ANALISI DEI DATI PER IL MARKETING 2014

Marco Riani
mriani@unipr.it
http://www.riani.it

LA CLASSIFICAZIONE CAP IX , pp.367-457

- Problema generale della scienza (Linneo, …)
- Analisi discriminante
- *Cluster Analysis* (analisi dei gruppi)

ANALISI DISCRIMINANTE

- Modello probabilistico
  - Dati *k* gruppi noti a priori, si devono assegnare nuovi elementi ad uno dei gruppi in base ad un vettore di dati, con una piccola probabilità di errore.
- Esempi di applicazione:
- Aziende sane ed a rischio fallimento
- Direct marketing: acquirenti, non acquirenti
- Quadri autentici e quadri falsi

CLUSTER ANALYSIS
(Analisi dei gruppi)

- **Metodo esplorativo**
- I gruppi non sono noti a priori e non si è neppure certi della loro esistenza
- I gruppi di unità vengono individuati in base ai valori di *p* variabili (quantitative e/o qualitative)

- I gruppi devono presentare:
- **Coesione interna** (gruppi **omogenei**)
- **Separazione esterna** (diversità da un gruppo all'altro)
- Esempi:
  - Classificazione di *n* province (o *n* comuni) in base a *p* indicatori di benessere economico
  - Segmentazione dei clienti o dei prodotti

PROBLEMI DI SCELTA NELLA CLUSTER ANALYSIS

- Scopo (esempi)
- Variabili
- Distanza (o similarità)
  - Euclidea o della città a blocchi
- Metodo di raggruppamento
  - Gerarchico
  - Non gerarchico
- Numero di gruppi
- Valutazione dei risultati
- Caratteristiche dei gruppi omogenei individuati
- Confronti tra classificazioni differenti

DISTANZE
variabili tutte quantitative

- Distanza euclidea: $d_{ij} = \sqrt{\sum [x_{is} - x_{js}]^2}$
- Distanza della città a blocchi: $d_{ij} = \sum | x_{is} - x_{js} |$

- **Caratteristiche (riepilogo):**
- Distanza city-block ≥ dist. Euclidea
- Distanza city-block risente in misura minore degli *outliers*
- Tra i due tipi di distanza non esiste sempre una relazione monotona (cioè gli ordinamenti o graduatorie possono essere diversi)

- CALCOLO SUGLI SCOSTAMENTI STANDARDIZZATI

METODI GERARCHICI procedura

- Metodi per passi (o stadi) successivi
- Inizialmente *n* unità distinte
- Nel primo passo si uniscono le unità con minore distanza che formano il primo gruppo
- Nel generico passo successivo:
- Si calcola la distanza CON UN CERTO CRITERIO tra il gruppo ottenuto e gli altri "gruppi" (anche formati da un solo elemento)
- Si uniscono i "gruppi" (elementi) con la minore distanza
- Si itera la procedura sino a quando tutti gli elementi sono riuniti in unico gruppo

