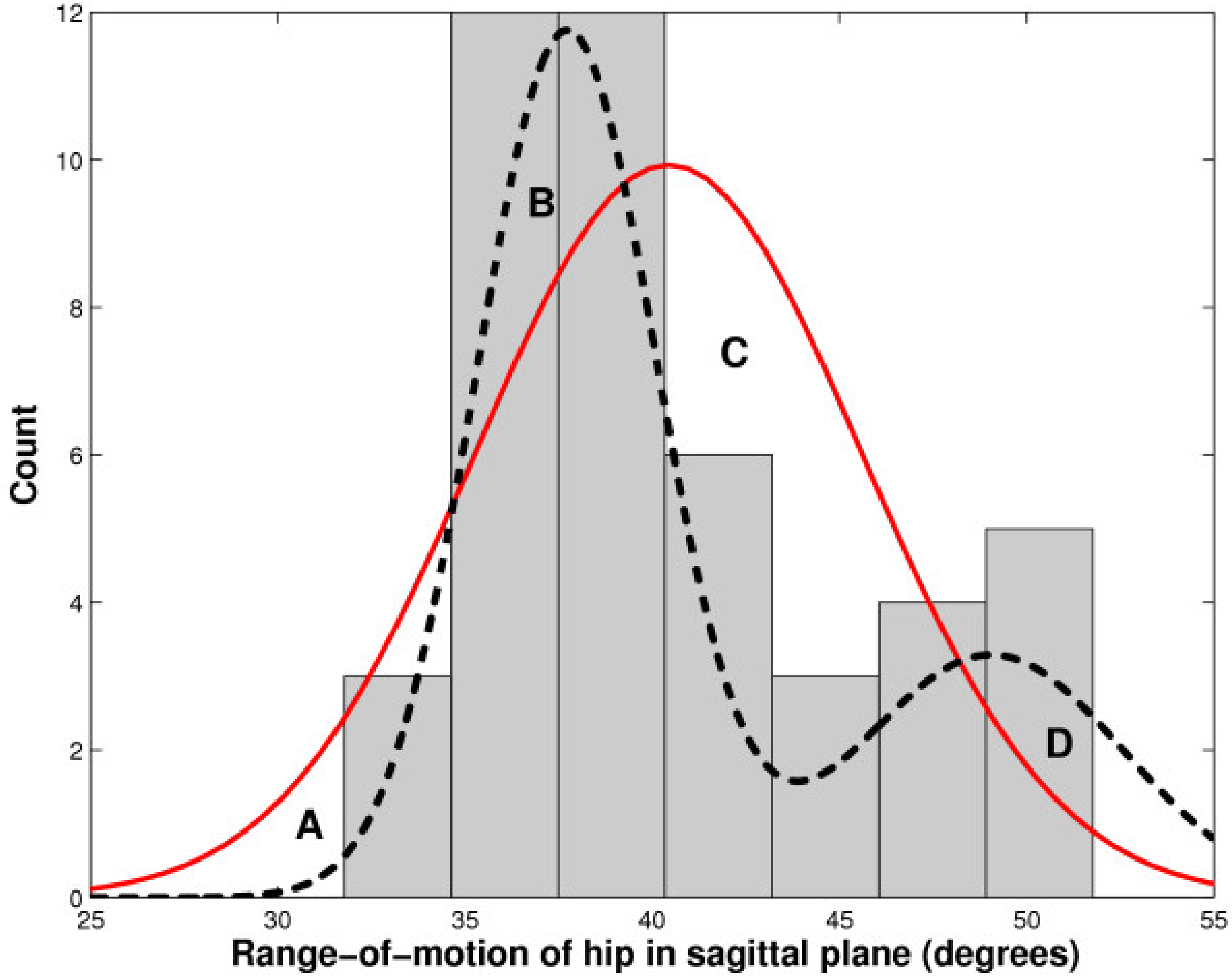
$\mathbf{x}^t$

64	2	2	21.4838 1	1298.8	57	6.3	0	170.3	1945	890	200	915	
76	2	1	23.87631	1	1032.5	50.1	15.8	0	75.8	2653	451	124	727
38	2	2	20.0108 2	2372.3	83.6	19.1	14.1	257.9	6321	660	328	721	
40	2	2	25.14062	3	2449.5	97.5	26.5	0.5	332.6	1061	864	153	615
72	2	1	20.98504	1	1952.1	82.6	16.2	0	170.8	2863	1209	92	799
40	2	2	27.52136	3	1366.9	56	9.6	1.3	154.6	1729	1439	148	654
65	2	1	22.01154	2	2213.9	52	28.7	0	255.1	5371	802	258	834
58	2	1	28.75702	1	1595.6	63.4	10.9	0	214.1	823	2571	64	825
35	2	1	23.07662	3	1800.5	57.8	20.3	0.6	233.6	2895	944	218	517
55	2	2	34.96995	3	1263.6	39.6	15.5	0	171.9	3307	493	81	562
66	2	2	20.94647	1	1460.8	58	18.2	1	137.4	1714	535	184	935
40	2	1	36.43161	2	1638.2	49.3	14.9	0	130.7	2031	492	91	741

