

# **Clustering (Data mining)**

Session 1: Introduction

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

## **Course outline: Session 1**

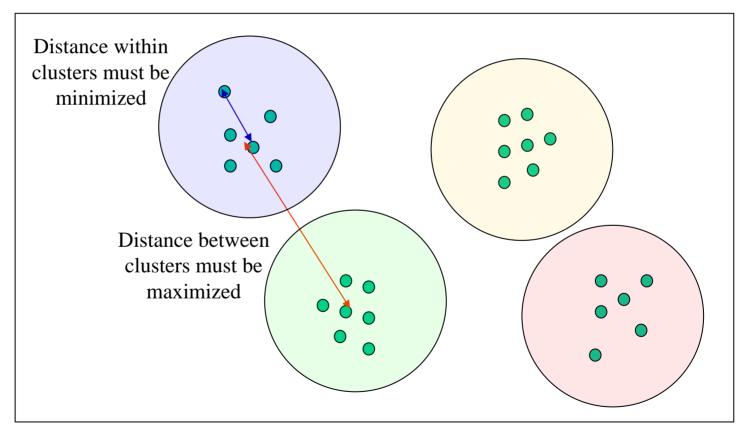
- 1. Introduction
  - 1.1 Problem formulation
  - 1.2 Types of features
  - 1.3 Feature extraction
  - 1.4 Graphical examination
  - 1.5 Data quality
  - 1.6 Distance measures
  - 1.7 Preprocessing
  - 1.8 Data reduction
  - 1.9 Types of clustering: partitional, hierarchical, probabilistic



