K-means praktična razmatranja

- Da li konvergira? Da li u lokalni ili u globalni minimum?
- ➢Kako da procenimo kvalitet dobijenih klastera?
- Kako da odredimo broj klastera *K*?
- Kako definisati metriku udaljenosti/sličnosti?

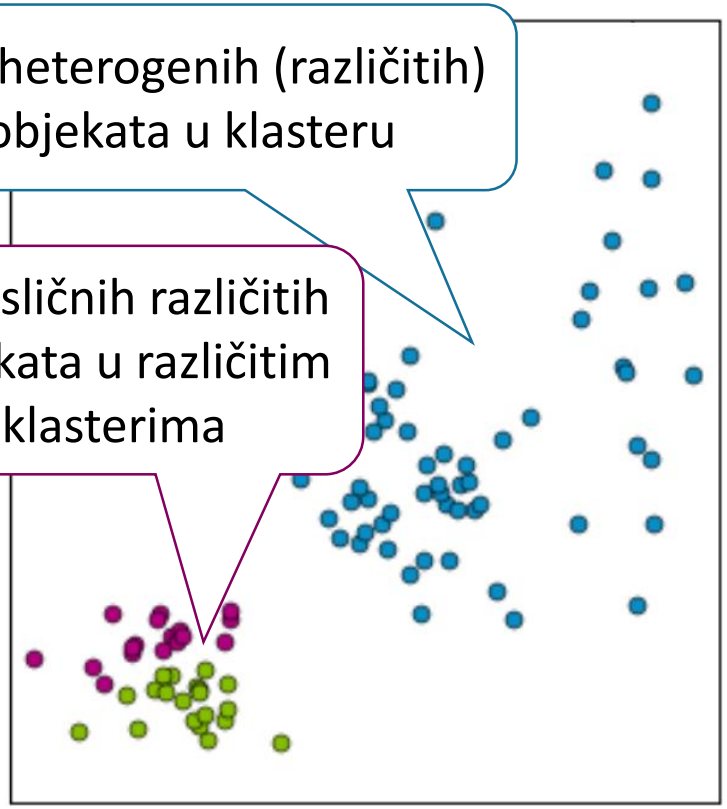
Kvalitet klastera

Neka je broj klastera fiksan (3)

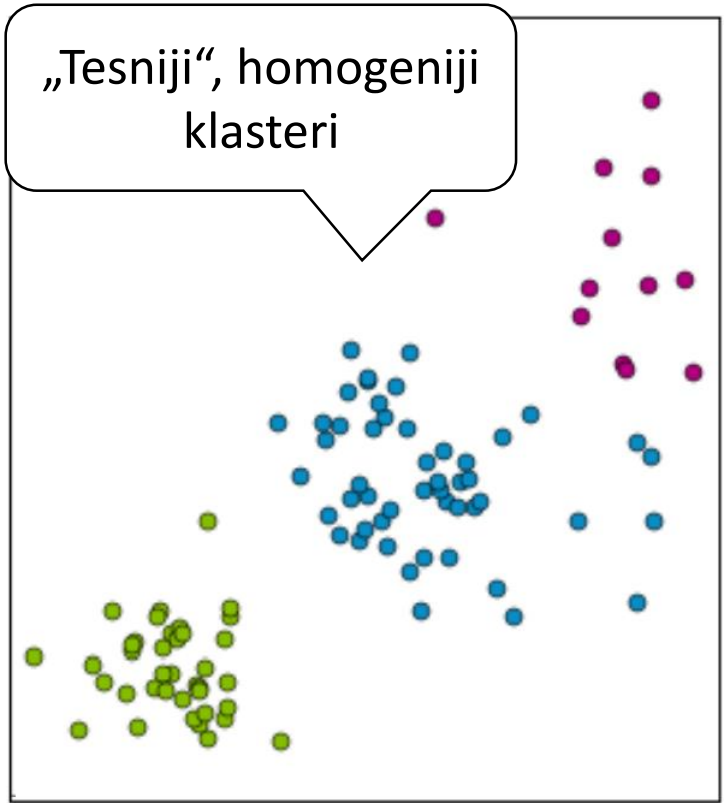
Koji rezultat preferiramo?