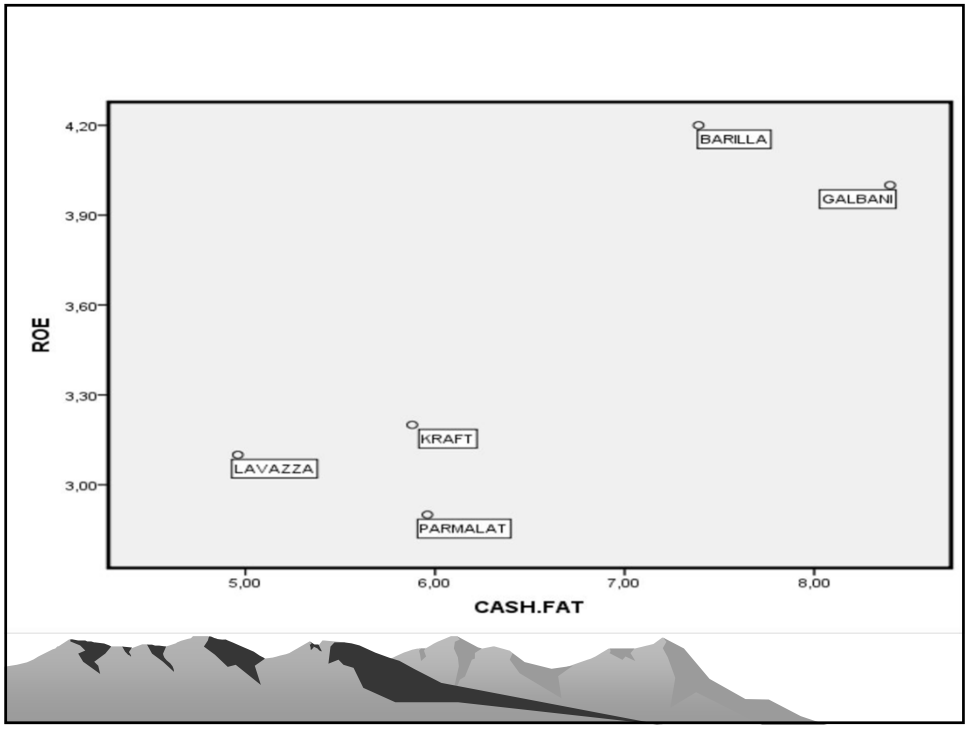
METODI GERARCHICI

- Caratteristiche comuni:
- Forniscono una "famiglia" di partizioni partendo da quella banale in cui tutti gli elementi sono distinti (i "gruppi" coincidono con le unità) sino a quella in cui tutte unità sono riunite in un unico gruppo.
- I gruppi che si ottengono ad un certo stadio possono solo essere riuniti nei passi successivi (ma non scissi)

I diversi metodi gerarchici differiscono per il CRITERIO DI CALCOLO DELLA DISTANZA TRA DUE GRUPPI

DISTANZA TRA DUE GRUPPI

- E’ generalmente calcolata in funzione delle distanze tra coppie di unità (che compaiono nella matrice delle distanze)
- Esempio: un gruppo di 3 ed uno di 2 unità



DISTANZA TRA DUE GRUPPI
Legame singolo e completo

- Legame singolo (*single linkage*):
  - Distanza = MINIMO delle distanze tra gli elementi di un gruppo e quelli dell'altro
- Legame completo:
  - Distanza = MASSIMO delle distanza tra gli elementi d'un gruppo e quelli dell'altro

Esempio legame singolo: 4 supermercati, 7 variabili standardizzate

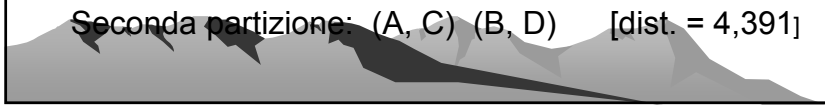
Matrice delle distanze

	Distanza euclidea			
	1:A	2:B	3:C	4:D
1:A	,000	13,123	2,506	17,060
2:B	13,123	,000	15,628	4,391
3:C	2,506	15,628	,000	19,542
4:D	17,060	4,391	19,542	,000

Questa è una matrice di dissimilarità

Prima partizione: (A, C) B , D [min dist. = 2,506]

Seconda partizione: (A, C) (B, D) [dist. = 4,391]



Esempio con SPSS: legame singolo

	1:A	2:B	3:C	4:D
1:A	,000	13,123	2,506	17,060
2:B	13,123	,000	15,628	4,391
3:C	2,506	15,628	,000	19,542
4:D	17,060	4,391	19,542	,000

Programma di agglomerazione

Stadio	Cluster accorpati		Coefficienti	Stadio di formazione del cluster		Stadio successivo
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	3	2,506	0	0	3
2	2	4	4,391	0	0	3
3	1	2	13,123	1	2	0

Distanza tra i gruppi (A, C) e (B, D) =
min dist. =min (A,D) (A,B) (C,B) (C,D) =
dist (A, B) = 13,123

Esempio con SPSS: legame completo

	1:A	2:B	3:C	4:D
1:A	,000	13,123	2,506	17,060
2:B	13,123	,000	15,628	4,391
3:C	2,506	15,628	,000	19,542
4:D	17,060	4,391	19,542	,000

Programma di agglomerazione

Stadio	Cluster accorpati		Coefficienti	Stadio di formazione del cluster		Stadio successivo
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	3	2,506	0	0	3
2	2	4	4,391	0	0	3
3	1	2	19,542	1	2	0

