FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Prof.dr.sc. N. Bogunović

Prof.dr.sc. B. Dalbelo Bašić

OTKRIVANJE ZNANJA U SKUPOVIMA PODATAKA

Multivarijantna analiza

bilješke za predavanja

ak.god. 2003/04

1. Uvod u multivarijatnu statistiku

Let the data speak!

«*The objective of the data analysis is to extract relevant information contained in the data which can then be used to solve a given problem*».

Exploratory data analysis, EDA vs. Hypothesis Testing

⇧

Data mining

Postoje različite klasifikacije multivarijatnih metoda.

Mjerne skale: *nominalna, uređajna, intervalna, racionalna* (zadnje dvije čine metričku skalu)

Broj varijabli: za varijable mjerene na zadnje tri skale broj varijabli je odgovarajući. Za nominalne varijable koje imaju 2 vrijednosti – definira se jedna «dummy» varijabla, (npr. varijabla spol, varijabla poprima vrijednosti: 0 – muški i 1 – ženski). Za nominalnu varijablu s 3 vrijednosti potrebno je formulirati tri varijable.

Neka je dano: *n* entiteta, *p* varijabli
Pretpostavimo podjelu tog skupa u dvije grupe.

DEPENDANCE METHODS – prisutnost ili odsutnost relacije između dva skupa (zavisne i nezavisne) varijable

INTERDEPENDANCE METHODS - ako je nemoguće unaprijed odrediti skup varijabli koje su zavisne i skup varijabli koje su nezavisne nego je potrebno odrediti kako i zašto su varijable međusobno u relaciji

DEPENDANCE METHODS

Dependance methods nadalje dijelimo prema:

- Broju nezavisnih varijabli (jedna ili više)
- Broju zavisnih varijabli (jedna ili više)
- Vrsti mjerne skale zavisne varijable
- Vrsti mjerne skale nezavisne varijable

Jedna zavisna varijabla i jedna nezavisna varijabla
(univarijatna statistika, za razliku od multivarijatne)

Jedna zavisna i više nezavisnih varijabli

Primjer: stručnjak za marketing želi utvrditi vezu između namjere kupnje (NK) nekog proizvoda i niza nezavisnih varijabli: prihoda(P), obrazovanja(O), godine(G), načina života(NŽ) itd.

Linearni model:

$$NK=\beta_0+ \beta_1P + \beta_2O+ \beta_3G +\beta_4NŽ + \varepsilon$$

REGRESIJA
Jedna zavisna i više nezavisnih varijabli – sve mjerene na metričkoj skali.

ANOVA (Analiza varijance)
Nezavisna varijabla mjerena na nominalnoj skali (primjer: umjesto da se bilježi točni prihod, prihod se kategorizira kao visok, srednji, nizak.)
ANOVA je tehnika za procjenu parametara linearnog modela kada su nezavisne varijable nominalne.
ANOVA je posebni slučaj regresije (nezavisne varijable su kategorizirane). U najjednostavnijem slučaju ANOVA se svodi na *t*-test ako nominalna varijabla poprima dvije vrijednosti.
(Primjer: Da li spol utječe na razinu kolesterola u krvi? Da li profesija utječe na razinu kolesterola u krvi? Da li spol i profesija zajedno utječu na razinu kolesterola u krvi?)

DISKRIMINANTNA ANALIZA

Pretpostavimo da namjeru kupnje mjerimo na nominalnoj skali (kupci i oni koji to nisu) dok su nezavisne varijable mjerene na metričkoj skali. Želimo odrediti da li se dvije grupe (kupci i oni koji to nisu) značajno razlikuju s obzirom na nezavisne varijable, i ako da, mogu li nezavisne varijable biti upotrebljene za predviđanje ili klasifikaciju potencijalnih kupaca u jednu od dvije grupe.
2- grupe DA je poseban slučaj multiple regresije.

LOGISTIČKA REGRESIJA

Pretpostavka diskriminantne analize je da podaci dolaze iz multivarijatne normalne distribucije. Logistička regresija se primjenjuje kada su te pretpostavke narušene i kada je zavisna varijabla kombinacija nominalne i metričke varijable.

Više od jedne zavisne i jedna ili više nezavisnih varijabli.

KANONSKA KORELACIJSKA ANALIZA

Je tehnika za analizu relacije između dviju skupova varijabli. U našem primjeru ako nas kao zavisna varijable uz namjeru kupnje prehrambenog proizvoda još interesira i mišljenje kupca o okusu proizvoda.
(Multipla regresija je poseban slučaj CCA)

MDA - DISKRIMINANTNA ANALIZA S VIŠE GRUPA

Pretpostavimo da potencijalne kupce podijelimo u tri grupe. Kako se te tri grupe razlikuju u odnosu na nezavisne varijable? Kako razviti metodu diskriminacije za buduće kupce?

INTERDEPENDANCE METHODS

Nema eksplicitno zadanih skupova zavisnih i nezavisnih varijabli. Potrebno je identificirati kako i zašto su varijable korelirane jedna s drugom.

METODA GLAVNIH KOMPONENATA

- metoda za redukciju podataka. Reducira veliki broj varijabli na mali broj kompozitnih varijabli.

FAKTORSKA ANALIZA

Pokušava identificirati mali broj faktora koji su odgovorni za korelaciju između velikog broja varijabli. FA – tehnika redukcije podataka. Identificira grupe varijabli tako da su korelacije varijabli unutar grupe veće nego one između grupa.
(Primjer školski psiholog pokušava analizirati korelaciju između ocjena različitih kolegija predmeta za učenike u školi)

GRUPIRANJE PODATAKA

Tehnika grupiranja elemenata (objekata, entiteta, opservacija) tako da su elementi unutar jednog klastera slični u odnosu na obilježja (varijable) koje ih opisuju.
Naročito interesantna u bio znanostima za razvijanje taksonomija.
Primjer: grupiranje prehrambenih artikala prema vrijednostima nutrijenata (vitaminima, mineralima, ugljikohidratima…), grupiranje potencijalnih kupaca prema kupovnim navikama.

2. METODA GLAVNIH KOMPONENATA

ili Karhunen-Loève transformacija

ili Hotellingova transformacija

(engl. *Principal Component Analysis - PCA*)

- Karl Pearson 1901. godine prvi opisao PCA
- Hotelling 1933. dao opis izračuna glavnih komponenti
- Primjena za više varijabli tek s razvojem računala

Jedna od najjednostavnijih metoda multivarijatne statistike.