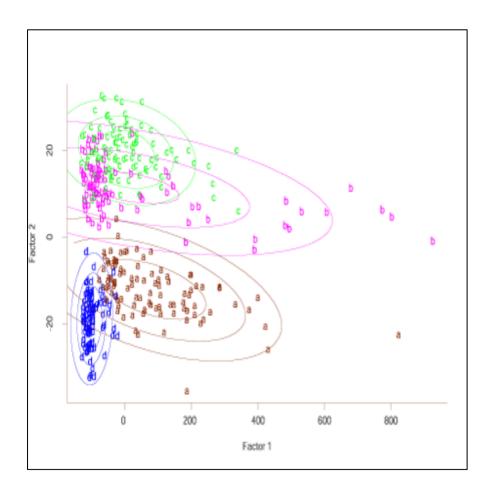
	dit View	Favorites	Tools Help											
G Bad	k 🕶 🕘	- 💌 😰	🏠 🔎 Search 🧧	Favorites (	3 🗟	1	📃 🎁	l 🔏 👘						
Address	🕘 http://lil	b.stat.cmu.ed	u/datasets/Plasma_Retinol	1									💌 🔁 Go	Li
Google	e <mark>G</mark> -		💌 Go 🚸 🍏 🚰 👻	🔂 Bookmarks	👻 🚳 32 bl	ocked 😽	Check 👻 🍕	🖏 Look for Ma	ap 👻 🔚 Ak	utoFill 🔒 Se	end to 👻 🏼 🏼		O Settings ▼	٩
Niere	nberg D	W, Stukel	l TA, Baron JA, 1	Dain BJ, Gr	eenberg	ER. Det	terminan	ts of pla	asma lev	els of b	eta-caro	tene and	d retinol.	j
Descr	iption:	This da	atafile contains	315 observ	ations o	on 14 va	riables.	This da	ata set	can be u	sed to d	emonstra	ate multip	le
Varis	ble Nam	es in ord	ler from left to	right:										
	AGE:	Age (yea	ars)	-										
	SEX:	Sex (1=1	Male, 2=Female).											
	SMOK	STAT: Smo	oking status (1=1	Never, 2=Fo	rmer, 3=	Current	Smoker)							
	QUET	ELET: Que	etelet (weight/()	height^2))										
	VITU	SE: Vitar	nin Use (1=Yes, :	fairly ofte	n, 2=Yes	, not o:	ften, 3=	No)						
	CALO	RIES: Nur	mber of calories	consumed p	er day.									
	FAT:	Grams of	f fat consumed p	er day.										
	FIBE	R: Grams	of fiber consume	ed per day.										
	ALCO	HOL • Numb	per of alcoholic	drinks con										
		110H. 140dia	Ser or arcomotic	urinks con	isumea pe	r week.								
			Cholesterol con:		-									
	CHOL	ESTEROL:		sumed (mg p	er day).									
	CHOL BETA	ESTEROL: DIET: Die	Cholesterol con	sumed (mg p ene consume	er day). d (mcg p	er day)								
	CHOL BETA RETD BETA	ESTEROL: DIET: Die IET: Diet PLASMA: H	Cholesterol cons etary beta-carot cary retinol cons Plasma beta-carot	sumed (mg p ene consume sumed (mcg tene (ng/ml	er day). d (mcg p per day)	er day)								
	CHOL BETA RETD BETA	ESTEROL: DIET: Die IET: Diet PLASMA: H	Cholesterol con: etary beta-carot ary retinol con:	sumed (mg p ene consume sumed (mcg tene (ng/ml	er day). d (mcg p per day)	er day)								
64	CHOL BETA RETD BETA	ESTEROL: DIET: Die IET: Diet PLASMA: H	Cholesterol cons etary beta-carot cary retinol cons Plasma beta-carot	sumed (mg p ene consume sumed (mcg tene (ng/ml	er day). d (mcg p per day)	er day)		170.3	1945	890	200	915		
64	CHOL BETA RETD BETA RETP	ESTEROL: DIET: Die IET: Diet PLASMA: H LASMA: PJ	Cholesterol com etary beta-caroto cary retinol com Plasma beta-caroto lasma Retinol (no	sumed (mg p ene consume sumed (mcg tene (ng/ml g/ml)	er day). d (mcg p per day) )	er day)		170.3	1945 75.8	890 2653	200 451	915 124	727	
	CHOL BETA RETD BETA RETP 2	ESTEROL: DIET: Diet IET: Diet PLASMA: H LASMA: P. 2	Cholesterol com etary beta-carot ary retinol com Plasma beta-carot lasma Retinol (na 21.4838 1	sumed (mg p ene consume sumed (mcg tene (ng/ml g/ml) 1298.8	er day). d (mcg p per day) ) 57	oer day) 6.3	0						727	
76	CHOL BETA RETD BETA RETP 2 2	ESTEROL: DIET: Diet IET: Diet PLASMA: P LASMA: P 2 1	Cholesterol cons stary beta-caroto ary retinol com- Plasma Beta-caroto [asma Retinol (no 21.4838 1 23.67691 20.0108 2 25.14062	sumed (mg p ene consume sumed (mcg tene (ng/ml g/ml) 1298.8 1	er day). d (mcg p per day) ) 57 <del>1032.5</del>	er day) 6.3 50.1 19.1	0 15.8		75.8	2653	451	124	<del>727</del> 615	
<del>76</del> 38	CHOL BETA RETD BETA RETP 2 2 2 2	ESTEROL: DIET: Diet IET: Diet PLASMA: P LASMA: P 2 1 2	Cholesterol com etary beta-carot lasma beta-carot lasma Retinol (nu 21.4838 1 23.87631 20.0108 2	sumed (mg p ene consume sumed (mcg tene (ng/ml g/ml) 1298.8 1 2372.3	er day). d (mcg p per day) ) 57 <u>1832.5</u> 83.6	6.3 50.1 19.1 97.5	0 15.8 14.1	0 257.9	<del>75.8</del> 6321	<del>2653</del> 660	451 328	<del>124</del> 721		
76 38 40 72 40	CHOL BETA RETD BETA RETP 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	ESTEROL: DIET: Diet IET: Diet PLASMA: P LASMA: P 2 1 2 2 2 2	Cholesterol cons etary beta-carot ary retinol com Plasma beta-carot lasma Retinol (ny 21.4838 1 23.87631 20.0108 2 25.14062 20.98504 27.52136	sumed (mg p ene consume sumed (mcg tene (ng/ml 1298.8 1 2372.3 3 1 3	er day). d (mcg p per day) ) 57 1032.5 83.6 2449.5 1952.1 1366.9	6.3 50.1 19.1 97.5 82.6 56	0 15.8 14.1 26.5 16.2 9.6	8 257.9 0.5	75.8 6321 332.6 170.8 154.6	<del>2653</del> 660 1061	451 328 864 1209 1439	124 721 153 92 148	615 799 654	