Više heterogenih (različitih) objekata u klasteru

Više sličnih različitih objekata u različitim klasterima



„Tesniji“, homogeniji klasteri



Kvalitet klastera

Za fiksan broj klastera

- *K*-means ima za cilj da minimizuje sumu kvadratnih rastojanja opservacija od centroida
- Možemo ovo direktno koristiti kao meru evaluacije (manja vrednost je bolja):

$$\sum_{j=1}^{k} \sum_{i:z^{(i)}=j} \left\| \mu_j - x^{(i)} \right\|_2^2$$

(merimo heterogenost klastera)

Šta ako uvećamo broj klastera?

- Možemo da rafiniramo klastere sve više i više prema podacima

- U ekstremnom slučaju $K = N$
  - Svaka tačka je poseban klaster (svaki centroid odgovara toj jednoj opservaciji u klasteru)
  - Heterogenost je 0!

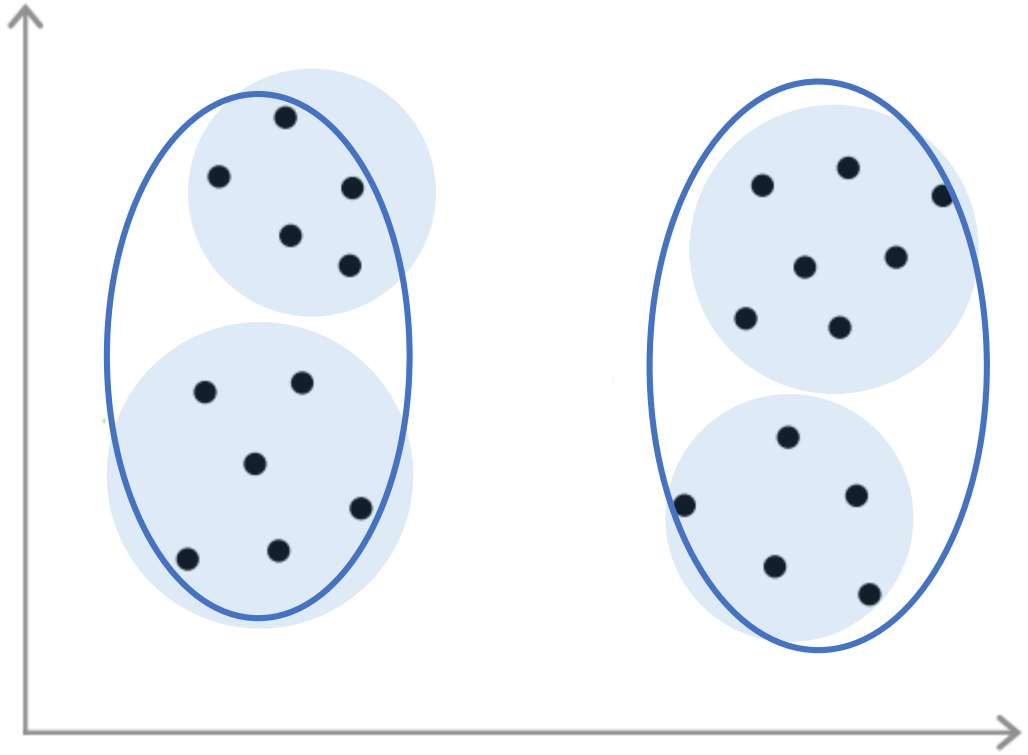
- Sa povećanjem *K* smanjuje se najmanja moguća heterogenost klastera

K-means praktična razmatranja

- Da li konvergira? Da li u lokalni ili u globalni minimum?
- Kako da procenimo kvalitet dobijenih klastera?
- ➢Kako da odredimo broj klastera *K*?
- Kako definisati metriku udaljenosti/sličnosti?