Distanza tra i gruppi (A, C) e (B, D) =
max dist (A,D) (A,B) (C,B) (C,D) = dist
(C, D) = 19,542

Interpretazione del “programma di agglomerazione” di SPSS

- Stadio = passo della classificazione gerarchica (da 1 a n-1)
- Cluster accorpati = “gruppi” (elementi) che si uniscono
- Coefficienti = livelli di distanza
- Stadio di formazione del cluster:
- se =0 indica che il “gruppo” è costituito da una singola unità;
- se =1, 2, 3, …indica un gruppo di più elementi che si è formato in precedenza, al passo corrispondente
- Stadio successivo = indica il passo in cui il gruppo ottenuto si riunirà ad altri gruppi

DISTANZA TRA DUE GRUPPI Legame medio

- Legame medio (fra gruppi):
- Distanza tra due gruppi = MEDIA delle $n_1.n_2$ distanze tra gli elementi d’un gruppo e quelli dell’altro
- ESEMPIO 4 supermercati, 7 variabili

Esempio con SPSS: legame medio

	1:A	2:B	3:C	4:D
1:A	,000	13,123	2,506	17,060
2:B	13,123	,000	15,628	4,391
3:C	2,506	15,628	,000	19,542
4:D	17,060	4,391	19,542	,000

Programma di agglomerazione

Stadio	Cluster accorpati		Coefficienti	Stadio di formazione del cluster		Stadio successivo
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	3	2,506	0	0	3
2	2	4	4,391	0	0	3
3	1	2	16,338	1	2	0

Distanza tra i gruppi (A, C) e (B, D)
= media 4 dist. = 16,338

Confronti fra i metodi

- Proprietà formali: pp. 384-387 (teoremi Zani)
- Legame singolo: effetto “catena” (p. 382)
  ma è l’unico che individua gruppi non circolari
- Legame completo: gruppi compatti, ma solo di forma approssimativamente sferica
- Legame medio: può essere un buon compromesso (nella pratica risulta spesso più simile al legame completo)

DENDROGRAMMA

- Diagramma ad albero, verticale o orizzontale (SPSS), per la rappresentazione della successione di partizioni
- Le “radici” dell’albero sono le unità iniziali
- A livelli crescenti di distanza si uniscono i gruppi (elementi) tra loro
- Costruzione manuale del dendrogramma
- In SPSS le distanze sono riscalate (ridimensionate) nell’intervallo 0 - 25 per rendere comparabili i dendrogrammi ottenuti con metodi diversi

Esempio: 5 modelli di auto

- 7 Variabili:
- Prezzo, cilindrata, lunghezza, peso, potenza, accelerazione, consumo
- Distanza della città a blocchi sugli scostamenti standardizzati
- Metodo del legame singolo

Matrice delle distanze

Matrice delle distanze

Caso	Distanza City-Block				
	1:PUNTO	2:BRAVO	3:FIESTA	4:CORSA	5:GOLF
1:PUNTO	,000	10,842	3,384	2,888	9,146
2:BRAVO	10,842	,000	13,777	12,646	6,356
3:FIESTA	3,384	13,777	,000	2,137	12,081
4:CORSA	2,888	12,646	2,137	,000	10,950
5:GOLF	9,146	6,356	12,081	10,950	,000

Questa è una matrice di dissimilarità

LEGAME SINGOLO costruzione manuale del dendrogramma

Matrice delle distanze

Caso	Distanza City-Block				
	1:PUNTO	2:BRAVO	3:FIESTA	4:CORSA	5:GOLF
1:PUNTO	,000	10,842	3,384	2,888	9,146
2:BRAVO	10,842	,000	13,777	12,646	6,356
3:FIESTA	3,384	13,777	,000	2,137	12,081
4:CORSA	2,888	12,646	2,137	,000	10,950
5:GOLF	9,146	6,356	12,081	10,950	,000

Questa è una matrice di dissimilarità

Programma di agglomerazione

Stadio	Cluster accorpati		Coefficienti	Stadio di formazione del cluster		Stadio successiv
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	3	4	2,137	0	0	2
2	1	(3,4) 3	2,888	0	1	4
3	2	5	6,356	0	0	4
4	(1,3,4) 1	(2,5) 2	9,146	2	3	0