Cilj je načiniti novi koordinatni sustav s manjim brojem dimenzija od izvornog koji naglašava glavne uzorke varijacija podataka

Primjena:

- **redukcija dimenzionalnosti podataka** (reducira broj izvornih varijabli na mali broj indeksa koji su linearna kombinacija izvornih varijabli i koji se zovu glavne komponente)
- **interpretacija podataka** (glavne komponente objašnjavaju varijabilnost podataka na najkoncizniji način, na taj način pokazuje neke skrivene povezanosti, međuodnose podataka. Podaci se prikazuju na način koji nije uobičajen, ali sadrži mnogo bitnih informacija o skupu izvornih podataka)

Cilj metode glavnih komponeneta:

X_1, X_2, ...X_p varijabli (svojstava), mjerenih na n objekata (sva mjerenja se prikazuju n x p matricom),

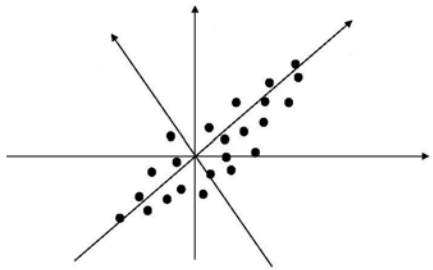
treba naći Y_1, Y_2, ...Y_p tako da su nekorelirani

(odsustvo korelacije – indeksi odražavaju različite «dimenzije» podataka)

i da vrijedi $Var(Y_1) \geq Var(Y_2) \geq \ldots\ldots \geq Var(Y_p)$

Y_i se nazivaju **glavne komponente**

- varijance većine Y_i zanemarivo male -> varijabilnost skupa podataka se može opisati s malim brojem glavnih komponenata Y_i
  - PCA provediva samo ako su izvorne varijable korelirane – najbolje ako su jako korelirane - tada ima redundancije u izvornim varijablama koje mjere istu stvar, na primjer 20-30 varijabli predstavi se sa 2-3 glavne komponente.



Osnovne definicije:

Neka je X slučajni vektor , elementi od X su slučajne varijable.

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \dots \\ X_p \end{bmatrix}$$

Tada je očekivanje slučajnog vektora vektor definiran sa:

$$E(X) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \dots \\ E(X_p) \end{bmatrix},$$

gdje je $E(X_i)$ očekivanje slučajne varijable X_i, označimo ga s μ_i.

Varijanca slučajnog vektora X je

$$Var(X) = \sigma^2 = E\left[(X - E(X))^2\right].$$

Za $i, j = 1,2,\dots p$ definirajmo realne brojeve:

$$c_{ij} = E\left[(X_i - E(X_i))(X_j - E(X_j))\right] = E(X_i X_j) - E(X_i)E(X_j).$$

Za $i \neq j$, c_{ij} zovemo **kovarijanca slučajnih varijabli** X_i, X_j i često je označavamo s $Cov(X_i, X_j)$. Simetričnu matricu Σ definiranu na slijedeći način:

$$\Sigma = \begin{bmatrix} c_{11} & c_{12} & \dots\dots\dots & c_{1p} \\ c_{21} & c_{22} & \dots\dots\dots & c_{2p} \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ c_{p1} & c_{p2} & \dots\dots\dots & c_{pp} \end{bmatrix}$$

nazivamo **kovarijaciona matrica slučajnog vektora** X. Kada je očekivanje slučajnog vektora nula (nul-vektor) tada je kovarijaciona matrica jednaka **autokorelacionoj matrici** slučajnog vektora X koja je definirana sa:

$$R = E(XX^T).$$

GLAVNE KOMPONENTE

Neka je $X = (X_1, \dots, X_p)^T$ slučajni vektor s **kovarijacionom matricom** Σ i neka su njene **svojstvene vrijednosti** dane s $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$.

Pogledajmo linearne kombinacije :

$$\begin{aligned} Y_1 &= l_1^T X = l_{11}X_1 + l_{21}X_2 + \dots + l_{p1}X_p \\ Y_2 &= l_2^T X = l_{12}X_1 + l_{22}X_2 + \dots + l_{p2}X_p \\ &\vdots \\ Y_h &= l_h^T X = l_{1h}X_1 + l_{2h}X_2 + \dots + l_{ph}X_p \\ &\vdots \\ Y_p &= l_p^T X = l_{1p}X_1 + l_{2p}X_2 + \dots + l_{pp}X_p \end{aligned}$$

Glavne komponente su nekorelirane linearne kombinacije $Y_1, Y_2, \dots, Y_p$ čije varijance su najveće moguće.

Linearne kombinacije Y_h, Y_k **su nekorelirane** ako vrijedi $Cov(Y_k, Y_h) = 0$.

Prva glavna komponenta je linarna kombinacija s najvećom varijancom, odnosno ona koja maksimizira izraz $Var(Y_1)$, uz uvjet da vrijedi $l_1^T l_1 = 1$.

Glavne komponente definiramo na slijedeći način:

- **Prva glavna komponenta** je linearna kombinacija $Y_1 = l_1^T X$ koja maksimizira izraz $Var\left(l_1^T X\right)$, uz uvjet $l_1^T l_1 = 1$.
- **Druga glavna komponenta** je linearna kombinacija $Y_2 = l_2^T X$ koja maksimizira izraz $Var\left(l_2^T X\right)$, uz uvjet $l_2^T l_2 = 1$ i $Cov\left(l_1^T X, l_2^T X\right) = 0$.

...

- ***h*-ta glavna komponenta** je linearna kombinacija $Y_h = l_h^T X$ koja maksimizira izraz $Var\left(l_h^T X\right)$, uz uvjet $l_h^T l_h = 1$ i $Cov\left(l_h^T X, l_k^T X\right) = 0$ za $k < h$.

...

- *p*-ta glavna komponenta je linearna kombinacija $Y_p = l_p^T X$ koja maksimizira izraz $Var\left(l_p^T X\right)$, uz uvjet $l_p^T l_p = 1$ i $Cov\left(l_p^T X, l_k^T X\right) = 0$ za $k < p$.

Objašnjenje metode glavnih komponenata

Varijance i kovarijance linearnih kombinacija Y_i (tj. glavnih komponenti) dane su formulama:

$$Var\left(Y_h\right) = l_h^T \Sigma l_h = \lambda_h \text{ za } h = 1,2,\ldots,p$$
$$Cov\left(Y_h, Y_k\right) = l_h^T \Sigma l_k = 0 \text{ za } h,k = 1,2,\ldots,p$$

Kovarijaciona matrica podataka je realna i simetrična tj. vrijedi

$$\Sigma^T = \Sigma,$$

te je pozitivno definitna , odnosno

$$x\Sigma x > 0, \forall x \ .$$

Kovarijaciona matrica je dimenzije *p* x *p* i ima *p* nenegativnih svojstvenih vrijednosti.

