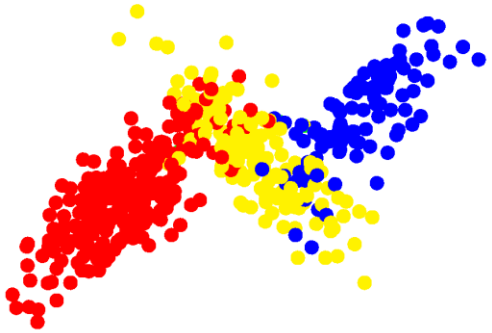
Klasterizacija

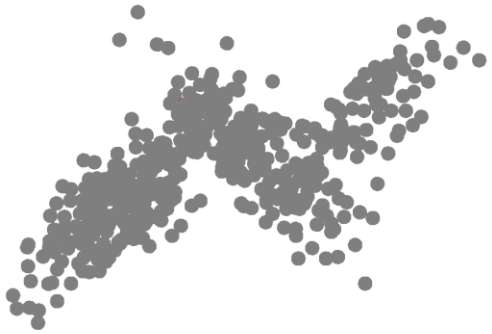
- Klasterizacija uzoraka
  - Mere sličnosti
  - Kriterijumske funkcije
  - Algoritmi nehijerarhijske klasterizacije
    - *k* srednjih vrednosti (eng. *k*-means)
    - ISODATA
  - Algoritmi hijerarhijske klasterizacije
    - Algoritmi zasnovani na podeli (eng. *top-down*)
    - Algoritmi zasnovani na povezivanju (eng. *bottom-up*)

Nenadgledano učenje

- Za razliku od nadgledanog učenja, kod nenadgledanog učenja uzorci iz skupa za obuku nemaju izlaznu vrednost (npr. oznaku pripadnosti klasi u slučaju klasifikacije)
- Parametarske metode
  - Ekvivalentne estimaciji gustine raspodele verovatnoće kao smeše Gaussovih komponenata
  - Kod *expectation maximization* (EM) algoritma, identitet komponente od koje je nastao svaki od uzoraka tretira se kao nedostajuća oznaka klase
- Neparametarske metode
  - Ove metode ne razmatraju gustine raspodele verovatnoće eksplicitno
  - Akcenat je na pronalaženju prirodnih grupacija (klastera) medju uzorcima



nadgledano učenje

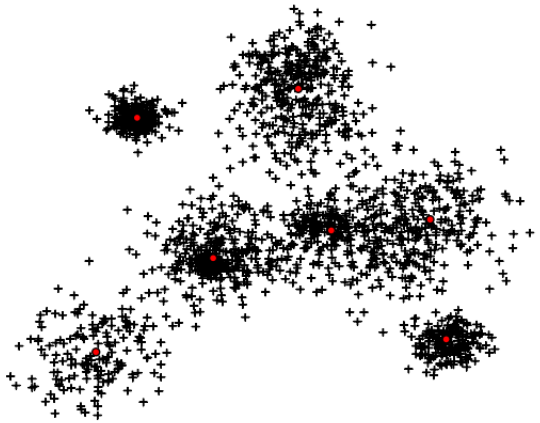


nenadgledano učenje

Neparametarska klasterizacija

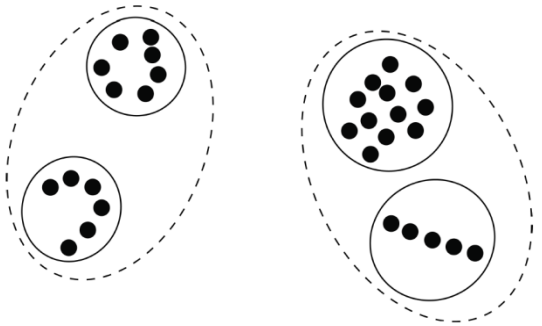
- Klasterizacija je jedna od osnovnih mentalnih aktivnosti čoveka, koja nam pomaže da se izborimo sa velikom količinom podataka
  - Procesiranje svake informacije kao posebnog entiteta bilo bi nemoguće
  - Svaki klaster opisan je uobičajenim karakteristikama entiteta koje poseduje
  - Ako za neki entitet znamo da pripada klasteru, možemo pretpostaviti da ima osobine koje imaju i ostali entiteti iz istog klastera

- Postoji više razloga za klasterizaciju podataka
  - Redukcija količine podataka (kompresija)
    - Daleko je efikasnije obrađivati klastere kao posebne entitete nego uzorke, jer je klastera mnogo manje
  - Generisanje ili testiranje određene hipoteze o prirodi podataka
  - Predikcija na osnovu grupisanja