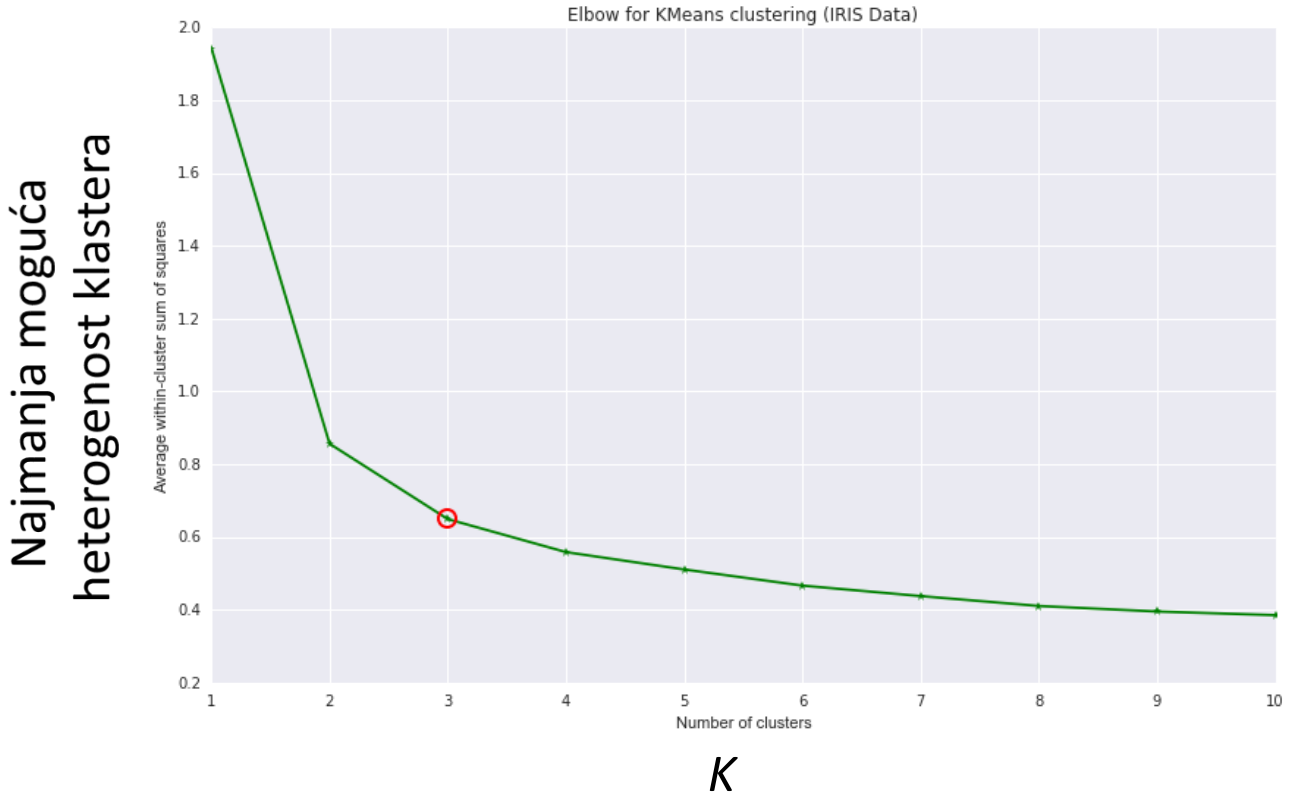
Kako da odaberemo *K*?