Svaka se simetrična matrica može napisati kao produkt svojih svojstvenih vektora i svojstvenih vrijednosti na slijedeći način:

$$\Sigma = \lambda_1 e_1 e_1^T + \lambda_2 e_2 e_2^T + \ldots + \lambda_p e_p e_p^T,$$

odnosno

$$\Sigma = Q^T \Lambda Q,$$

gdje je

Q matrica svojstvenih vektora matrice Σ,

Λ je dijagonalna matrica koja na dijagonali ima svojstvene vrijednosti matrice Σ.

Svojstvene vrijednosti (λ) definirane kao nul-točke jednadžbe

$$\det\left(\lambda \mathbf{I} - \Sigma\right) = 0,$$

a **svojstveni vektori** (e) se dobivaju iz jednadžbe

$$\Sigma \cdot e = \lambda e.$$

Tvrdnja 1. Neka je B pozitivno definitna matrica sa svojstvenim vrijednostima $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ i pripadnim normaliziranim svojstvenim vektorima $e_1, e_2, \ldots, e_p$. Tada je

$$\max_{x \neq 0} \frac{x^T B x}{x^T x} = \lambda_1 \quad \text{(postiže se za } x = e_1\text{) i vrijedi također}$$

$$\max_{x \perp e_1, \ldots e_k} \frac{x^T B x}{x^T x} = \lambda_{k+1} \quad \text{(postiže se za } x = e_{k+1}, k = 1,2,\ldots,p-1\text{).}$$

Tvrdnja 2. Neka je Σ kovarijaciona matrica slučajnog vektora $X = (X_1, \ldots, X_p)^T$ i neka su dani parovi svojstvena vrijednost - svojstveni vektor $(\lambda_1, e_1), (\lambda_2, e_2), \ldots, (\lambda_p, e_p)$ kovarijacione matrice Σ, gdje je $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. (Ako su neke svojstvene vrijednosti λ_h jednake, tada izbor pripadnog svojstvenog vektora e_h i Y_h nije jedinstven.) Označimo koordinate vektora e_h ovako: $e_h = [e_{1h}, e_{2h}, \ldots e_{ph}]^T$. Tada je *h*-ta glavna komponenta dana sa

$$Y_h = e_h^T X = e_{1h} X_1 + e_{2h} X_2 + \ldots + e_{ph} X_p \text{ za } h = 1,2,\ldots,p$$

i vrijedi

$$Var(Y_h) = e_h^T \Sigma e_h = \lambda_h \text{ za } h = 1,2,\ldots,p$$

$$Cov(Y_k, Y_h) = e_h^T \Sigma e_k = 0 \text{ za } h \neq k.$$

Tvrdnja 3. Ukupna varijanca je jednaka

$$tr(\Sigma) = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_p^2 = \sum_{j=1}^{p} Var(X_j)$$

$$= \lambda_1 + \lambda_2 + \ldots + \lambda_p = \sum_{h=1}^{p} Var(Y_h)$$

Primjer. Pretpostavimo da slučajne varijable X_1, X_2, X_3 imaju kovarijacionu matricu:

$$\begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

Može se provjeriti da tada parovi svojstvena vrijednost-svojstveni vektor izgledaju ovako:

$\lambda_1 = 5.83 \quad e_1^T = [0.383, -0.924, 0]$

$\lambda_2 = 2.00 \quad e_2^T = [0,0,1]$

$\lambda_3 = 0.17 \quad e_3^T = [0.924, 0.383, 0]$

Glavne komponente su tada :

$Y_1 = e_1^T X = 0.383 X_1 - 0.924 X_2$

$Y_2 = e_2^T X = X_3$

$Y_3 = e_3^T X = 0.924 X_1 + 0.383 X_2$

Varijanca prve glavne komponente je

$$Var(Y_1) = Var(0.383 X_1 - 0.924 X_2) = 5.83 = \lambda_1,$$

kovarijanca između prve i druge glavne komponente je

$$Cov(Y_1, Y_2) = Cov(0.383 X_1 - 0.924 X_2, X_3) = 0.$$

Sada računamo trag:

$$\sigma_1^2 + \sigma_2^2 + \sigma_3^2 = 1 + 5 + 2 = \lambda_1 + \lambda_2 + \lambda_3 = 5.83 + 2.00 + 0.17 = 8$$

Prve dvije komponente sudjeluju s udjelom $\frac{(5.83 + 2)}{8} = 0.98$ od ukupne varijance. U ovom slučaju je jasno da bi komponente Y_1, Y_2 mogle dobro zamjeniti tri originalne varijable s vrlo malo gubitaka informacije.

Geometrijska interpretacija metode glavnih komponenata

Želimo li vidjeti što bi bile glavne komponente nekog konkretnog skupa uzoraka moramo definirati neke pojmove *deskriptivne statistike*. Neka je $X = \{x_1, x_2, \ldots x_n\}$ neki skup uzoraka, tada je *srednja vrijednost* dana s

$$\mu = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i.$$

Uzorci mogu biti višedimenzionalni podaci, odnosno svaki uzorak x_i može biti *p*-dimenzionalni vektor

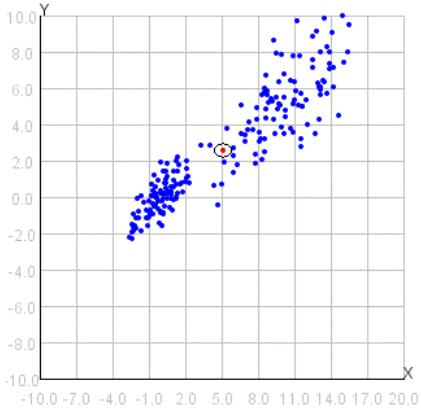
$$\mathbf{x}_i = \begin{bmatrix} x_i^1 \\ x_i^2 \\ \ldots \\ x_i^p \end{bmatrix}.$$

Tada μ *vektor srednjih vrijednosti* definiramo kao:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_i^1 \\ \frac{1}{n}\sum_{i=1}^{n} x_i^2 \\ \ldots \\ \frac{1}{n}\sum_{i=1}^{n} x_i^p \end{bmatrix} = \begin{bmatrix} \mu^1 \\ \mu^2 \\ \ldots \\ \mu^p \end{bmatrix}.$$

Ako su podaci dvodimenzionalni, tada je vektor srednjih vrijednosti

PRIKAZ VEKTORA SREDNJIH VRIJEDNOSTI



Varijanca skupa uzoraka S je *p*-dimenzionalni vektor dan izrazom:

$$\mathbf{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^2 = \begin{bmatrix} \frac{1}{n-1}\sum_{i=1}^{n}(x_i^1 - \mu^1)^2 \\ \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 - \mu^2)^2 \\ \ldots \\ \frac{1}{n-1}\sum_{i=1}^{n}(x_i^p - \mu^p)^2 \end{bmatrix}.$$

Komponente ovog vektora mjere raširenost (*spread*) skupa uzoraka duž svih *p* osi koje razapinju *p*-dimenzionalni prostor.