Neparametarska klasterizacija

- Neparametarska klasterizacija obuhvata tri koraka:
  - Definisanje mera sličnosti/razlike među uzorcima
  - Definisanje kriterijumske funkcije za klasterizaciju
    - Predstavlja meru kvaliteta određene klasterizacije
  - Definisanje algoritma za minimizaciju/maksimizaciju kriterijumske funkcije
    - Čest pristup je *iterativna optimizacija* – polazeći od određene početne particije premeštati uzorke iz jednog klastera u drugi tako da se vrednost kriterijumske funkcije optimizuje
- Različit izbor mere bliskosti, kriterijumske funkcije i algoritma za minimizaciju/ maksimizaciju kriterijumske funkcije može dovesti do bitno različitih rezultata
- Koliko uopšte ima „razumnih“ klastera na slici?
  - U klaster analizi uvek postoji određena doza subjektivnosti, pa je interpretaciju rezultata uvek najbolje prepustiti ekspertu
  - To je sasvim suprotno nadgledanom učenju, gde je funkcija cilja jasna (Bayesov rizik)



Mere bliskosti

- Funkcija $d(\mathbf{x},\mathbf{y})$ koja predstavlja rastojanje između vektora **x** i **y** naziva se *normom* (i obeležava sa $\|\mathbf{x}-\mathbf{y}\|$) ako važi sledeće:

$$
\begin{aligned}
&d(\mathbf{x},\mathbf{y}) \geq 0 \\
&d(\mathbf{x},\mathbf{y}) = 0 \text{ akko je } \mathbf{x} = \mathbf{y} \\
&d(\mathbf{x},\mathbf{y}) \leq d(\mathbf{x},\mathbf{z}) + d(\mathbf{z},\mathbf{y}) \\
&d(a\mathbf{x},a\mathbf{y}) = |a|\, d(\mathbf{x},\mathbf{y})
\end{aligned}
$$

- Često korišćen opšti oblik norme je ***L_p*-norma** (norma Minkowskog)

$$\|\mathbf{x}-\mathbf{y}\|_p = \left(\sum_{i=1}^{D} |x_i - y_i|^p \right)^{1/p}$$

pri čemu izbor parametra *p* utiče na to koliko se značaja pridaje većim razlikama po pojedinim dimenzijama

Mere bliskosti

- Najčešće korišćene mere rastojanja su posebni slučajevi L_p-norme:
  - **L_1-norma** (*Manhattan* ili *city-block* rastojanje)

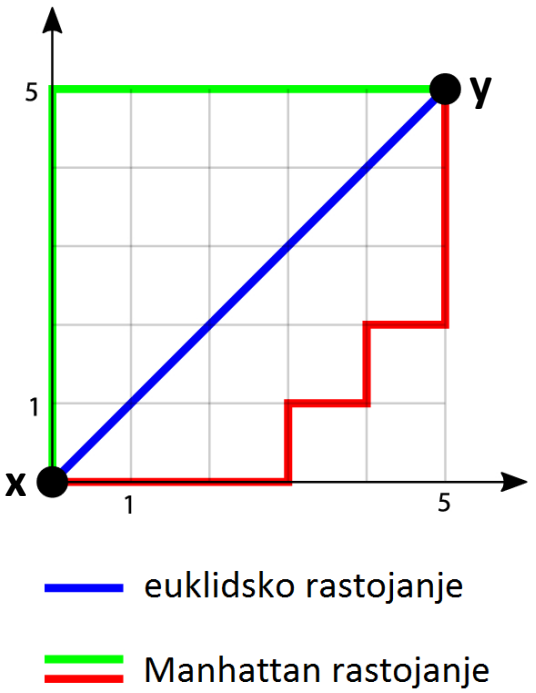
$$\|\mathbf{x}-\mathbf{y}\|_{CB}=\sum_{i=1}^{D}|x_i-y_i|$$

    - Kada se odnosi na binarne vektore, L_1-norma predstavlja Hammingovo rastojanje
  - **L_2-norma** (euklidsko rastojanje)

$$\|\mathbf{x}-\mathbf{y}\|_2=\sqrt{\sum_{i=1}^{D}(x_i-y_i)^2}$$

  - **L_∞-norma** (Čebiševljevo rastojanje)

$$\|\mathbf{x}-\mathbf{y}\|_\infty=\max_{1\le i\le D}|x_i-y_i|$$



Mere bliskosti

- Najčešće korišćene mere rastojanja su posebni slučajevi L_p-norme:
  - **L_1-norma** (*Manhattan* ili *city-block* rastojanje)