76 38 40 72 40 65	CHOL BETA RETD BETA RETP 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	ESTEROL: DIET: Diet IET: Diet PLASMA: P 2 2 2 2 1 2 1 2 1 2 1 2 1	Cholesterol com etary beta-carot ary retinol com Plasma beta-carot lasma Retinol (ne 21.4838 1 23.87631 20.0108 2 25.14062 20.98504 27.52136 22.01154	sumed (mg p ene consume sumed (mcg tene (ng/ml g/ml) 1298.8 1 2372.3 3 1 3 2	er day). d (mcg p per day). 57 1032.5 83.6 2449.5 1952.1 1366.9 2213.9	6.3 50.1 19.1 97.5 82.6 56 52	0 15.8 14.1 26.5 16.2 9.6 28.7	0 257.9 0.5 0	75.8 6321 332.6 170.8 154.6 255.1	2653 660 1061 2863 1729 5371	451 328 864 1209 1439 802	124 721 153 92 148 258	615 799 654 834	
76 38 40 72 40 65 58	CHOL BETA RETD BETA RETP 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	ESTEROL: DIET: Diet IET: Diet PLASMA: P 2 2 2 2 1 2 2 1 2 1 1 1 1	Cholesterol cons etary beta-carot ary retinol com Plasma beta-carot lasma Retinol (ny 21.4838 1 23.87631 20.0108 2 25.14062 20.98504 27.52136	sumed (mg p ene consume sumed (mcg tene (ng/ml 1298.8 1 2372.3 3 1 3 2 2 1	er day). d (mcg p per day) ) 57 1032.5 83.6 2449.5 1952.1 1366.9	6.3 50.1 19.1 97.5 82.6 56 52	0 15.8 14.1 26.5 16.2 9.6	8 257.9 0.5 0 1.3	75.8 6321 332.6 170.8 154.6	2653 660 1061 2863 1729	451 328 864 1209 1439	124 721 153 92 148	615 799 654	
76 38 40 72 40 65 58 35	CHOL BETA RETD BETA RETP 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	ESTEROL: DIET: Diet IET: Diet PLASMA: P 2 2 2 2 1 2 1 2 1 2 1 2 1	Cholesterol com etary beta-carot ary retinol com Plasma beta-carot lasma Retinol (ne 21.4838 1 23.87631 20.0108 2 25.14062 20.98504 27.52136 22.01154	sumed (mg p ene consume sumed (mcg tene (ng/ml g/ml) 1298.8 1 2372.3 3 1 3 2	er day). d (mcg p per day). 57 1032.5 83.6 2449.5 1952.1 1366.9 2213.9	6.3 50.1 19.1 97.5 82.6 56 52 63.4	0 15.8 14.1 26.5 16.2 9.6 28.7	0 257.9 0.5 0 1.3 0	75.8 6321 332.6 170.8 154.6 255.1	2653 660 1061 2863 1729 5371	451 328 864 1209 1439 802	124 721 153 92 148 258	615 799 654 834	
76 38 40 72 40 65 58	CHOL BETA RETD BETA RETP 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	ESTEROL: DIET: Diet IET: Diet PLASMA: P 2 2 2 2 1 2 2 1 2 1 1 1 1	Cholesterol com etary beta-carot ary retinol com lasma beta-carot 21.4838 1 23.87631 20.0108 2 25.14062 20.98504 27.52136 22.01154 28.75702	sumed (mg p ene consume sumed (mcg tene (ng/ml 1298.8 1 2372.3 3 1 3 2 2 1	er day). d (mcg p per day). 57 1032.5 83.6 2449.5 1952.1 1366.9 2213.9 1595.6	6.3 50.1 19.1 97.5 82.6 56 52 63.4 57.8	0 15.8 14.1 26.5 16.2 9.6 28.7 10.9	8 257.9 0.5 0 1.3 0 0	75.8 6321 332.6 170.8 154.6 255.1 214.1	2653 660 1061 2863 1729 5371 823	451 328 864 1209 1439 802 2571	124 721 153 92 148 258 64	615 799 654 834 825	
76 38 40 72 40 65 58 35	CHOL BETA RETD BETA RETP 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	ESTEROL: DIET: Diet IET: Diet PLASMA: P 2 2 1 2 2 1 2 1 1 2 1 1 1 1	Cholesterol cons stary beta-caroto ary retinol com Plasma beta-caroto (marked section) 21.4838 1 20.0108 2 25.14062 20.98504 27.52136 22.01154 28.75702 23.07662	sumed (mg p ene consume sumed (mcg tene (ng/ml g/ml) 1298.8 1 2372.3 3 1 3 1 3 1 3 1 3 1 3 3 1 3 3 1 3 3 1 3 3 1 3 3 1 3 3 1 3 3 1 3 3	er day). d (mcg p per day). 57 1032.5 83.6 2449.5 1952.1 1366.9 2213.9 1595.6 1800.5	6.3 50.1 19.1 97.5 82.6 56 52 63.4 57.8 39.6	0 15.8 14.1 26.5 16.2 9.6 28.7 10.9 20.3	8 257.9 0.5 0 1.3 0 0 0.6	75.8 6321 332.6 170.8 154.6 255.1 214.1 233.6	2653 660 1061 2863 1729 5371 823 2895	451 328 864 1209 1439 802 2571 944	124 721 153 92 148 258 64 218	615 799 654 834 825 517	