Različite komponente uzoraka mogu biti međusobno u korelaciji, npr. vrijednost varijable x^a raste kada raste vrijednost varijable x^b . Ovo svojstvo je sadržano u kovarijanci cov_{ab} od x^a i x^b definiranoj kao:

$$\mathrm{cov}_{ab} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i^a - \mu^a)(x_i^b - \mu^b).$$

Matrica $\mathbf{C}$ dimenzije *p x p* dana sa $\mathbf{C} = [\mathrm{cov}_{ab}]_{a,b=1,\ldots,n}$, odnosno

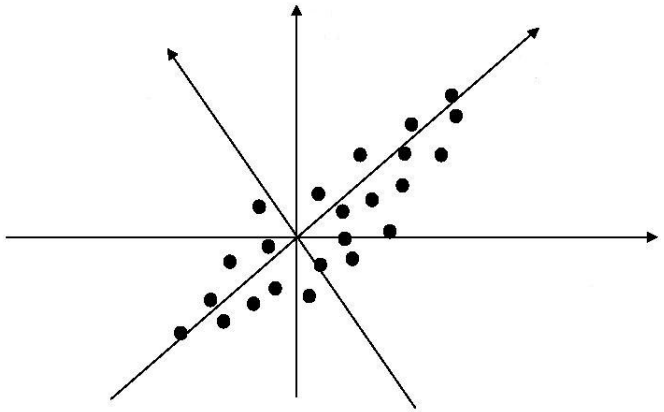
$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \ldots\ldots & c_{1p} \\ c_{21} & c_{22} & \ldots\ldots & c_{2p} \\ \ldots\ldots & & & \\ c_{p1} & c_{p2} & \ldots\ldots & c_{pp} \end{bmatrix} = \begin{bmatrix} \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^1-\mu^1\right)^2 & \cdots & \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^1-\mu^1\right)\left(x_i^p-\mu^p\right) \\ \vdots & \ddots & \vdots \\ \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^1-\mu^1\right)\left(x_i^p-\mu^p\right) & \cdots & \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^p-\mu^p\right)^2 \end{bmatrix}$$

Naziva se ***kovarijaciona matrica uzoraka***. Kovarijaciona matrica za skup uzoraka koji ima vektor srednjih vrijednosti nula postaje *autokorelaciona matrica* definirana ovako:

$$R = \begin{bmatrix} \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^1\right)^2 & \cdots & \frac{1}{n-1}\sum_{i=1}^{n}x_i^1x_i^p \\ \vdots & \ddots & \vdots \\ \frac{1}{n-1}\sum_{i=1}^{n}x_i^1x_i^p & \cdots & \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^p\right)^2 \end{bmatrix}$$

(Napomene: Nazivi SS i SSCP za *sum of squares* i *cross product;* $R = XX^T$)
Geometrijski gledano, metoda glavnih komponenata je izbor novog koordinatnog sustava dobivenog ortogonalnom transformacijom originalnog sustava

GEOMETRIJSKA INTERPRETACIJA GLAVNIH KOMPONENATA



- **Prva glavna komponenta** je smjer duž kojeg je varijanca podataka najveća. **Druga glavna komponenta** je smjer maksimalne varijance podataka u prostoru okomitom na prvu glavnu komponentu.
- **Novi koordinatni sustav** razapinju pripadni svojstveni vektori najvećih svojstvenih vrijednosti kovarijacione matrice skupa podataka.

Redukcija dimenzionalnosti metodom glavnih komponenata

Primjer slike u boji.
Metoda glavnih komponenata -> za redukciju dimenzionalnosti podataka, (uz što manje bitnih gubitaka).
Izvorni, p-dim podaci se projekcijom prevode u k-dim pri čemu vrijedi,

$$k < p$$

Ideja: napraviti projekciju tih *n* uzoraka iz p-dim prostora N u *k* -dim potprostor *M,* ali tako da ti projicirani uzorci budu što sličniji originalnim uzorcima.

Projekcija uzoraka iz prostora *N* u potprostor *M* dobija se množenjem uzorka transponiranom matricom matrice **V**, (*p x k matrica)* čiji stupci predstavljaju bazu potprostora *M* izraženu preko baze *N izvornog prostora.* Odnosno,

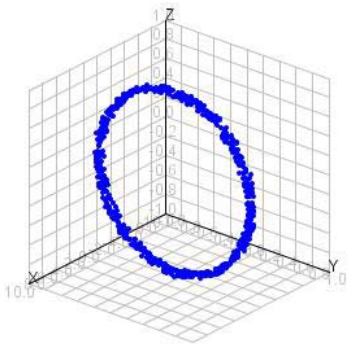
$$V^T \cdot x_i = y_i ,$$

gdje je x_i uzorak u prostoru *N,* a y_i uzorak u prostoru *M.*

Potprostor u koji se vrši projekcija treba biti tako odabran da je pogreška rekonstrukcije najmanja moguća, tj. da se projekcijom izgubi što je manje moguće informacije o izvornom podatku.

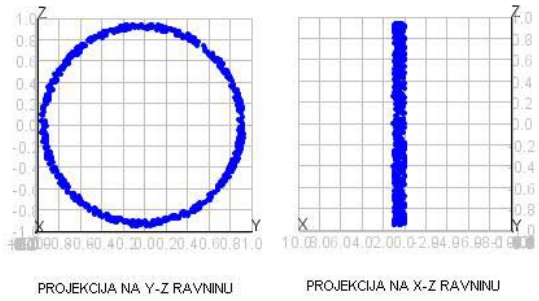
Primjer. Dan je skup točaka u 3-dim prostoru. Tražimo 2-dim prikaz podataka koji što vjernije opisuje originalni skup podataka.

PRIKAZ PODATAKA U ORIGINALNOM PROSTORU



Projekcija na Y-Z ravninu, vjernije čuva izvorne podatke nego projekcija na X-Z.

PROJEKCIJA PODATAKA NA RAVNINU



PROJEKCIJA NA Y-Z RAVNINU PROJEKCIJA NA X-Z RAVNINU

PCA najbolje oodređuje potprostor koji čuva najviše informacija!

Neka je dan p-dimenzionalni prostor uzoraka i *X* skup n-uzoraka iz tog prostora. Vektor srednjih vrijednosti uzoraka dan je izrazom:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

Ako srednja vrijednost skupa uzoraka nije nula, tada svakom uzorku iz S oduzmemo vektor srednjih vrijednosti, odnosno

$$\mathbf{t}_i = (\mathbf{x}_i - \boldsymbol{\mu}) .$$

Tada će taj dobiveni skup uzoraka T imati vektor srednjih vrijednosti nula.