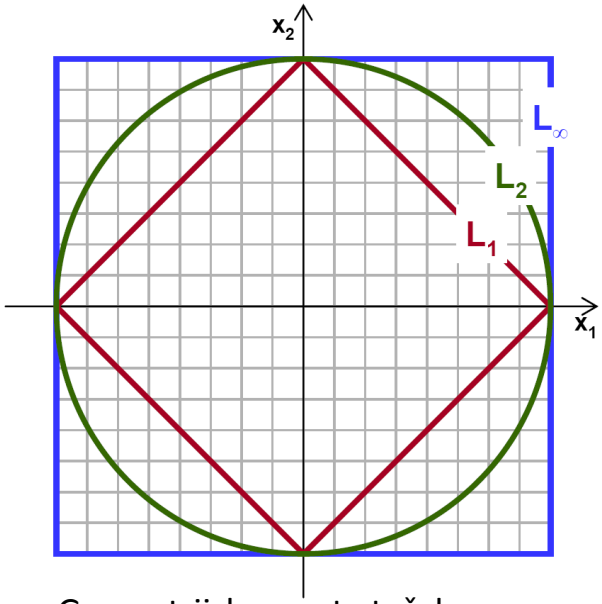
$$\|\mathbf{x}-\mathbf{y}\|_{CB}=\sum_{i=1}^{D}|x_i-y_i|$$

    - Kada se odnosi na binarne vektore, L_1-norma predstavlja Hammingovo rastojanje
  - **L_2-norma** (euklidsko rastojanje)

$$\|\mathbf{x}-\mathbf{y}\|_2=\sqrt{\sum_{i=1}^{D}(x_i-y_i)^2}$$

  - **L_∞-norma** (Čebiševljevo rastojanje)

$$\|\mathbf{x}-\mathbf{y}\|_\infty=\max_{1\le i\le D}|x_i-y_i|$$



Geometrijska mesta tačaka
na jednakom rastojanju od (0,0)

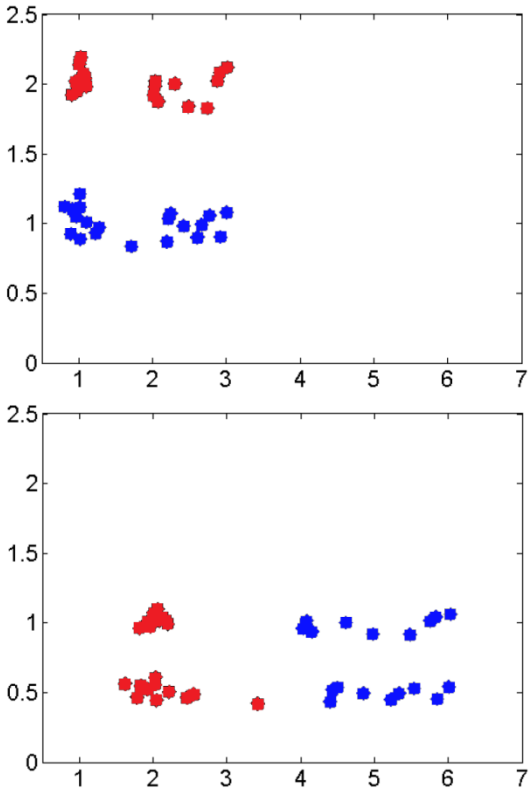
Mere bliskosti

- Postoje i razne druge mere rastojanja između uzoraka, npr:
  - Kvadratno rastojanje

    $$d(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^{\mathsf{T}}\mathbf{B}(\mathbf{x}-\mathbf{y})}$$

    - Specijalan slučaj je Mahalanobisovo rastojanje
- Izbor mere rastojanja zavisi od konkretnog problema
  - Euklidsko rastojanje je opravdano ako su uzorci jednako rasuti po svim dimenzijama
  - Čak i prosto skaliranje osa može dovesti do sasvim drukčije podele na klastere
- Pored mera rastojanja (mera različitosti) mogu se koristiti i mere *sličnosti*
  - Često korišćena mera sličnosti je **skalarni proizvod**:
    $d(\mathbf{x},\mathbf{y}) = \langle \mathbf{x},\mathbf{y} \rangle = \mathbf{x}^{\mathsf{T}}\mathbf{y}$
    - Skalarni proizvod se koristi kada su vektori norme 1, a ako nisu, koristi se **kosinusna mera sličnosti**:

      $$d(\mathbf{x},\mathbf{y}) = \cos\angle(\mathbf{x},\mathbf{y}) = \frac{\langle \mathbf{x},\mathbf{y} \rangle}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}$$