76 38 40 72 40 65 58 35 55	CHOL BETA RETD BETA RETP 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	ESTEROL: DIET: Diet IET: Diet PLASMA: P 2 2 2 2 1 2 1 2 1 1 2 1 1 2 1 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 2 1 2 2 2 1 2 2 2 1 2 2 2 2 1 2 2 2 2 1 2 2 2 2 1 2 2 2 2 1 2 2 2 2 2 2 2 1 2	Cholesterol cons atary beta-carot ary retinol com Plasma beta-carot 21.4838 1 20.0108 2 25.14062 20.98504 27.52136 22.01154 28.75702 23.07662 34.96995	sumed (mg p ene consume sumed (mcg tene (ng/ml g/ml) 1298.8 1 2372.3 3 1 3 2 1 3 3 2 1 3 3 3 3 3 3 3 3 3 3	er day). d (mcg p per day). 57 1032.5 83.6 2449.5 1952.1 1366.9 2213.9 1595.6 1800.5 1263.6	6.3 50.1 19.1 97.5 82.6 56 52 63.4 57.8 39.6 58	0 15.8 14.1 26.5 16.2 9.6 28.7 10.9 20.3 15.5	8 257.9 0.5 0 1.3 0 0 0.6 0	75.8 6321 332.6 170.8 154.6 255.1 214.1 233.6 171.9	2653 660 1061 2863 1729 5371 823 2895 3307	451 328 864 1209 1439 802 2571 944 493	124 721 153 92 148 258 64 218 81	615 799 654 834 825 517 562	
76 38 40 72 40 65 58 35 55 55 66	CHOL BETA RETD BETA RETP 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	ESTEROL: DIET: Diet IET: Diet PLASMA: P 2 2 1 2 1 2 1 2 1 1 2 1 2 1 2 2 1 2 2 1 2 2 1 2 2 2 1 2	Cholesterol cons stary beta-carot ary retinol com plasma beta-carot 21.4838 1 23.87691 20.0108 2 25.14062 20.98504 27.52136 22.01154 28.75702 23.07662 34.96995 20.94647	sumed (mg p ene consume sumed (mcg tene (ng/ml 1298.8 1 2372.3 3 1 3 2 1 3 2 1 3 2 1 3 3 1 3 2 1 3 3 1	er day). d (mcg p per day). 57 1092.5 83.6 2449.5 1952.1 1366.9 2213.9 1595.6 1800.5 1263.6 1460.8	6.3 50.1 19.1 97.5 82.6 56 52 63.4 57.8 39.6 58 49.3	0 15.8 14.1 26.5 16.2 9.6 28.7 10.9 20.3 15.5 18.2	8 257.9 0.5 0 1.3 0 0 0.6 0 1	75.8 6321 332.6 170.8 154.6 255.1 214.1 233.6 171.9 137.4	2653 660 1061 2863 1729 5371 823 2895 3307 1714	451 328 864 1209 1439 802 2571 944 493 535	124 721 153 92 148 258 64 218 81 184	615 799 654 834 825 517 562 935	
76 38 40 72 40 65 58 35 55 66 40	CHOL BETA RETD BETA 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	ESTEROL: DIET: Diet IET: Diet PLASMA: P 2 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1	Cholesterol cons etary beta-carot ary retinol com Plasma beta-carot lasma Retinol (ny 21.4838 1 23.07631 20.0108 2 25.14062 20.98504 27.52136 22.01154 28.75702 23.07662 34.96995 20.94647 36.43161	sumed (mg p ene consume sumed (mcg tene (ng/ml 1298.8 1 2372.3 3 1 3 2 1 3 2 1 3 1 3 2 1 3 3 1 2 1 3 3 1 2 1 2	er day). d (mcg p per day). 57 1032.5 83.6 2449.5 1952.1 1366.9 2213.9 1595.6 1800.5 1263.6 1460.8 1638.2	6.3 50.1 19.1 97.5 82.6 56 52 63.4 57.8 39.6 58 49.3	0 15.8 14.1 26.5 9.6 28.7 10.9 20.3 15.5 18.2 14.9	8 257.9 0.5 0 1.3 0 0 0.6 0 1 0	75.8 6321 332.6 170.8 154.6 255.1 214.1 233.6 171.9 137.4 130.7	2653 660 1061 2863 1729 5371 823 2895 3307 1714 2031	451 328 864 1209 1439 802 2571 944 493 535 492	124 721 153 92 148 258 64 218 81 184 91	615 799 654 834 825 517 562 935 741	
76 38 40 72 40 65 58 35 55 66 40 57	CHOL BETA RETD BETA RETP 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	ESTEROL: DIET: Diet IET: Diet PLASMA: P 2 2 1 2 2 1 2 2 1 1 1 1 2 2 1 1 1 1 2 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1	Cholesterol cons stary beta-caroto ary retinol cons Plasma beta-caroto lasma Retinol (n 21.4838 1 20.0108 2 25.14062 20.98504 27.52136 22.01154 28.75702 23.07662 34.96995 20.94647 36.43161 31.73039	sumed (mg p ene consume sumed (mcg tene (ng/ml g/ml) 1298.8 1 2372.3 3 1 3 2 1 3 3 1 3 1 3 2 1 3 3 1 2 3 3 3 3	er day). d (mcg p per day). 57 1032.5 83.6 2449.5 1952.1 1366.9 2213.9 1595.6 1800.5 1263.6 1460.8 1638.2 2072.9	6.3 50.1 19.1 97.5 82.6 52 63.4 57.8 39.6 58 39.6 58 49.3 106.7	0 15.8 14.1 26.5 16.2 9.6 28.7 10.9 20.3 15.5 18.2 14.9 9.6	8 257.9 0.5 0 1.3 0 0.6 0 1 0 0.9	75.8 6321 332.6 170.8 154.6 255.1 214.1 233.6 171.9 137.4 130.7 420	2653 660 1061 2863 1729 5371 823 2895 3307 1714 2031 1982	451 328 864 1209 1439 802 2571 944 493 535 492 1105	124 721 153 92 148 258 64 218 81 184 91 120	615 799 654 834 825 517 562 935 741 679	