Da bi odredili potprostor M u koji će se projicirati skup uzoraka T potrebno je odrediti kovarijacionu matricu skupa uzoraka T, te njene svojstvene vrijednosti i jedinične svojstvene vektore.
Kovarijaciona matrica računa se prema formuli:

$$\mathbf{C}(i,j) = \frac{1}{n-1}\sum_{l=1}^{n}\left(x_{i,l} - \mu_i\right)\cdot\left(x_{j,l} - \mu_j\right)^T = \frac{1}{n-1} t_i t_j .$$

Svojstvene vrijednosti (λ) definirane su kao nul-točke jednadžbe,

$$\det(\lambda\mathbf{I} - \mathbf{C}) = 0 .$$

Broj svojstvenih vrijednosti kovarijacione matrice C (dim *p* x *p*) je *p*.

Svojstveni vektori e i svojstvene vrijednosti λ su povezani jednadžbom,

$$\mathbf{C} \cdot e = \lambda e .$$

- Svakom svojstvenom vektoru odgovara jedna svojstvena vrijednost,
- Jednoj svojstvenoj vrijednosti može odgovarati beskonačno mnogo svojstvenih vektora, (međusobno kolinearni)
- Svakoj svojstvenoj vrijednosti pripada samo jedan jedinični svojstveni vektor.
- Svojstveni vektori koji pripadaju različitim svojstvenim vrijednostima međusobno su ortogonalni.

Baza k-dimenzionalnog potprostora M **određena je pomoću 'vodećih' k jediničnih svojstvenih vektora kovarijacione matrice** $\mathbf{C}$ **(inače ih ima p!).**
Pod pojmom 'vodeći' jedinični svojstveni vektori podrazumijevaju se jedinični svojstveni vektori koji pripadaju najvećim svojstvenim vrijednostima.

$$B_M = \{\mathbf{e_1}, \mathbf{e_2}, ..., \mathbf{e}_k\}$$
$$|\lambda(\mathbf{e_1})| > |\lambda(\mathbf{e_i})|, \ (\forall i) \ (1 < i <= p)$$
$$|\lambda(\mathbf{e}_2)| > |\lambda(\mathbf{e_i})|, \ (\forall i) \ (2 < i <= p)$$
$$.....$$
$$|\lambda(\mathbf{e}_k)| > |\lambda(\mathbf{e_i})|, \ (\forall i) \ (k < i <= p) \ ,$$

gdje je B_M baza vektorskog potprostora M, $\mathbf{e_i}$ su jedinični svojstveni vektori, a $\lambda(\mathbf{e_i})$ su svojstvene vrijednosti koje pripadaju jediničnim svojstvenim vektorima.

Stupci matrice *V* (*p* x *k* matrica) sadržavat će vektore iz B_M, to je **zapis k svojstvenih vektora u terminima p originalnih varijabli.**

$$V = \begin{bmatrix} e_1^1 & e_2^1 & ..... & e_k^1 \\ e_1^2 & e_2^2 & ..... & e_k^2 \\ ..... & & & \\ e_1^p & e_2^p & ..... & e_k^p \end{bmatrix}.$$

Sada ovu matricu ***V*** koristimo za proiciranje podataka iz prostora *N* u prostor *M*.
Neka je sada $\mathbf{x}_i$ neki uzorak iz prostora *N*, tada je njegova projekcija y_i:

$$y_i = \mathbf{V}^T \cdot x_i = \begin{bmatrix} e_1^1 & e_1^2 & ..... & e_1^p \\ e_2^1 & e_2^2 & ..... & e_2^p \\ ..... & & & \\ e_k^1 & e_k^2 & ..... & e_k^p \end{bmatrix} \begin{bmatrix} x_i^1 \\ x_i^2 \\ ... \\ x_i^p \end{bmatrix} = \begin{bmatrix} y_i^1 \\ y_i^2 \\ ... \\ y_i^k \end{bmatrix}.$$

Sada je y_i^k ***k*- ta glavna komponenta**.

Dobili smo *p*-dimenzionalan vektor $\mathbf{x}_i$ zapisan kao *k*-dimenzionalan **vektor glavnih komponenti** $\mathbf{y}_i$ (k < p).

Sada tu projekciju primjenimo na sve elemente skupa *uzoraka*. Ovo proiciranje podataka je sada na neki način kompresija skupa uzoraka.

Matrica **U (dim n x k)**, reci su zapisi n uzoraka izvornog prostora N u k-dim potprostoru M.

nove koordinate = matrica transformacije x uzorci izraženi su starim koordinatama

$$\mathbf{U}^T_{(k \times n)} = \mathbf{V}^T_{(k \times p)} \mathbf{X}^T_{(p \times n)}$$

Rekonstrukcija podataka i pripadna pogreška

U slučaju da metodu glavnih komponenata želimo koristiti za kompresiju podataka ili za slanje podataka kanalima nedostatne širine (manje od dimenzije podataka), tada će nas zanimati i rekonstrukcija podataka nakon slanja (kompresije) i greška koja pri tome nastaje.

Formula za rekonstrukciju uzorka x_i iz vektora glavnih komponenata je:

$$x_i^{'} = V \cdot y_i = \begin{bmatrix} \mathbf{e}_1^1 & \mathbf{e}_2^1 & \ldots\ldots\ldots & \mathbf{e}_k^1 \\ \mathbf{e}_1^2 & \mathbf{e}_2^2 & \ldots\ldots\ldots & \mathbf{e}_k^2 \\ \ldots\ldots\ldots & & & \\ \mathbf{e}_1^p & \mathbf{e}_2^p & \ldots\ldots\ldots & \mathbf{e}_k^p \end{bmatrix} \begin{bmatrix} y_i^1 \\ y_i^2 \\ \ldots \\ y_i^k \end{bmatrix}.$$

$$\mathbf{X^T}_{(p \times n)} = \mathbf{V}_{(p \times k)}\, \mathbf{U^T}_{(k \times n)}$$

Uslijed gubitka informacije koji je uzrokovan projekcijom, javlja se pogreška rekonstrukcije (udaljenost između uzoraka), a njen kvadrat je točno jednak sumi svih svojstvenih vrijednosti koje su odbačene:

$$\varepsilon = \left\| \mathbf{x}_i - \mathbf{x}_i^{'} \right\| = \left\| \mathbf{x}_i - \mathbf{V^T V} \cdot \mathbf{x}_i \right\| = \sum_{i=k+1}^{p} \lambda_i \, .$$