- Ovo je netrivijalan problem koji može biti težak i za (ljudskog) eksperta

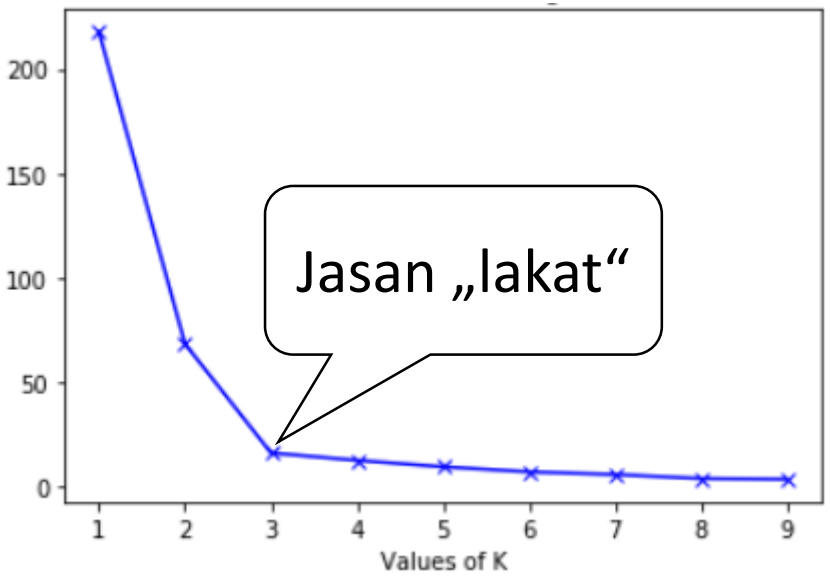


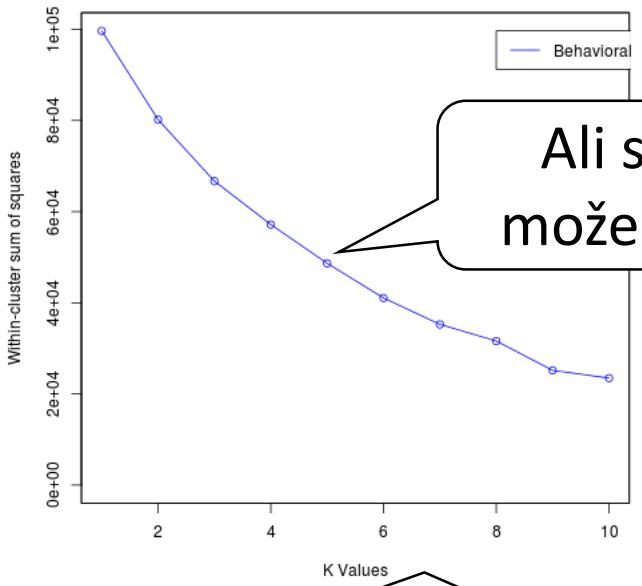
Kako da odaberemo *K*?

- Rekli smo da je manja heterogenost bolja
- Ali ne želimo previše klastera jer onda ne opisujemo dobro strukturu koja postoji u podacima
- Treba nam nagodba između ova dva cilja – heuristika „lakat na krivoj“



Heuristika „lakat na krivoj“





Ali se u praksi može i ovo desiti

Proverite:

- Algoritam
- Pretprocesiranje
- Vizualizujte rezultate
- Razmotrite druge mere poput *Silhouette function*

Heuristika „lakat na krivoj“

- Metod je i računski zahtevan
- Treba da crtamo najmanju moguću heterogenost, što znači da bismo za svaku vrednost *K* trebali isprobati sve moguće podele na *K* klastera
- U praksi , za isto *K* primenićemo više slučajnih inicijalizacija algoritma (čak i ako koristimo *K-means++*) i na grafiku zabeležiti najmanju dobijenu heterogenost

Kako da odaberemo *K*?

- Imajte u vidu da je „lakat na krivoj“ samo heuristika
- Želimo da evaluiramo klasterovanje iz konteksta konkretnog problema koji rešavamo
- Šta želimo od dobre klasterizacije?
  - Objekti unutar istog klastera treba da su slični
  - Objekti iz različitih klastera treba da su manje slični

Kako da odaberemo *K*?

- Na primer, ako klasterujemo tekstualne dokumente, *loš* klastering će imati sledeće osobine:
  - Dokumenti unutar istog klastera imaju mešoviti sadržaj
  - Dokumenti sličnog sadržaja su razbacani u više različitih klastera

- Pokazaćemo ovo na primeru klasterizacije *Wikipedia* stranica
  - Pronaći ćemo članke najbliže centroidima (smatraćemo da reprezentuju klaster) i pregledaćemo njihove naslove i prve rečenice
  - Za svaki klaster ćemo pronaći top 5 reči (prema broju pojavljivanja u klasteru)

Primer – klasterovanje Wikipedia članaka

$K = 2$	Cluster 0	artists, songwriters, professors, politicians, writers, …
	Cluster 1	baseball players, hockey players, soccer (association football) players, …