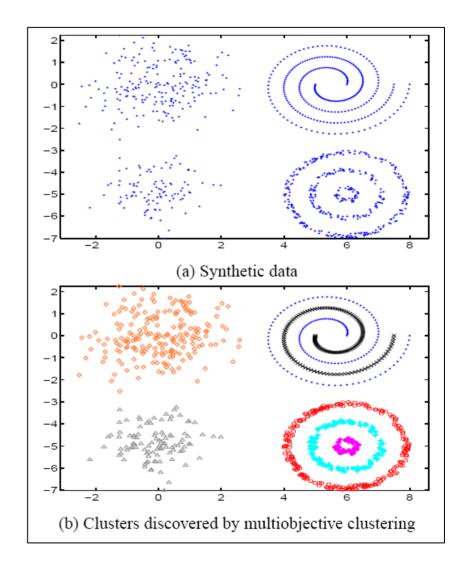


Find groups of points that are close to each other within the cluster and far from the rest of clusters











#### 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

Distance

definition

- Collect different attributes of customers based on their geographical and lifestyle related information.
- Define an apropriate distance measure between a pair of customers.
- Cluster algorithm

Cluster

validation

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



#### 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

Distance definition

Cluster validation

- $\rightarrow$  Identify frequently occurring terms in each document.
  - Form a similarity measure based on the frequencies of different terms.
- Cluster \_\_\_\_\_ Use it to cluster.
  - Do newly arrived documents fit in the clusters?



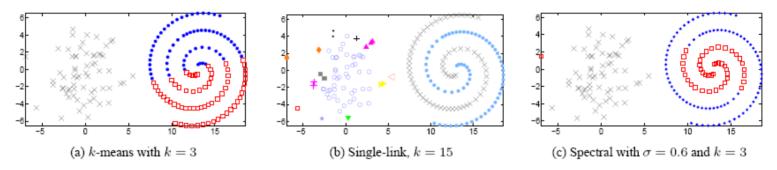


Figure 1: The resulting partitions by (a) k-means, (b) single-link and (c) spectral clustering on this "globular-spiral" data set.