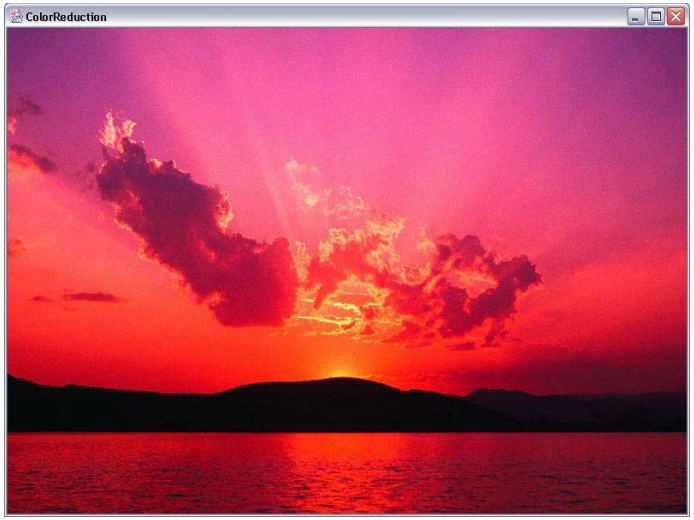
Srednja kvadratna pogreška rekonstrukcije svih uzoraka iz skupa *S* je:

$$\bar{\varepsilon} = \frac{1}{n} \sum_{i=1}^{n} \left\| \mathbf{x}_i - \mathbf{V^T V} \cdot \mathbf{x}_i \right\| .$$

Primjer primjene metode glavnih komponenata redukciju dimenzionalnosti podataka u obradi slike

- Boja u RGB zapisu je predstavljena kao vektor u trodimenzionalnom prostoru čiju bazu čine vektori R, G i B koji odgovaraju *crvenoj, plavoj i zelenoj* boji. Dakle, svaki slikovni element (engl. *pixel)* je jedan vektor u prostoru koji razapinju vektori R, G i B.
- Slika je skup 3-dimenzionalnih podataka.

ORIGINALNA SLIKA U BOJI

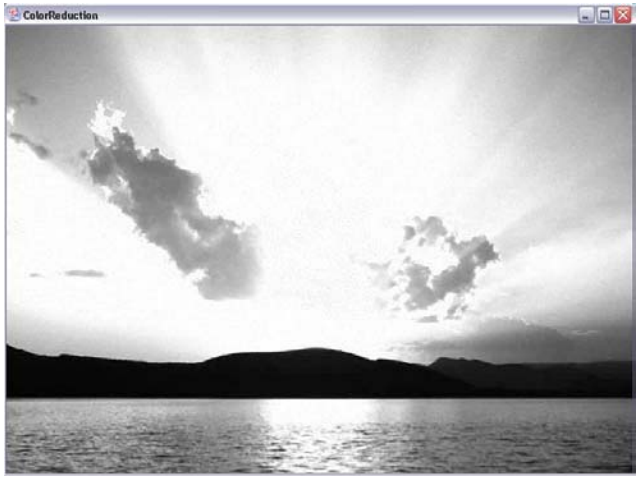


Pretvorbu slike u boji u crno bijelu sliku, odnosno u nijanse sive, možemo gledati kao projekciju elemenata skupa iz 3-dimenzionalnog (R, G i B) prostora u 1-dimenzionalan prostor.

- PCA određuje smjer u kojem će projekcija imati najveću varijancu, odnosno crno-bijela projekcija slike će zadržati najviše informacija o boji. (nije najbolji način pretvorbe!)
- smjer prve glavne komponente je vektor u prostoru RGB - boja čijih različitih nijansi na slici ima najviše.

Primjer - na slici koja većinom ima nijanse crvene boje, bolje će izgledati projekcija na os R (crvena), nego projekcija na G (zelena) ili B (plava).

PROJEKCIJA NA OS R (CRVENA)



PROJEKCIJA NA OS B (PLAVA)

PROJEKCIJA U SMJERU PRVE GLAVNE KOMPONENTE

Projekcija u smjeru prve glavne komponente uvijek daje najvjerniju crno-bijelu sliku.

Literatura:

Johnson, R. A.; Wichern, D. W.: *Applied Multivariate Statistical Analysis*, Prentice Hall; 5th edition, 2002.

Poljak, T., Metoda glavnih komponenata, diplomski rad, Matematički odjel Prirodoslovno-matematičkog fakulteta, 2003.

3. GRUPIRANJE PODATAKA

(*engl. CLUSTER ANALIZA*)

engl. Taxonomy analysis

Cilj: Pridružiti objekte u grupe na temelju *sličnosti* objekata.

Sličnost je predefinirani kriterij koji se računa iz opažanja (mjerenja) na objektima.

Pitanja:

- Koju mjeru sličnosti ili različitosti (engl. similarity, dissimilarity) koristiti ?
- Koji algoritam grupiranja koristiti?

Za grupiranje objekata – metrika, za grupiranje varijabli – korelacijski koeficijenti

Mjera udaljenosti (engl. *dissimilarity measure*) je mjera različitosti podataka

Mjera udaljenosti ili metrika d je funkcija sa X x X u R koja zadovoljava uvjete:

- $D(x_k, x_l) \geq 0$, za $x_k = x_l$, $D(x_k, x_l) = 0$ (pozitivna definitnost)
- $D(x_k, x_l) = D(x_l, x_k)$ (simetričnost)
- $D(x_k, x_l) \leq D(x_k, x_j) + D(x_j, x_l)$ (pravilo trokuta)

Metrika:

- L_2, **Euklidska**, $D(x_k, x_j) = || x_k - x_j || = (\sum_i (x_{ki} - x_{ji})^2)^{1/2}$ specijalni slučaj metrika **Minkowski** za r = 2

$$D(x_k, x_j) = (\sum_i | x_{ki} - x_{ji}|^r)^{1/r}$$

(primjer: skup točaka u 2-dim prostoru koji je od neke čvrste točke, središta, udaljen za odabranu konstantnu vrijednost *r* je kružnica)

- L_1, **Manhattan ili Cityblock** specijalni slučaj metrika Minkowski za r = 1

(primjer: skup točaka u 2-dim prostoru koji je od neke čvrste točke, središta, udaljen za odabranu konstantnu vrijednost *r* je «dijamant»)

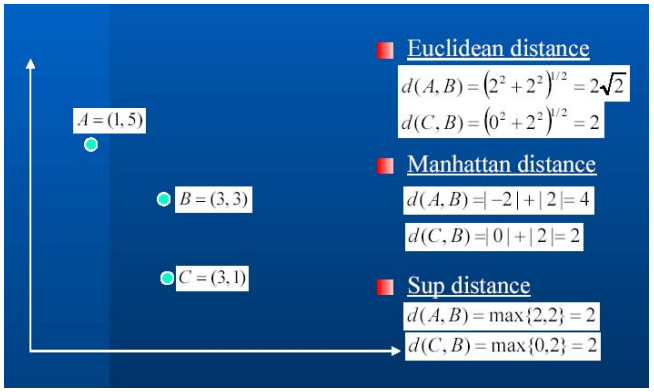
Za binarne vektora L_1 je Hammingova udaljenost

- L_∞, za $r \to \infty$ formula se naziva **Čebiševljeva** udaljenost:

$$D(x_k, x_l) = Max_{1 \leq j \leq N} \{ | x_{kj} - x_{ij}| \}$$

(primjer: skup točaka u 2-dim prostoru koji je od neke čvrste točke, središta, udaljen za odabranu konstantnu vrijednost *r* je kvadrat)

Primjer:



Statistička udaljenost:

Mahalonobisova udaljenost (1948.g.)

$$d(x,y) = \text{sqrt}\ [\ (x-y)'\Sigma^{-1}(x-y)],$$

gdje je Σ^{-1} inverz matrice varijanci-kovarijanci.