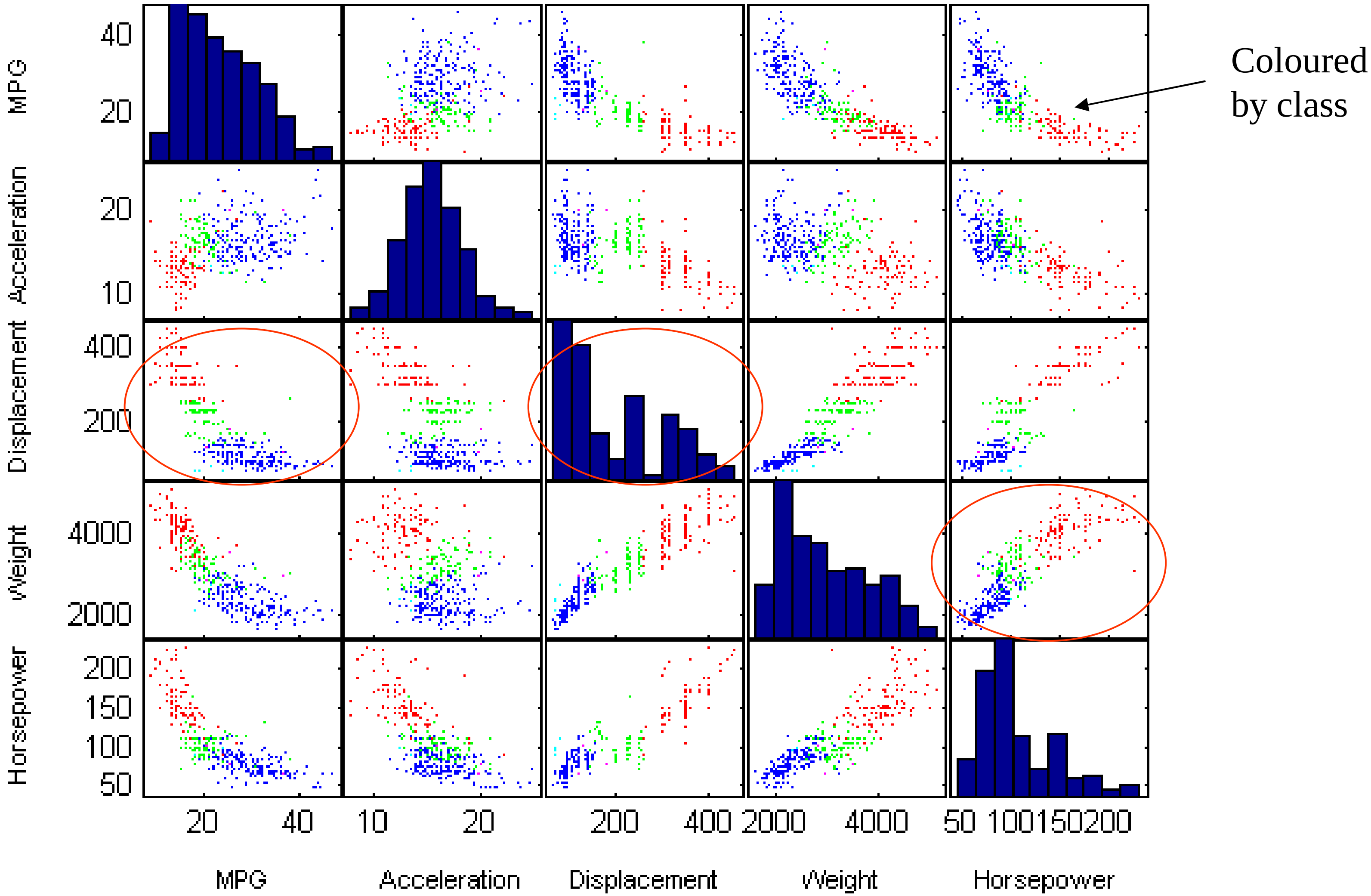
1.4 Graphical examination

- Univariate distribution plots
- (Bivariate distribution plots)
- Pairwise plots
  - Scatter plots
  - Boxplots
- (Multivariate plots)
  - (Chernoff faces)
  - (Star plots)