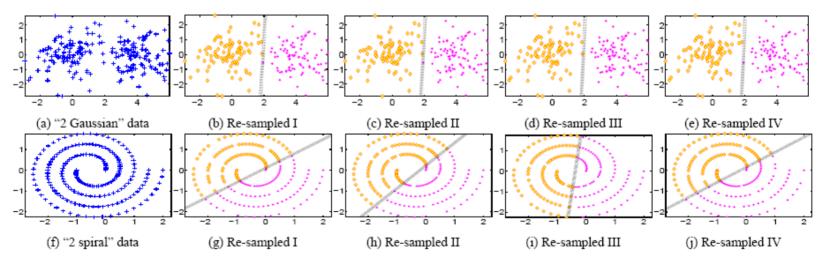
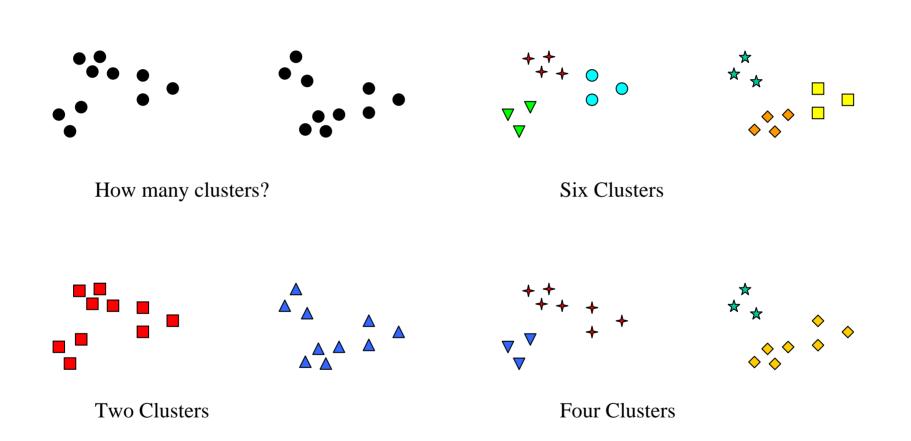
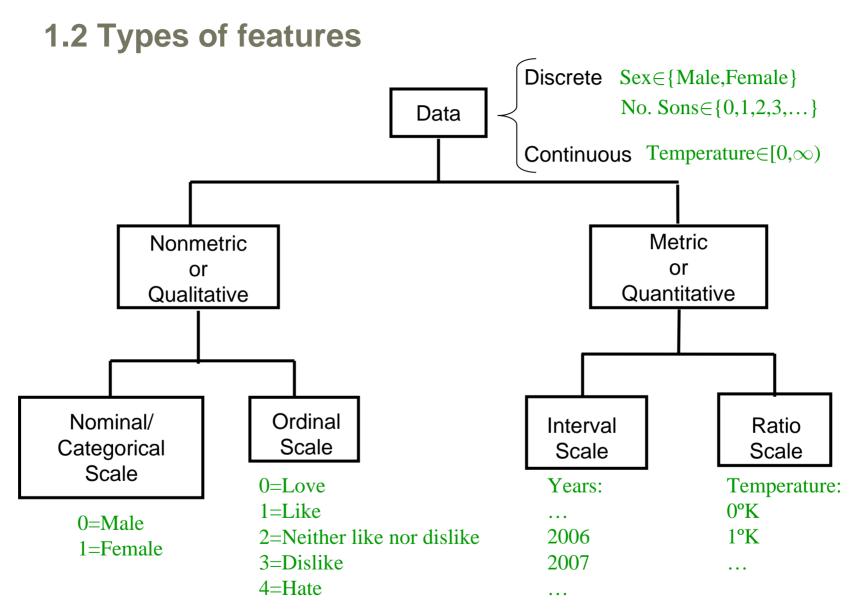


Figure 2: Results of k-means with k = 2 for different re-sampled versions of two data sets. Dotted lines in the figures correspond to the cluster boundaries. The partitions of "2 Gaussian" data set are almost the same for different re-sampled versions, suggesting that k-means with k = 2 gives good clusters.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











## **1.2 Types of features**

#### Coding of categorical variables

Hair Colour {Brown, Blond, Black, Rec	$\xrightarrow{\text{No order}} \left( x_{Brown}, x_{Blond}, x_{Black}, x_{Red} \right) \in \left\{ 0, 1 \right\}^{4}$
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}	$\xrightarrow{\text{Implicit order}} x_{size} \in \{0, 1, 2\}$
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

			lasma beta-caro	tinol consumed (mcg per day) beta-carotene (ng/ml)									
+			lasma Retinol (no										
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	15.0	-0	75.0	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		18.2	1	137.4	1714	535	184	935
40	2	1	36.43161	2	1638.2		14.9	Ô	130.7	2031	492	91	741