LEGAME COMPLETO

costruzione manuale del dendrogramma

Programma di agglomerazione

Stadio	Cluster accorpati		Coefficient	Stadio di formazion del cluster		Stadio successivo
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	3	4	2,137	0	0	2
2	1	3	3,384	0	1	4
3	2	5	6,356	0	0	4
4	1	2	13,777	2	3	0

LETTURA DI UN DENDROGRAMMA

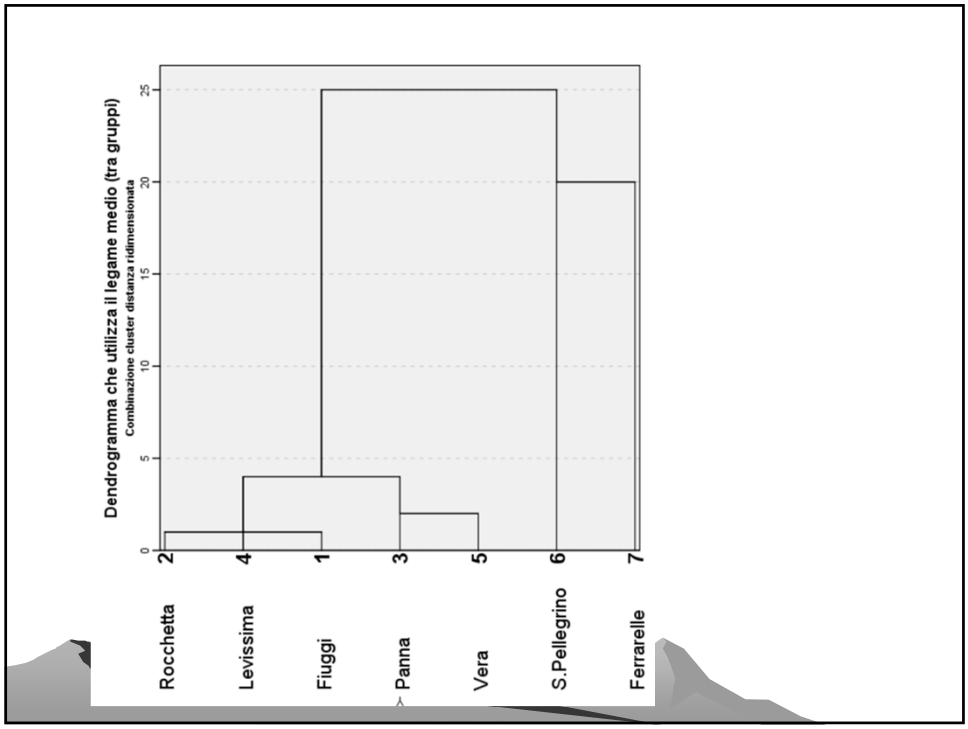
- Mostra le partizioni che si ottengono a livelli crescenti di distanza
- Si può proporre un **“taglio”** in corrispondenza d’un salto grande *(individuazione del **numero ottimo di gruppi**)*
- Vi possono essere più tagli ragionevoli
- Le ultime “radici” che si uniscono al tronco possono essere *outliers*

ESERCIZIO

- Si è applicata la *cluster analysis* a 7 marche di acque minerali, considerando 6 variabili inerenti alla composizione chimica.
- Utilizzando SPSS con il metodo del legame medio e la distanza euclidea sugli scostamenti standardizzati si è ottenuta la tabella riportata di seguito.
- Si costruisca il corrispondente dendrogramma, proponendo anche, con adeguato commento, un "taglio" opportuno per il medesimo e scrivendo la corrispondente partizione.