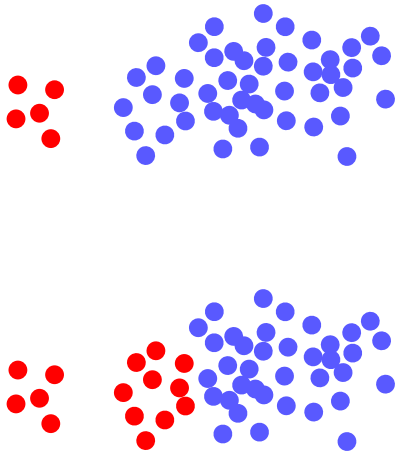


Kriterijumske funkcije za klasterizaciju

- Kada je definisana mera sličnosti/razlike, treba definisati kriterijumsku funkciju koja će biti optimizovana
  - Najčešće korišćena kriterijumska funkcija je suma kvadrata grešaka:

$$J_{\text{MSE}} = \sum_{i=1}^{C} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2, \quad \boldsymbol{\mu}_i = \frac{1}{N_i} \sum_{\mathbf{x} \in C_i} \mathbf{x}$$

    - Ovaj kriterijum opisuje koliko je dobro skup podataka $X = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N)}\}$ reprezentovan centrima klastera $M = \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_C\}$ ($C < N$)
    - Metode klasterizacije koje koriste ovaj kriterijum nazivaju se metode minimalne varijanse
    - Problem se javlja ako se klasteri značajno razlikuju po broju uzoraka, a kriterijumska funkcija je dosta osetljiva i na pojedinačne uzorke koji značajno odstupaju od ostalih (eng. *outliers*)



  - Kriterijumska funkcija može se zasnivati i na matrici unutarklasnog rasipanja:

$$\mathbf{S}_{\text{W}} = \sum_{i=1}^{C} \mathbf{S}_i = \sum_{i=1}^{C} \sum_{\mathbf{x} \in C_i} (\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^{\text{T}}$$

    pri čemu se za skalarnu meru rasipanja koristi trag ili determinanta ove matrice

Optimizacija kriterijumske funkcije

- Kada je pronađena kriterijumska funkcija, potrebno je odrediti particiju skupa uzoraka koja optimizuje izabrani kriterijum
  - Optimalno rešenje može se dobiti jedino ispitivanjem svih mogućih particija, što je računarski neizvodljivo jer ih ima previše
    - Mogućih particija N elemenata na m podskupova ima $m^N/m!$
- Iz ovih razloga **iterativna optimizacija** je uobičajen pristup (iako rezultuje suboptimalnim rešenjem), i ona obuhvata sledeće korake:
  1. Pronaći razumnu početnu particiju
  2. Premeštati uzorke iz jednog klastera u drugi sa ciljem minimizacije/maksimizacije kriterijumske funkcije
- Dve osnovne grupe iterativnih metoda su:
  - Algoritmi **nehijerarhijske klasterizacije** (eng. *flat algorithms*)
    - Ovi algoritmi proizvode skup disjunktnih klastera (npr. *k-means* ili ISODATA)
  - Algoritmi **hijerarhijske klasterizacije**
    - Rezultat primene ovih algoritama je hijerarhija ugnežđenih klastera
    - Mogu biti zasnovani na *povezivanju* (eng. *bottom-up*) ili na *podeli* (eng. *top-down*)

Iterativna optimizacija

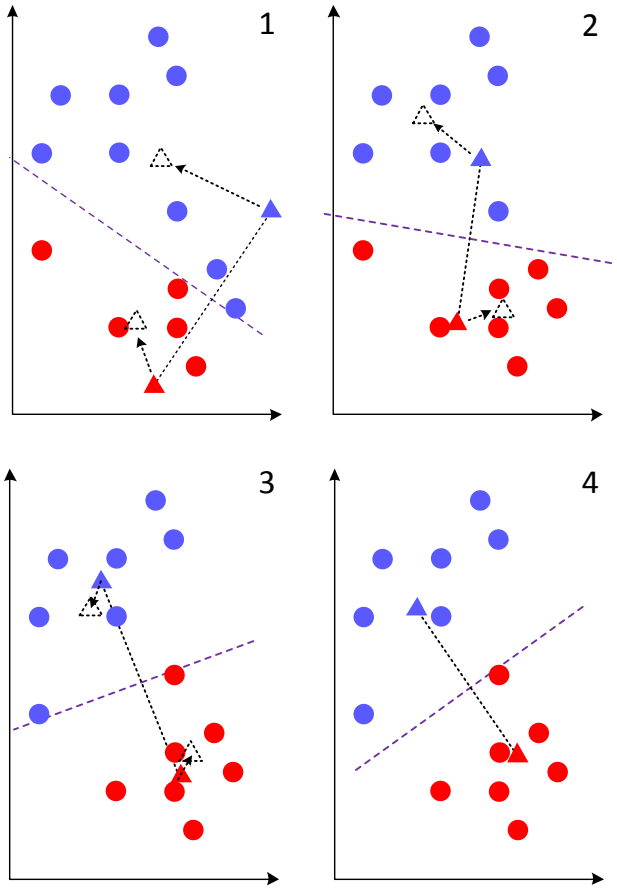
- Primena metoda iterativne optimizacije podrazumeva da su određeni:
  - Mera sličnosti između uzoraka
  - Kriterijumska funkcija za klasterizaciju
    - Predstavlja meru kvaliteta određene klasterizacije
- Dve osnovne grupe iterativnih metoda su:
  - Algoritmi **nehijerarhijske klasterizacije** (eng. *flat algorithms*)
    - Ovi algoritmi proizvode skup disjunktnih klastera (npr. *k-means* ili ISODATA)
  - Algoritmi **hijerarhijske klasterizacije**
    - Rezultat primene ovih algoritama je hijerarhija ugnežđenih klastera
    - Mogu biti zasnovani na *povezivanju* (eng. *bottom-up*) ili na *podeli* (eng. *top-down*)