Ta je **udaljenost pozitivno definitna kvadratna forma** oblika x'Ax, sdje je A= Σ^{-1} i poopćenje je euklidske udaljenosti ako varijable imaju različite standardne devijacije i korelirane su!
Na primjer ako se Mahalanobisova udaljenost koristi za računanje udaljenosti jedne multivarijatne opservacije od centra populacije:

$$D^2 = \sum_{i=1}^{p}\sum_{j=1}^{p}(x_i - \overline{x_i})v_{ij}(x_j - \overline{x}_j)$$

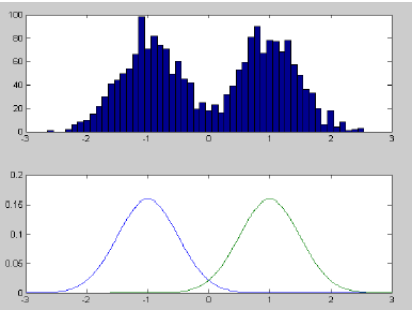
gdje su (x_1, x_2, … , x_p) vrijednosti varijabli X_1, X_2, … , X_p, a v_{ij} je element u i-tom retku i j-tom stupcu inverzne matrice varijanci kovarijanci.

(Primjer: skup točaka u 2-dim prostoru koji je od neke čvrste točke, središta, udaljen za odabranu konstantnu vrijednost *r* je elipsa)

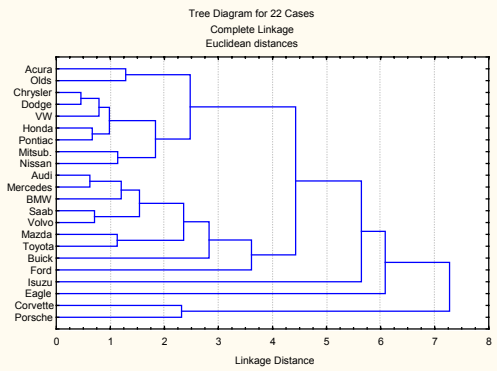
Vrste grupiranja:

Particijska **Hijerarhijska**

Parametarska



Hijerarhijska grupiranja – rezultat grupiranja DENDOGRAM



- Aglomerativna (*bottom-up*) (počinju individualnim objektom, inicijalno n objekat – n grupa, najsličniji objekti se grupiraju, grupe se stapaju u skladu s odabranim kriterijem)
- Divizivna (*top-down*) (rade suprotno, inicijalno svih n podataka je jedna grupa, koja se dijeli na podgrupe, podgrupe se dijele dalje u skladu s odabranim kriterijem)

Particijska grupiranja – nisu hijerarhijske (*engl. flat*)

- K srednjih vrijednosti , (k –means)
- SOM

Parametarski model

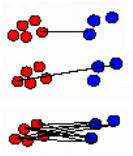
- EM algoritam

Aglomerativna hijerarhijska grupiranja

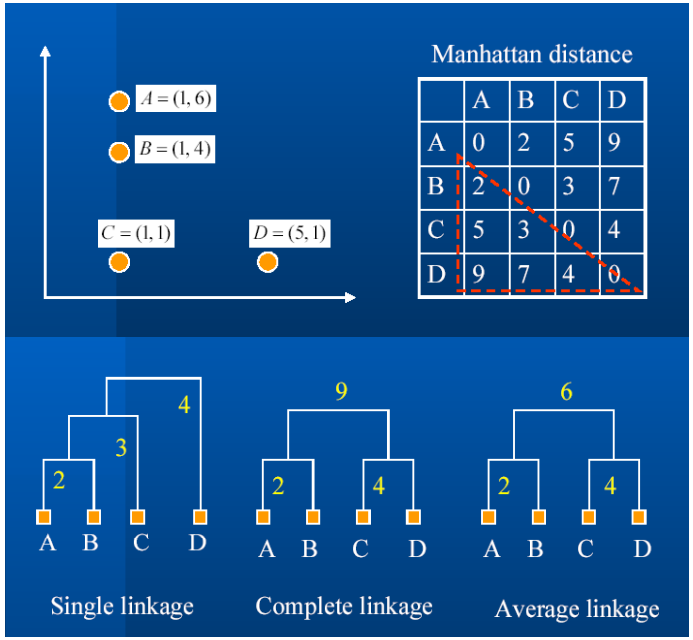
Metode povezivanja (linkage methods)

- pogodne za varijable i objekte

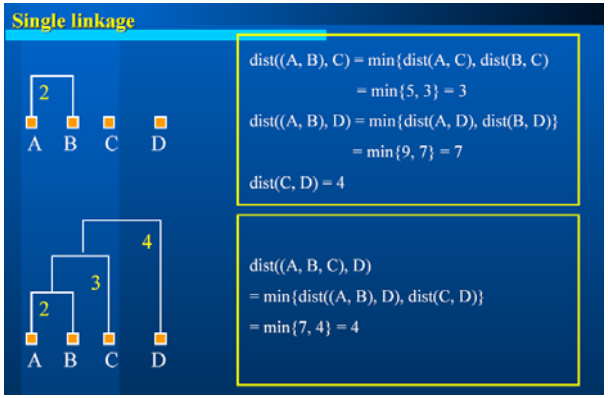
- single linkage
- complete linkage
- average linkage



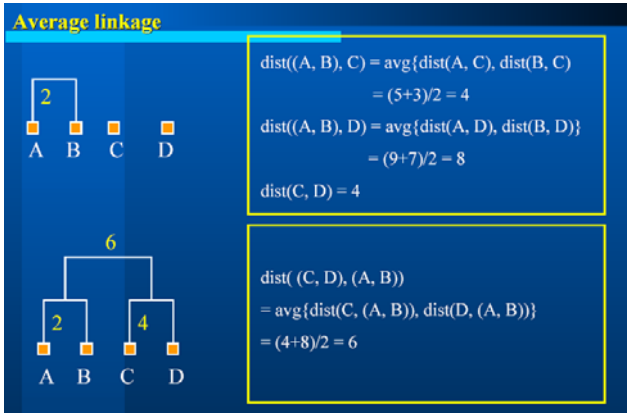
Primjer: Grupiranje 4 podataka u 2-dim prostoru