PROGRAMMA DI AGGLOMERAZIONE
(1=Fiuggi; 2=Rocchetta; 3=Panna; 4=Levissima; 5=Vera; 6=S. Pellegrino; 7=Ferrarelle

Stadio	Cluster accorpati		Coefficienti	Stadio di formazione		Stadio successivo
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	2	4	0.622	0	0	2
2	1	2	0.768	0	1	4
3	3	5	0.975	0	0	4
4	1	3	1.263	2	3	6
5	6	7	4.074	0	0	6
6	1	6	5.108	4	5	0



Metodi che utilizzano anche la matrice dei dati:

METODO DEL CENTROIDE

Centroide di n_1 punti e p variabili = punto con coordinate uguali alle medie delle p variabili

- Metodo di *cluster analysis* del centroide: Distanza tra due gruppi = distanza tra il centroide d'un gruppo ed il centroide dell'altro
- E' simile al metodo del legame medio

METODO DI WARD

- Obiettivo della cluster analysis è ottenere gruppi omogenei (con piccola variabilità all'interno)

Il metodo di Ward minimizza la **devianza** nei gruppi (funzione obiettivo):

dev. tot. = dev. nei + dev. fra

T = W + B

- Ad ogni passo si uniscono i "gruppi" (elementi) che comportano il minore incremento di W
- Non richiede il calcolo preliminare della matrice delle distanze
- Deve essere calcolato sugli **scostamenti standardizzati**

Scomposizione della devianza

(p. 390)

n unità stat. divise in g gruppi, p variabili

$$T = \sum_{s=1}^{p}\sum_{i=1}^{n}\left(x_{is} - M(X_s)\right)^2 \quad \text{DEVIANZA TOTALE}$$

W (within) = DEVIANZA NEI g GRUPPI

(funzione delle differenze tra i singoli valori

e le rispettive medie di gruppo)

B (between) = DEVIANZA FRA I GRUPPI

(funzione delle differenze fra le medie di gruppo

e la media generale)

SCELTA DEL NUMERO DI GRUPPI

- Un metodo gerarchico fornisce una successione di partizioni delle *n* unità
- **Occorre scegliere la migliore**:
- *Gruppi omogenei: piccola devianza nei gruppi*
- *Numero di gruppi rispondente alle finalità della classificazione*
- *Possibilità di più soluzioni*
- **Tabella "programma di agglomerazione" di SPSS:**
- Colonna "coefficienti" (distanze):
- *"Salto" notevole tra il valore allo stadio i ed il successivo (i+1): scelgo la partizione dello stadio i*
- Esame del dendrogramma

Valutazioni delle classificazioni: tre domande fondamentali (p. 406)

- Qual è la validità della partizione individuata con un certo metodo?
- Quali analogie esistono tra classificazioni ottenute con i vari metodi?
- Quali sono le caratteristiche di ogni segmento?

GIUDIZIO SULLA BONTA' DELLA CLASSIFICAZIONE: indice R^2

- dev. tot. = dev. nei + dev. fra
- **T = W + B**
- Calcolo di **$R^2 = 1 - W/T = B/T$**
- $R^2 = 1$: Classificazione perfetta
- $R^2 = 0$: Classificazione pessima
- Non è fornito direttamente da SPSS
- Serve anche per la scelta del numero di gruppi
- Tornare ai rapporti di correlazione (p. 143)

CLUSTER DI APPARTENENZA

- Fornito da SPSS:
- *Per un numero di gruppi fisso*
- *Per numeri di gruppi compresi in un intervallo*
- Può essere salvato come nuova variabile
- Può essere utilizzato per partizionare la matrice dei dati, **individuare le caratteristiche di ogni gruppo e calcolare R^2**

STATISTICHE SULLE CARATTERISTICHE DEI CLUSTER

- Procedura SPSS: matrice partizionata in base a “cluster di appartenenza”
- Dati – Distingui
- Distingui i risultati per gruppo (confronta)
- Statistiche descrittive – descrittive
- LE MEDIE DI OGNI VARIABILE IN CIASCUN GRUPPO NE SINTETIZZANO LE CARATTERISTICHE

CALCOLO DI R^2 CON SPSS

(P. 407-412)

$R^2 = 1 - W/T = B/T$

Procedura:

Analizza – confronta medie – ANOVA univariata

Variabili: gli scostamenti standardizzati

Fattore: la nuova variabile aggiunta nella matrice dei dati che contiene le “etichette” dei gruppi individuati con la cluster analysis

Si ottiene la tabella dell’ANOVA, che presenta le Devianze (Somma dei quadrati *degli scostamenti*) TOTALE, Fra e NEI