Algoritam *k* srednjih vrednosti (eng. *k-means*)

- Iterativna minimizacija J_{MSE}

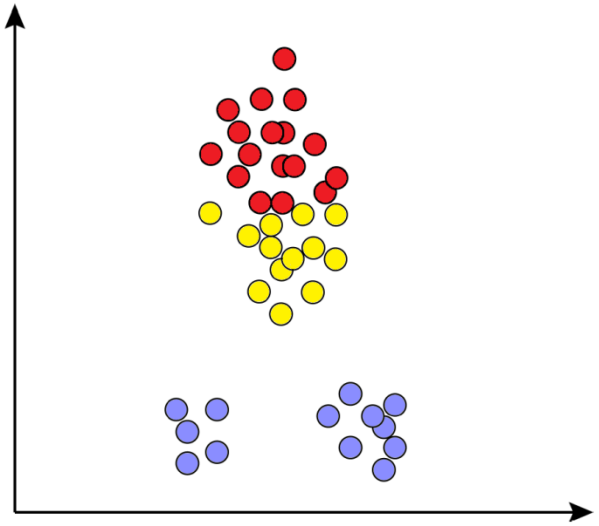
$$J_{\text{MSE}} = \sum_{i=1}^{C} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2, \quad \boldsymbol{\mu}_i = \frac{1}{N_i} \sum_{\mathbf{x} \in C_i} \mathbf{x}$$

1. Definisati broj klastera *k*
2. Inicijalizovati klastere:
   a) Proizvoljnom dodelom uzoraka klasterima, ili
   b) Proizvoljnim postavljanjem centroida klastera
3. Naći uzoračku sredinu svakog klastera i proglasiti tu vrednost novim centroidom
4. Preraspodeliti uzorke tako da pripadnu klasteru čiji im je centroid najbliži
5. Ako je u koraku 4. došlo do bilo kakve promene, vratiti se na korak 3, inače KRAJ



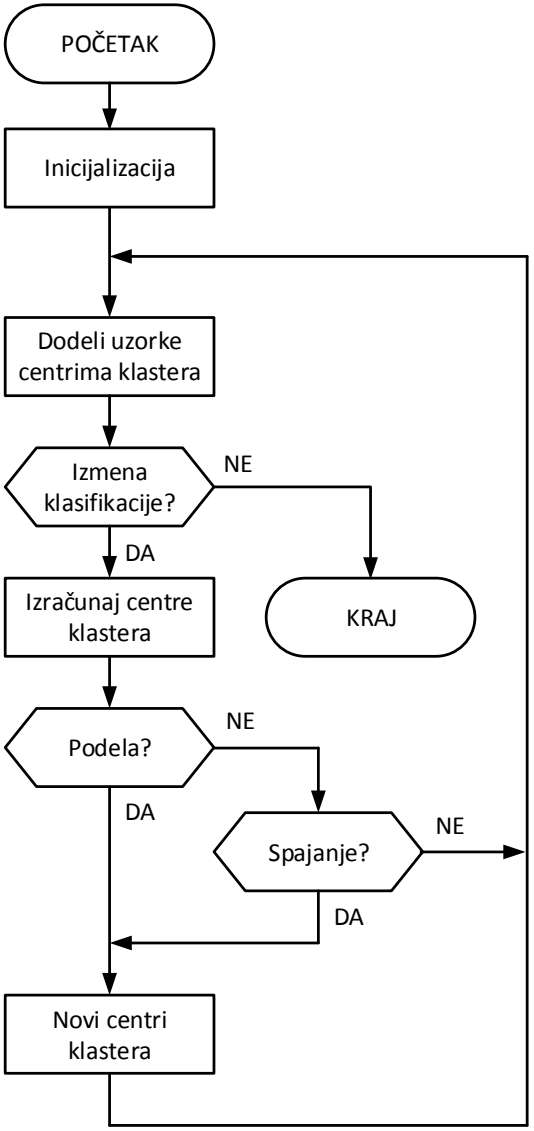
Nedostaci algoritma *k* srednjih vrednosti