1.4 Graphical examination: Univariate distribution



1.4 Graphical examination: Scatter plots



1.5 Data quality: Missing data

Types of missing data:

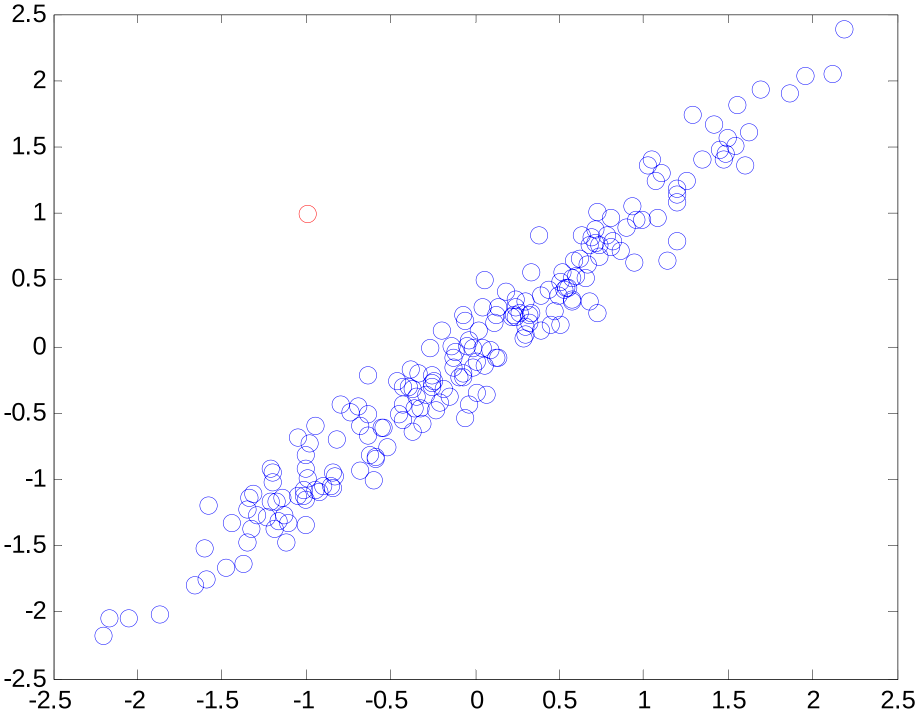
- Missing Completely At Random (MCAR)
- Missing at Random (MAR)

Strategies for handling missing data:

- use observations with complete data only
- delete case(s) and/or variable(s)
- estimate missing values (imputation):

+ All-available
+ Mean substitution
+ Cold/Hot deck
+ Regression (preferred for MCAR): Linear, Tree
+ Expectation-Maximization (preferred for MAR)
+ Multiple imputation (Markov Chain Monte Carlo, Bayesian)

1.5 Data quality: Outliers



Univariate detection

$$\frac{|x_i - median(x)|}{MAD(x)} > 4.5$$

$$MAD(x) = median\left(\left|x - median(x)\right|\right)$$

Multivariate detection

$$d^2(\mathbf{x}_i, \overline{\mathbf{x}}) = (\mathbf{x}_i - \overline{\mathbf{x}})^t S^{-1} (\mathbf{x}_i - \overline{\mathbf{x}}) > p + 3\sqrt{2p}$$

Covariance matrix

Number of variables

1.5 Data Quality: Duplicate data

- Data set may include data objects that are duplicates, or almost duplicates of one another
- This is a major issue when merging data from heterogeneous sources