Si può quindi calcolare R^2 per ogni variabile e globale

8 AUTO, 7 VARIABILI, 2 gruppi

Prezzo: $R^2 = 0.865$

Cilindrata : $R^2 = 0.425$

ANOVA univariata

		Somma dei quadrati	df	Media dei quadrati
Punteg(PREZZO)	Fra gruppi	6,054	1	6,054
	Entro gruppi	,946	6	,158
	Totale	7,000	7	
Punteg(CILINDRATA)	Fra gruppi	2,975	1	2,975
	Entro gruppi	4,025	6	,671
	Totale	7,000	7	

R^2 per ogni variabile e globale (2 gruppi)

- PREZZO: $R^2 = 6.054 : 7 = 0.865$
- CILINDRATA: $R^2 = 2.975 : 7 = 0.425$
- LUNGHEZZA: $R^2 = 0.621$
- PESO: $R^2 = 0.698$
- POTENZA: $R^2 = 0.800$
- ACCEL. : $R^2 = 0.751$
- CONSUMO: $R^2 = 0.847$
- ----------
- **GLOBALE: $R^2 = 0.715$**

R^2 per ogni variabile e globale (3 gruppi)

- PREZZO: $R^2 = 0.912$
- CILINDRATA: $R^2 = 0.641$
- LUNGHEZZA: $R^2 = 0.874$
- PESO: $R^2 = 0.781$
- POTENZA: $R^2 = 0.914$
- ACCEL. : $R^2 = 0.858$
- CONSUMO: $R^2 = 0.883$
- ----------
- **GLOBALE: $R^2 = 0.838$**

CONFRONTI TRA CLASSIFICAZIONI

- Classificazioni diverse d'un medesimo insieme di *n* unità:
- Diversi insiemi di variabili (es. caratteristiche dei clienti e variabili comportamentali)
- Stesso insieme di variabili ma:
  - Diversa distanza
  - Diverso metodo di raggruppamento

CONFRONTI TRA CLASSIFICAZIONI (continua)

- Tabella di contingenza delle *n* unità riferita ai gruppi delle due partizioni
- Classificazioni identiche: una sola frequenza non nulla per ogni riga e ogni colonna
- Indici di associazione
- Metodi ad hoc

3 gruppi: legame singolo e completo

Tavola di contingenza Single Linkage * Complete Linkage

Conteggio

		Complete Linkage			
		1	2	3	Totale
Single Linkage	1	2	4	0	6
	2	0	0	1	1
	3	0	0	1	1
Totale		2	4	2	8

3 gruppi: legame medio e completo

Tavola di contingenza Average Linkage (Between Groups) * Complete Linkage

Conteggio

		Complete Linkage			
		1	2	3	Totale
Average Linkage (Between Groups)	1	2	0	0	2
	2	0	4	0	4
	3	0	0	2	2
Totale		2	4	2	8

RIEPILOGO PROCEDURA SPSS

Classificazione – Cluster gerarchica

- **SCELTE - Metodo:**
- ***Metodo di raggruppamento***
- ***Misura di distanza o similarità***
- ***Standardizzazione***
- **OUTPUT:**
- ***Statistiche:Programma di agglomerazione***
- ***Grafici: dendrogramma***
- ***Salva: cluster di appartenenza***
- ***R quadro: ANOVA***
- **CONFRONTI:**
- ***Statistiche descrittive – Tavole di contingenza***

sulle colonne "cluster di appartenenza" con metodi diversi

OSSERVAZIONI FINALI

- Un medesimo insieme di unità e variabili (stessa matrice **X**) può essere classificato in modi diversi
- La scelta della partizione "migliore" dipende dalle finalità dell'analisi
- Considerando criteri diversi (distanze, metodi di raggruppamento), i gruppi che si mantengono nelle varie classificazioni corrispondono alla reale "struttura" (pattern) dei dati, non dipendendo dall'algoritmo scelto
- Se n è molto grande (>1.000): metodi non gerarchici (tecniche di DATA MINING)

CLASSIFICAZIONE NON GERARCHICA

Metodo applicabile anche in presenza di *data sets* molto numerosi
Tecnica di DATA MINING

METODI NON GERARCHICI (p.424 e seg.)