- Neophodno je definisati broj klastera na početku
  - Ako je ovaj parametar pogrešan, rezultat klasterizacije će takođe biti pogrešan
- Algoritam je osetljiv na šum, kao i na pojedinačne uzorke koji značajno odstupaju od ostalih
- Neadekvatan izbor centara klastera pri inicijalizaciji takođe može dovesti do pogrešnog rezultata



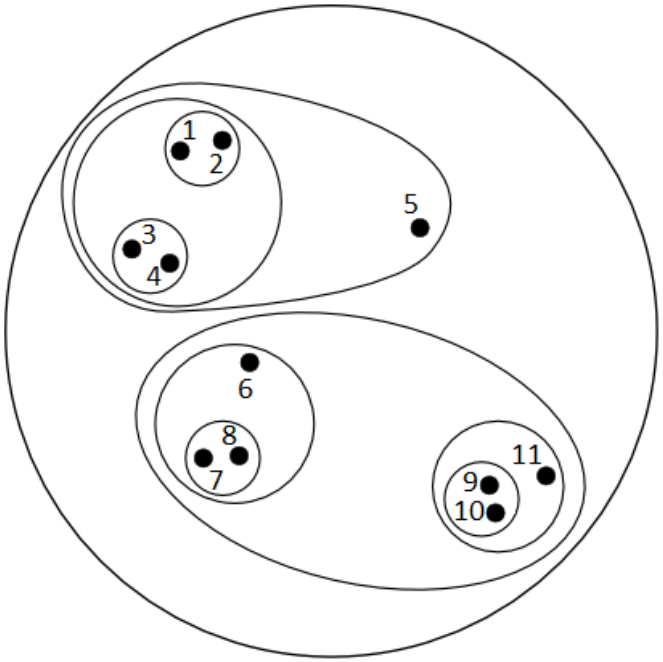
ISODATA algoritam

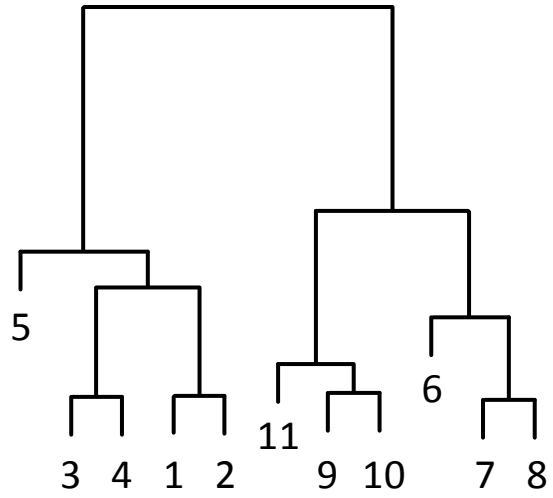
- ISODATA (eng. *Iterative Self-Organizing Data Analysis Technique Algorithm*) predstavlja proširenje *k-means* algoritma određenim heuristikama u cilju automatskog određivanja broja klastera (pri čemu se definiše željeni, približni broj klastera N_D)
- ISODATA podržava mogućnosti:
  - podele klastera koji se po određenoj dimenziji rasipaju više od unapred zadatog parametra σ_s^2 (u slučaju da je broj trenutno identifikovanih klastera premali, odnosno, manji od $N_D/2$)
  - spajanja klastera koji se nalaze veoma blizu, odnosno, čiji su centroidi na rastojanju manjem od unapred zadatog parametra D_{MERGE} (u slučaju da je broj identifikovanih klastera prevelik, odnosno, veći od $2N_D$)
- Algoritam sigurno daje dobar rezultat samo kada su podaci linearno razdvojivi
- Rezultat zavisi od početnih uslova pa treba izvršiti algoritam više puta, s različitim početnim uslovima



Hijerarhijska klasterizacija

- Rezultat primene ovih algoritama je hijerarhija ugnežđenih klastera
- Mogu biti zasnovani na povezivanju (eng. *bottom-up*) ili na podeli (eng. *top-down*)