- Glavne reči u klasteru 1 se odnose na sport
- Glavne reči u klasteru 0 ne pokazuju jasan šablon
- Gruba podela je sportisti/ostalo
- Želeli bismo da klaster 0 podelimo na više kategorija pa ćemo uvećati *K*

Primer – klasterovanje Wikipedia članaka

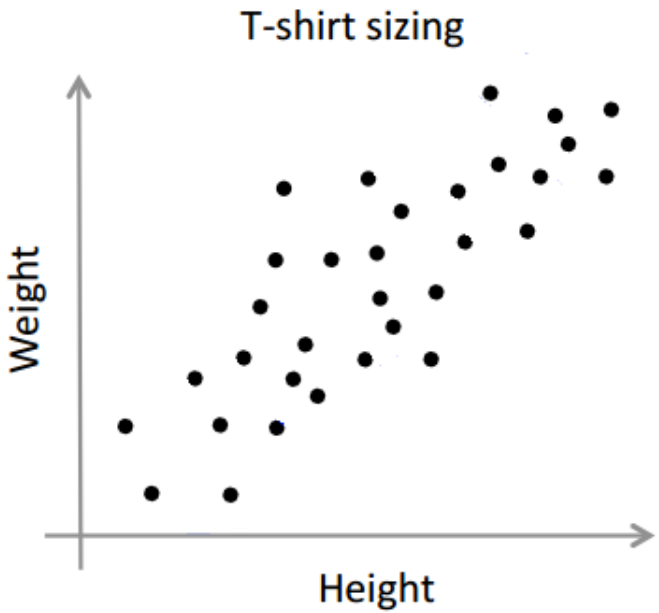
$K = 10$	Cluster 0	artists, actors, film directors, playwrights
	Cluster 1	soccer (association football) players, rugby players
	Cluster 2	track and field athletes
	Cluster 3	baseball players
	Cluster 4	professors, researchers, scholars
	Cluster 5	Austrailian rules football players, American football players
	Cluster 6	female figures from various fields
	Cluster 7	composers, songwriters, singers, music producers
	Cluster 8	composers, songwriters, singers, music producers
	Cluster 9	ice hockey players
	Cluster 10	politicians

- Klasteri 0, 1 i 5 su i dalje mešovitog sadržaja, ali ostali izgledaju dosta konzistentno
- Neki su „veći“ od drugih. Na primer, klaster 4 je sveobuhvatniji od klastera 3 – možda bi uvećavanje broja klastera razbilo veće klastere

Primer – klasterovanje *Wikipedia* članaka

- Visoke vrednosti *K* bi rezultovale čistijim klasterima, ali ne možemo stalno uvećavati *K*
- Moramo se zapitati koliko granularnosti želimo u klasterima
  - Ako želimo grub pregled *Wikipedia* članaka – ne želimo suviše sitnu podelu
  - Ako želimo da se fokusiramo na određeni deo *Wikipedia* članaka – treba da uvećamo*K*

Primer – podela majica na veličine



- Recimo da želimo da prodajemo majice
  - Prikupili smo podatke o visini i težini mušterija
  - Pitamo se koje veličine majica treba da ponudimo (S/M/L ili XS/S/M/L/XL)
  - Možemo probati da klasterujemo mušterije u 3 i u 5 klastera i onda da za svaku podelu procenimo koliko dobro će majice odgovarati mušterijama
  - Dakle, u ovom slučaju evaluiramo broj klastera na osnovu cilja klasterovanja (domena)

Kako da odaberemo *K* – zaključak

- Zaključak: ne postoji zlatno pravilo za izbor *K* – sve zavisi od konkretnog domena i zadatka

- Možemo:
  - Probati da odredimo broj klastera vizuelizacijom skupa podataka (ali ovo ne mora uvek biti moguće)
  - Probati da primenimo heuristiku „lakat na krivoj“, ali ovo ne mora uvek da rezultuje jasnim (ili optimalnim) brojem klastera
  - Analizirati rezultate klasterovanja dobijene za različite vrednosti *K* (primer *Wikipedia* članaka)
  - Proceniti broj klastera na osnovu poznavanja domena/cilja klasterovanja (primer veličina majica)