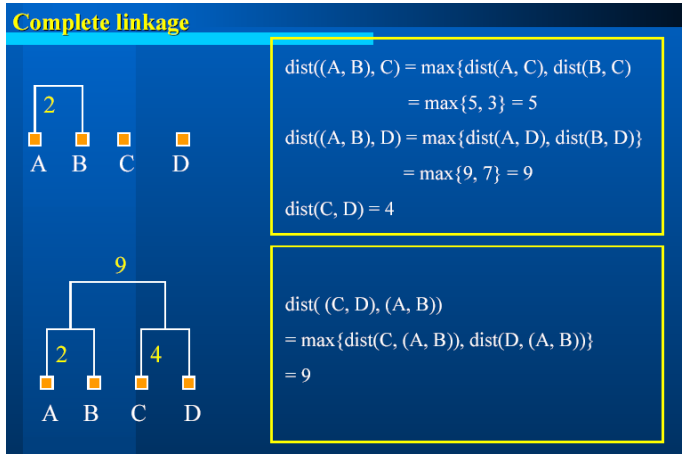
Single linkage – **Povezivanje na temelju minimalne udaljenosti**
ili povezivanje najbližeg susjeda
Podaci u proceduri mogu biti udaljenosti ili sličnosti između objekata. Najbliži susjed određuje najmanju udaljenost ili najveću sličnost između podataka.
Zbog načina spajanja ne može razlikovati slabo odjeljive grupe, ali može odijeliti ne-elipsoidalne grupe.
Ima tendenciju stvaranja duljih lanaca na čijim se krajevima jedinke mogu bitno razlikovati.



Average Linkage – Povezivanje na temelju srednje udaljenosti između grupa. Udaljenost je srednja vrijednost udaljenosti svih parova u grupama.



Povezivanje na temelju maksimalne udaljenosti – udaljenost između dvije grupe (elementa) je određena najvećom udaljenošću.
Osigurava da su svi objekti u grupi unutar neke maksimalne udaljenosti.



Uočava se sličnost dendograma *complete linkage* i *average linkage*, ali se povezivanje dešava na različitim razinama udaljenosti.

Ulaz u postupak povezivanja može biti i korelacijske matrica. Sličnost između dviju varijabli mjeri se produkt-moment korelacijskim koeficijentom. Varijable s velikim negativnim korel. koef. smatraju se jako udaljenima, a one s većim pozitivnim smatraju se bliskima.

Zaključci:
- hijerarhijske aglomerativne metode su osjetljive na *outliere*
- nema mogućnosti preispitivanja već pridjeljenih (krivo) objekata grupama
- dobro je probati više metoda i više mjera udaljenosti te provjeriti konzistentnost rješenja
- stabilnost grupiranja može se provjeriti dodavanjem perturbacija. Ako su grupe jasno odjeljive grupiranje prije i poslije perturbacija se trebaju slagati

Particijske metoda: Algoritam k srednjih vrijednosti – najpoznatiji

ALGORITAM k – SREDNJIH VRIJEDNOSTI

Odnosi se na particiju objekata, a ne varijabli.
Ne koristi matricu sličnosti pa je zahvalnija metoda za veći skup podataka.

Ukratko:

1. odabere se k početnih centara grupa
2. sve se vrijednosti rasporede u k grupa po pravilu minimalne udaljenosti
3. računa se novih k centroida
4. ponavljaj korake 2 i 3 dok više nama promjena

Algoritam k - srednjih vrijednosti (engl. *k – means algoritam*) je postupak grupiranja na temelju **minimizacije kriterijske funkcije**:

$$J = \Sigma_{j=1}^{N_c} J_j, \quad \text{pri čemu je} \quad J_j = \Sigma_{x \in S_j} || x - Z_j ||^2.$$

N_c predstavlja broj elemenata od k grupa, dok S_j predstavlja skup uzoraka čiji je centar Z_j.

Cilj algoritma je naći k središta grupa $Z_1, Z_2, ..., Z_k$ za N početnih neraspodjeljenih uzoraka. Broj k se zadaje na početku, zajedno sa uzorcima, i za njega vrijedi:

$$0 < k < N.$$

Specifičnost algoritma je ta da ovisi o redoslijedu uzimanja uzoraka.

Algoritam:

1. izabiremo k središta grupa Z_1 (1), Z_2 (1), ..., Z_k (1). Metoda izbora početnih središta grupa je proizvoljna. Postoji nekoliko tipova uobičajenih izbora pa prema tome i nekoliko tipova algoritma k – srednjih vrijednosti.

2. u m – tom koraku (iteraciji) razdjeljujemo uzorke $x_1, x_2, ..., x_N$ u k grupa pomoću relacije:

   $x \in S_j$ (m) ako je $|| x - Z_j (m) || < || x - Z_i (m) ||$, i = 1, 2, .., N; $i \neq j$.

   S_j (m) predstavlja skup uzoraka u m – tom koraku čiji je centar Z_j.

3. izračunavamo nova središta grupa Z_j (m+1), j = 1, 2, ..., k tako da je kriterijska funkcija

   $J = \Sigma_{j=1}^{k} \Sigma_{x \in Sj(m)} || x - Z_j (m+1) ||^2$ minimalna.

   Središta grupa koja minimiziraju kriterijsku funkciju u m – toj iteraciji su aritmetičke srednje vrijednosti uzoraka pojedinih grupa

   $Z_j (m+1) = 1/N_j (\Sigma_{x=Sj(m)} x)$ za j = 1, 2, ..., k; N_j je broj uzoraka u grupi.

4. ako je $Z_j (m+1) = Z_j (m)$ za sve j =1, 2, ..., k, postupak završava. Ukoliko taj uvjet nije ispunjen, ponavljamo postupak od koraka 2.

Na rezultat grupiranja pomoću algoritma k – srednjih vrijednosti utječe:

- broj grupa
- izbor početnih središta grupa

Algoritam zahtjeva eksperimentiranje sa različitim vrijednostima k i različitim početnim konfiguracijama centara.

Nema općenitog dokaza o konvergenciji algoritma.

Metoda glavnih komponenata i grupiranje

Može se raditi PCA prije grupiranja kako bi se reducirao veliki broj varijabli i time smanjilo ukupno računanje. Rezultati se sa i bez predprocesiranja s PCA mogu razlikovati!

Literatura:

Hartigan, J.A., Clustering Algorithms, John Wiley & Sons, 1975.