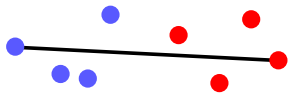


Hijerarhijska klasterizacija zasnovana na podeli

- Postupak obuhvata sledeće korake:
  - Početi od jednog klastera koji obuhvata svih N uzoraka
  - Pronaći najbolju particiju iz skupa od 2^{N-1} mogućih particija i na taj način podeliti klaster na dva klastera
  - Ponavljati prethodni korak za svaki novodobijeni klaster sve dok broj klastera ne bude jednak N
- Glavni problem u ovom pristupu je što je računarski teško izvodljivo ispitati sve moguće particije čak i za relativno male vrednosti N
  - U praksi se ne ispituju sve moguće particije već samo one koje zadovoljavaju određene unapred definisane osobine

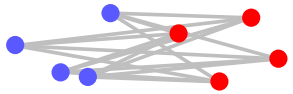
Hijerarhijska klasterizacija zasnovana na povezivanju

- Postupak obuhvata sledeće korake:
  - Početi od *N* zasebnih klastera
  - Pronaći dva najbliža klastera i povezati ih u jedan
  - Ponavljati prethodni korak sve dok broj klastera ne bude jednak 1
- Bliskost dva klastera može se određivati na osnovu različitih mera:
  - **Maksimalno rastojanje** $d_{\max}(C_i, C_j) = \max_{\mathbf{x}\in C_i, \mathbf{y}\in C_j} \|\mathbf{x}-\mathbf{y}\|$

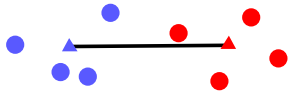


  - **Minimalno rastojanje** $d_{\min}(C_i, C_j) = \min_{\mathbf{x}\in C_i, \mathbf{y}\in C_j} \|\mathbf{x}-\mathbf{y}\|$

  - **Prosečno rastojanje** $d_{\text{avg}}(C_i, C_j) = \frac{1}{N_i N_j} \sum_{\mathbf{x}\in C_i} \sum_{\mathbf{y}\in C_j} \|\mathbf{x}-\mathbf{y}\|$



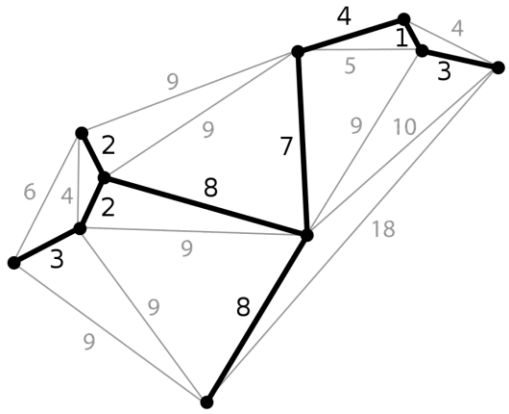
  - **Srednje rastojanje** $d_{\text{mean}}(C_i, C_j) = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|, \quad \boldsymbol{\mu}_i = \sum_{\mathbf{x}\in C_i} \mathbf{x}, \quad \boldsymbol{\mu}_j = \sum_{\mathbf{y}\in C_j} \mathbf{y}$



naročito osetljiva na outliere

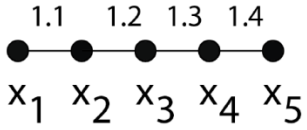
Hijerarhijska klasterizacija zasnovana na povezivanju

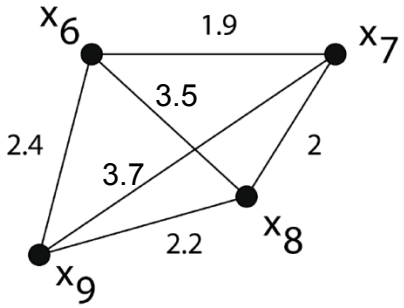
- Ako se koristi minimalno rastojanje:
  - Ova varijanta algoritma se naziva i algoritmom *najbližeg suseda*
  - Ako se algoritam izvršava dok ne preostane samo jedan klaster, rezultat je *minimalni razapinjući graf*
  - Ova varijanta algoritma favorizuje izdužene klastere