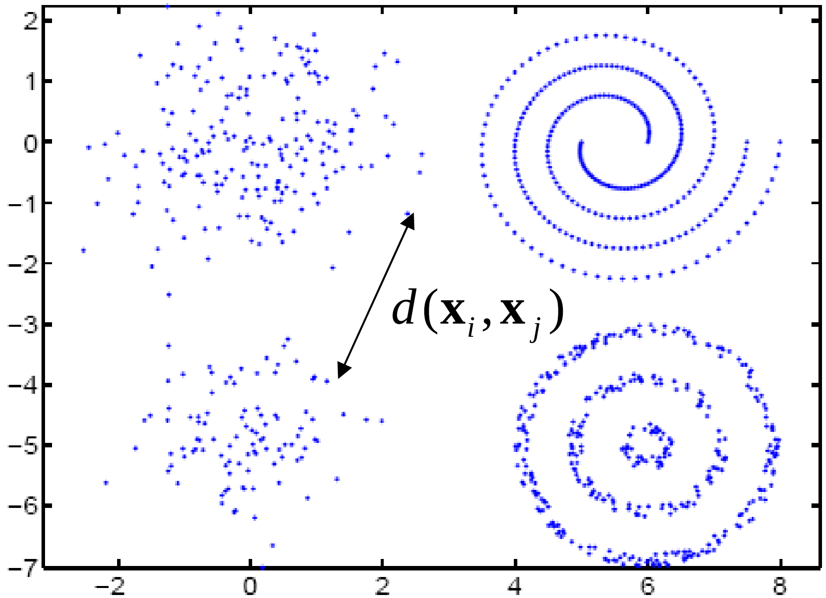
Example:

- Same person with multiple email addresses

Data cleansing:

- Remove duplicates (by partial distance, by classification)

1.6 Distance measures: Generic



$$d(\mathbf{x}_i, \mathbf{x}_j)$$

1-norm (Manhattan) $$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{s=1}^{n} \left| x_{is} - x_{js} \right|$$

Most used → p-norm (Euclidean p=2) Minkowski $$d(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{s=1}^{n} \left(x_{is} - x_{js} \right)^p \right)^{\frac{1}{p}}$$

Infinity (Chebyshev) norm $$d(\mathbf{x}_i, \mathbf{x}_j) = \max_s \left| x_{is} - x_{js} \right|$$

1.6 Distance measures: Generic

	Height (m)	Weight (kg)
Juan	1.80	80
John	1.70	72
Jean	1.65	81

$d(\mathbf{x}_i, \mathbf{x}_j)$	Juan	John	Jean
Juan	-----	8.0004	1.0112

	Height (cm)	Weight (kg)
Juan	180	80
John	170	72
Jean	165	81

$d(\mathbf{x}_i, \mathbf{x}_j)$	Juan	John	Jean
Juan	-----	11.3137	15.0333

Matrix-based distance $d(\mathbf{x}_i, \mathbf{x}_j) = \left(\mathbf{x}_i - \mathbf{x}_j\right)^t M^{-1} \left(\mathbf{x}_i - \mathbf{x}_j\right)$

Euclidean distance $d(\mathbf{x}_i, \mathbf{x}_j) = \left(\mathbf{x}_i - \mathbf{x}_j\right)^t I^{-1} \left(\mathbf{x}_i - \mathbf{x}_j\right) = \left(\mathbf{x}_i - \mathbf{x}_j\right)^t \left(\mathbf{x}_i - \mathbf{x}_j\right)$

Mahalanobis distance $d(\mathbf{x}_i, \mathbf{x}_j) = \left(\mathbf{x}_i - \mathbf{x}_j\right)^t \Sigma^{-1} \left(\mathbf{x}_i - \mathbf{x}_j\right)$

1.6 Distance measures: Generic

Mahalanobis distance $d(\mathbf{x}_i,\mathbf{x}_j)=\left(\mathbf{x}_i-\mathbf{x}_j\right)^t\Sigma^{-1}\left(\mathbf{x}_i-\mathbf{x}_j\right)$

$$\Sigma=\begin{pmatrix}\sigma^2_{height} & r\sigma_{height}\sigma_{weight}\\ r\sigma_{height}\sigma_{weight} & \sigma^2_{weight}\end{pmatrix}=\begin{pmatrix}100 & 70\\ 70 & 100\end{pmatrix}$$