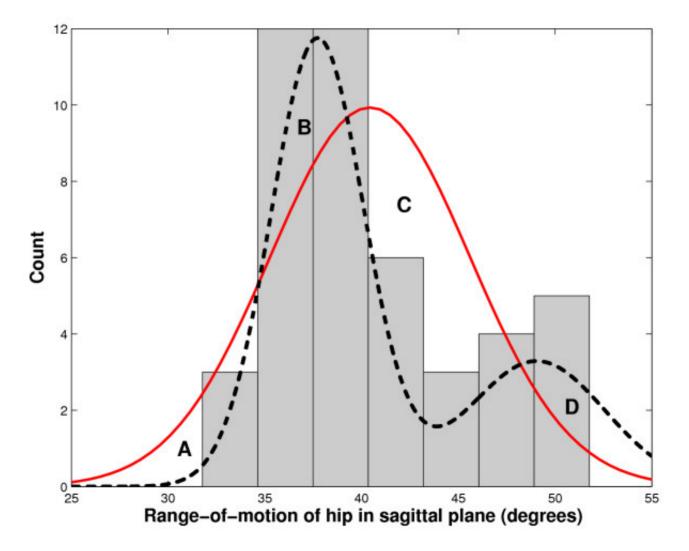


## **1.4 Graphical examination**

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

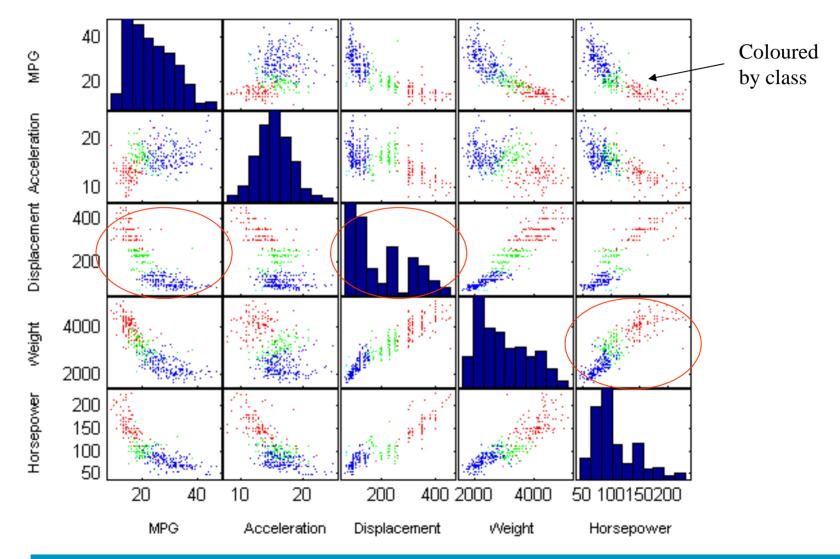


## **1.4 Graphical examination: Univariate distribution**





#### **1.4 Graphical examination: Scatter plots**





## 1.5 Data quality: Missing data

Types of missing data:

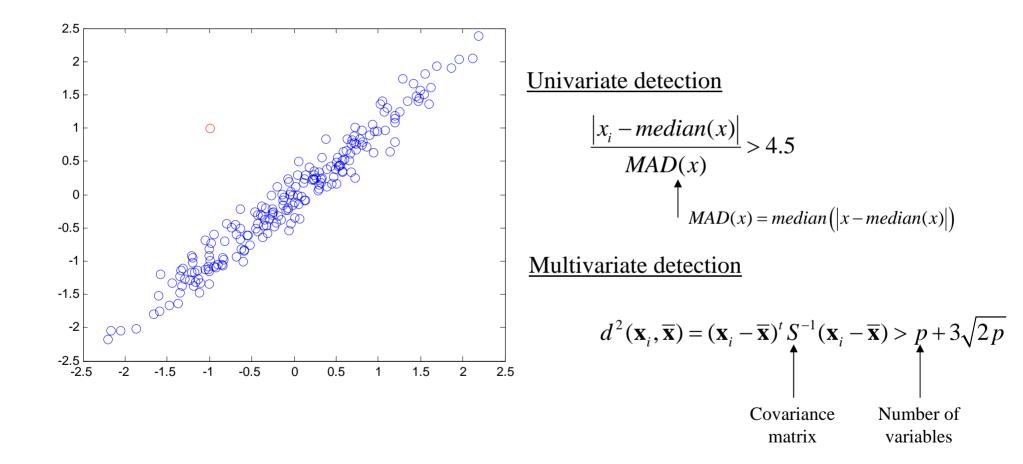
- Missing Completely At Random (MCAR)
- Missing at Random (MAR)