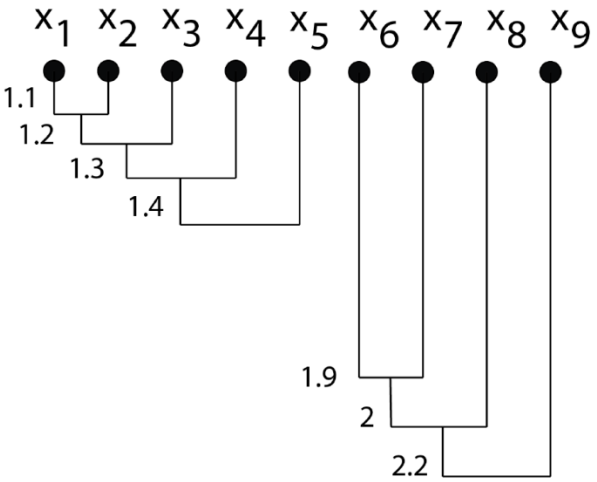
- Ako se koristi maksimalno rastojanje:
  - Ova varijanta algoritma favorizuje kompaktne klastere
- Ako se koriste prosečno ili srednje rastojanje:
  - Osetljivost na outliere mnogo je manja
  - Od svih ovih rastojanja srednje rastojanje je računarski najmanje zahtevno
    - Izračunavanje svih ostalih rastojanja zahteva izračunavanje $N_i N_j$ rastojanja između pojedinačnih uzoraka

Primer 1

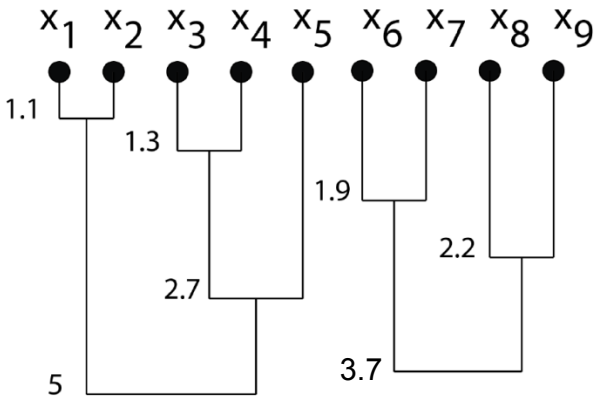




Skup uzoraka koji sadrži jedan prirodan izduženi klaster i jedan prirodan kompaktni klaster (rastojanja koja nisu obeležena imaju veoma velike vrednosti)



Dendrogram dobijen na osnovu minimalnog rastojanja (eng. *single link*) pokazuje da je prilikom klasterizacije prvo formiran izduženi klaster a zatim kompaktni klaster



Dendrogram dobijen na osnovu maksimalnog rastojanja (eng. *complete link*) pokazuje da je prilikom klasterizacije prvo formiran kompaktni klaster a zatim izduženi klaster

Primer 2

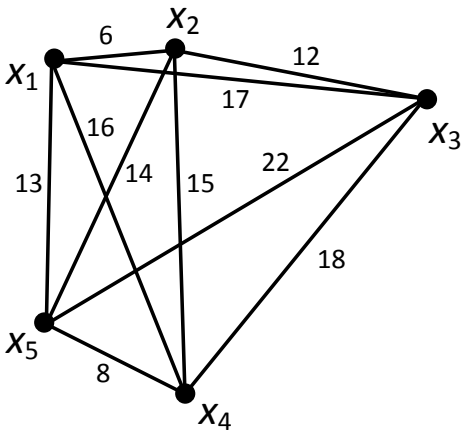
- Korišćenjem *single-link* algoritma odrediti hijerarhijsku klasterizaciju 5 uzoraka na osnovu matrice njihovih rastojanja:

$$\mathbf{P} = \begin{bmatrix} 0 & 6 & 17 & 16 & 13 \\ 6 & 0 & 12 & 15 & 14 \\ 17 & 12 & 0 & 18 & 22 \\ 16 & 15 & 18 & 0 & 8 \\ 13 & 14 & 22 & 8 & 0 \end{bmatrix}$$

Primer 2

- Korišćenjem *single-link* algoritma odrediti hijerarhijsku klasterizaciju 5 uzoraka na osnovu matrice njihovih rastojanja:

$$\mathbf{P} = \begin{bmatrix} 0 & 6 & 17 & 16 & 13 \\ 6 & 0 & 12 & 15 & 14 \\ 17 & 12 & 0 & 18 & 22 \\ 16 & 15 & 18 & 0 & 8 \\ 13 & 14 & 22 & 8 & 0 \end{bmatrix}$$



$$\mathbf{P}_1 = \begin{bmatrix} 0 & 12 & 15 & 13 \\ 12 & 0 & 18 & 22 \\ 15 & 18 & 0 & 8 \\ 13 & 22 & 8 & 0 \end{bmatrix}$$

$$\mathbf{P}_2 = \begin{bmatrix} 0 & 12 & 13 \\ 12 & 0 & 18 \\ 13 & 18 & 0 \end{bmatrix}$$

$$\mathbf{P}_3 = \begin{bmatrix} 0 & 13 \\ 13 & 0 \end{bmatrix}$$