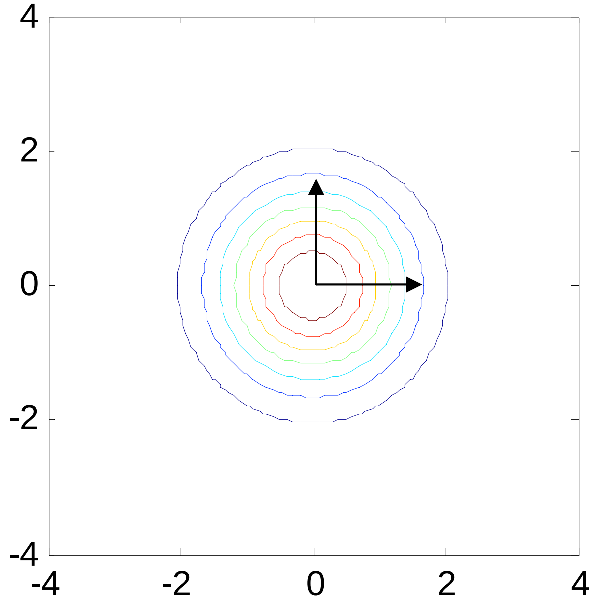
$\sigma_{height}=10cm$

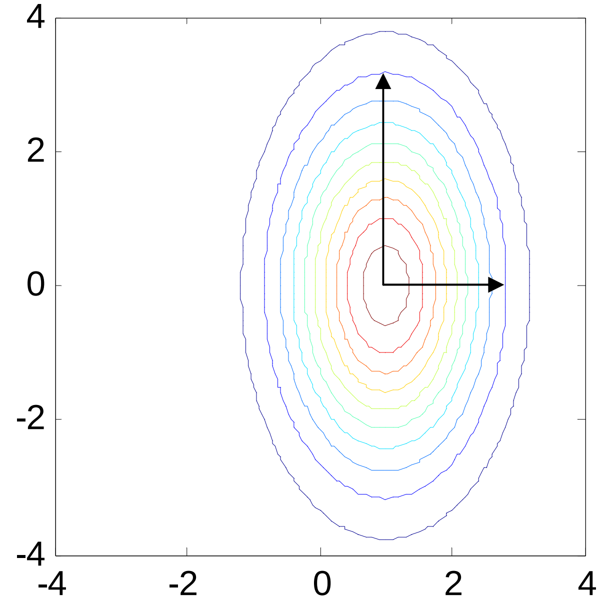
$\sigma_{weight}=10kg$

$r=0.7$

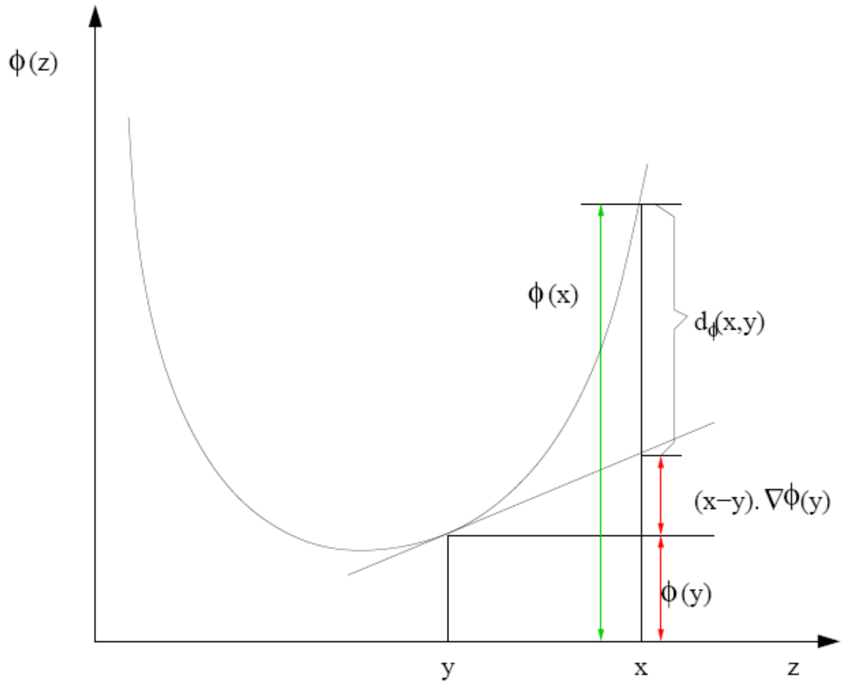
$d(\mathbf{x}_i,\mathbf{x}_j)$	Juan	John	Jean
Juan	-----	0.7529	4.8431

Independently of units!!





1.6 Distance measures: Generic



Bregman divergence

$$d_\phi(\mathbf{x},\mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \nabla\phi(\mathbf{y}) \rangle$$

Strictly convex, differentiable

$$\phi(\mathbf{x}) = \|\mathbf{x}\|^2 \longrightarrow$$ Euclidean distance

$$d_\phi(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|^2$$

$$\phi(\mathbf{x}) = \sum_{i=1}^{p} -x_i \log x_i \longrightarrow$$ Kullback-Leibler divergence

$$d_\phi(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p} x_i \log \frac{x_i}{y_i}$$

$$\phi(\mathbf{x}) = \sum_{i=1}^{p} -\log x_i \longrightarrow$$ Itakura-Saito distance

$$d_\phi(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p} \left(\frac{x_i}{y_i} - \log \frac{x_i}{y_i} - 1 \right)$$

CEU Universidad San Pablo