- Caratteristica generale:
- Date *n* unità e *p* variabili si cerca un'unica partizione con *g* gruppi (*g* fissato a priori) che soddisfi una condizione di ottimo (ad esempio: min W)
- Non è possibile una soluzione analitica se *n* è anche solo moderatamente grande
  - Esempio: *n* = 20 ; *g* = 4; numero di partizioni possibili > 45 miliardi!
- Procedure iterative approssimate

Metodo delle k-medie

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

- $\boldsymbol{\mu}_i$ è la media dei punti che si trovano nel cluster i S_i

PREGI E DIFETTI

- Adatti anche per *n* molto elevato
- Individuano “tipologie” di unità, ma non pongono in luce l’aggregazione successiva degli elementi
- I risultati possono dipendere dall’ordine in cui sono elencate le unità

PROCEDURA ITERATIVA
METODO DELLE K (*g*) MEDIE

- Scelta (**appropriata**) dei *g* “poli” (centroidi, centri, *seeds*) iniziali (in automatico in SPSS)
- Per ogni unità, calcolo della distanza da ciascuno dei *g* poli ed assegnazione al gruppo più vicino (con minore distanza)
- Calcolo del centroide di ogni gruppo (nuovi poli)
- Ripetizione dei passi due e tre sino ad una regola di stop

REGOLE DI STOP

- Convergenza (assenza di variazioni tra il passo *t* e *t+1*):
  - Raggiungimento d'un ottimo (locale)
- Variazione minore d'una "soglia"
- Numero massimo di iterazioni
  - Adatto nel caso di *n* molto grande

VANTAGGI E SVANTAGGI

- **Procedura veloce (tempi brevi di elaborazione)**
- **Consente di classificare in *g* gruppi anche *data sets* molto numerosi (*n* ≈ 1.000.000)**
- **Può essere ripetuta più volte con diversi valori di *g* (numero di gruppi)**
- **Applicabile quando non interessa seguire l'aggregazione delle singole unità, bensì creare delle "tipologie"**
- **Può condurre a soluzioni di ottimo relativo, condizionate dalla scelta iniziale dei poli e dall'ordine delle unità**

SCELTA DEL NUMERO DI GRUPPI

- **Intervallo di valori di *g* (ad es.: da 2 a 10)**
- **Classificazione (partizione) corrispondente per ogni *g***
- **Valutazione delle partizioni mediante:**
  $R^2 = 1 – W/T$ e confronti
- **Aumentando *g* migliora R^2**

ULTERIORI PROBLEMI:

- **In SPSS: per variabili quantitative, calcolo preliminare degli scostamenti standardizzati**
- **Confronti tra metodi gerarchici e non gerarchici (per *n* piccolo)**

SCELTE NELLE K-MEDIE DI SPSS

- Calcolo degli scost. standar. delle variabili quantitative: *analizza-statistiche descrittive-descrittive-salva val. stand. come variabili*
- *Analizza-classificazione-cluster k-medie*
- Inserire i “punteggi” delle variabili (scost. stand.)
- Scegliere il numero di gruppi
- Itera e classifica
- Salva: cluster di appartenenza (eventualmente distanza dal centro)
- OPZIONI: centri iniziali, tabella ANOVA

STATISTICHE SULLE CARATTERISTICHE DEI CLUSTER

- Procedura SPSS: matrice partizionata in base a “cluster di appartenenza”
- Dati – Distingui
- Distingui i risultati per gruppo (confronta)
- Statistiche descrittive – descrittive
- **LE MEDIE DI OGNI VARIABILE (O FREQ. REL. PER VARIAB. BINARIE) IN CIASCUN GRUPPO NE SINTETIZZANO LE CARATTERISTICHE**

CARATTERISTICHE DI OGNI GRUPPO

- Matrice partizionata in base al cluster di appartenenza
- *Dati – distingui – distingui i risultati per gruppo – gruppi basati su*: numero di cluster del caso
- *Analizza - Statistiche descrittive – descrittive*: si ottengono i valori medi, min, max, stand. dev. di ogni variabile per ciascun gruppo

Caratteristiche dei gruppi individuati

- Valutazione della bontà della classificazione: indice R^2
- Matrice partizionata in base a: *numero di cluster del caso*
- Calcolo delle “medie” (frequenze relative) per ciascun gruppo e ciascuna variabile