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**





## **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 heterogeous sources

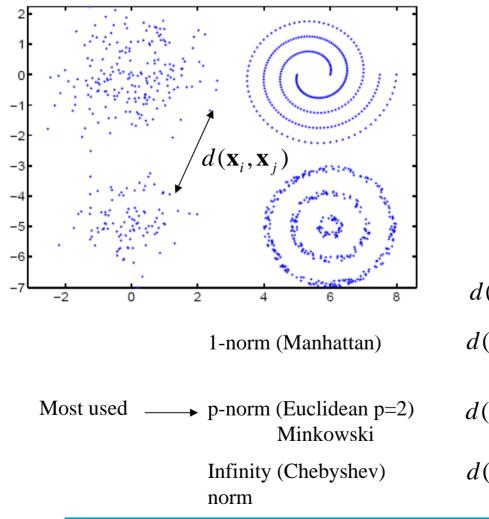
Example:

- Same person with multiple email addresses

Data cleansing:

- Remove duplicates (by partial distance, by classification)





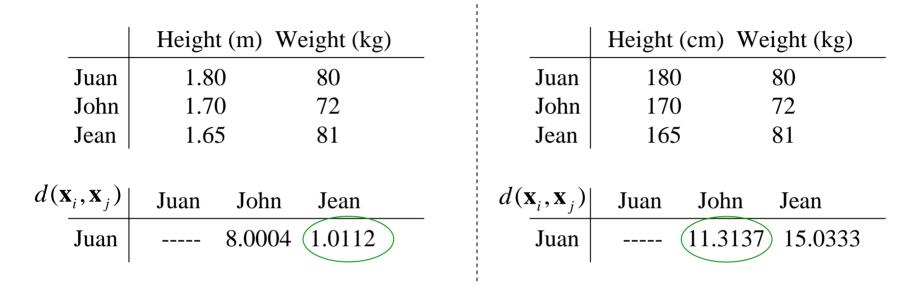
$$d(\mathbf{x}_{i}, \mathbf{x}_{j})$$

$$d(\mathbf{x}_{i}, \mathbf{x}_{j}) = \sum_{s=1}^{n} |x_{is} - x_{js}|$$

$$d(\mathbf{x}_{i}, \mathbf{x}_{j}) = \left(\sum_{s=1}^{n} (x_{is} - x_{js})^{p}\right)^{\frac{1}{p}}$$

$$d(\mathbf{x}_{i}, \mathbf{x}_{j}) = \max_{s} |x_{is} - x_{js}|$$



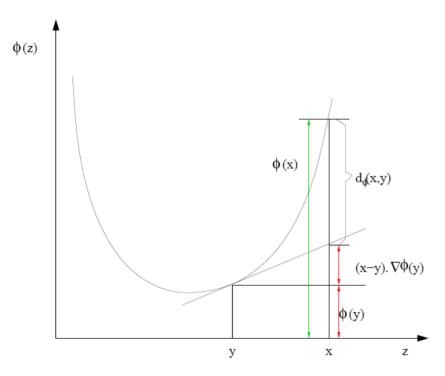


Matrix-based distance  $d(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^t M^{-1} (\mathbf{x}_i - \mathbf{x}_j)$ Euclidean distance  $d(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^t I^{-1} (\mathbf{x}_i - \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^t (\mathbf{x}_i - \mathbf{x}_j)$ Mahalanobis distance  $d(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^t \Sigma^{-1} (\mathbf{x}_i - \mathbf{x}_j)$ 



Mahalanobis distance  $d(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^t \Sigma^{-1} (\mathbf{x}_i - \mathbf{x}_j)$  $\Sigma = \begin{pmatrix} \sigma_{height}^2 & r\sigma_{height}\sigma_{weight} \\ r\sigma_{height}\sigma_{weight} & \sigma_{weight}^2 \end{pmatrix} = \begin{pmatrix} 100 & 70 \\ 70 & 100 \end{pmatrix} \qquad \frac{d(\mathbf{x}_i, \mathbf{x}_j)}{\text{Juan John}} \qquad \frac{Juan}{Juan} \qquad \frac{Juan}{Juan}$ Jean 4.8431  $\sigma_{height} = 10cm$ Independently of units!!  $\sigma_{\scriptscriptstyle weight} = 10 kg$ 4 r = 0.72 2 0 0 -2 -2 -4 -4 -2 2 -2 0 0 2 -4 4 -4 4





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  

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

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

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

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

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

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

