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?

  - Euklidska udaljenost
  - Kosinusna udaljenost

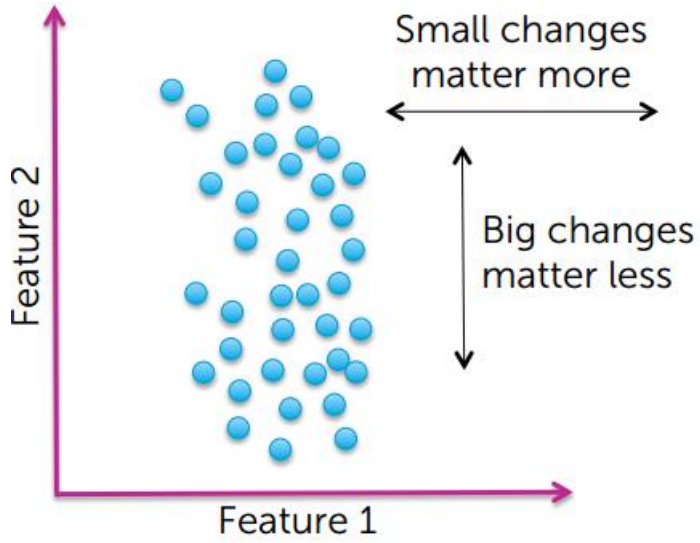
Euklidska udaljenost

- U jednodimenzionom prostoru

$$distance\left(x^{(i)}, x^{(q)}\right) = \sqrt{(x^{(i)} - x^{(q)})^2}$$

- U višedimenzionom prostoru možemo definisati mnogo različitih metrika udaljenosti (npr. dodala težine različitim dimenzijama)

Neka obeležja su relevantnija za problem



Otežinjena Euklidska udaljenost

$$distance\left(x^{(i)}, x^{(q)}\right) = \sqrt{\sum_{d=1}^{D} w_d \left(x_d^{(i)} - x_d^{(q)}\right)^2}$$

w_d - težina svakog obeležja (relativna važnost)

- Ako želimo da skaliramo obeležja koja se kreću u različitom opsegu:

$$w_d = \frac{1}{\max x_d - \min x_d}$$

- Ako specificiramo težine tako da $w_d \in \{0, 1\}$ ovo je ekvivalentno selekciji obeležja
- Kako dizajniramo/selektujemo obeležja je veoma važno, ali i veoma teško jer zahteva domensko znanje

Euklidska udaljenost

Originalni dokumenti:

1	0	0	0	5	3	0	0	1	0	0	0	0
3	1	0	0	2	0	0	1	0	1	0	0	0

$similarity = 5.1$

Isti dokumenti, pri čemu je svakom dupliran broj reči:

2	0	0	0	10	6	0	0	2	0	0	0	0
6	2	0	0	4	0	0	2	0	2	0	0	0

$similarity = 10.2$

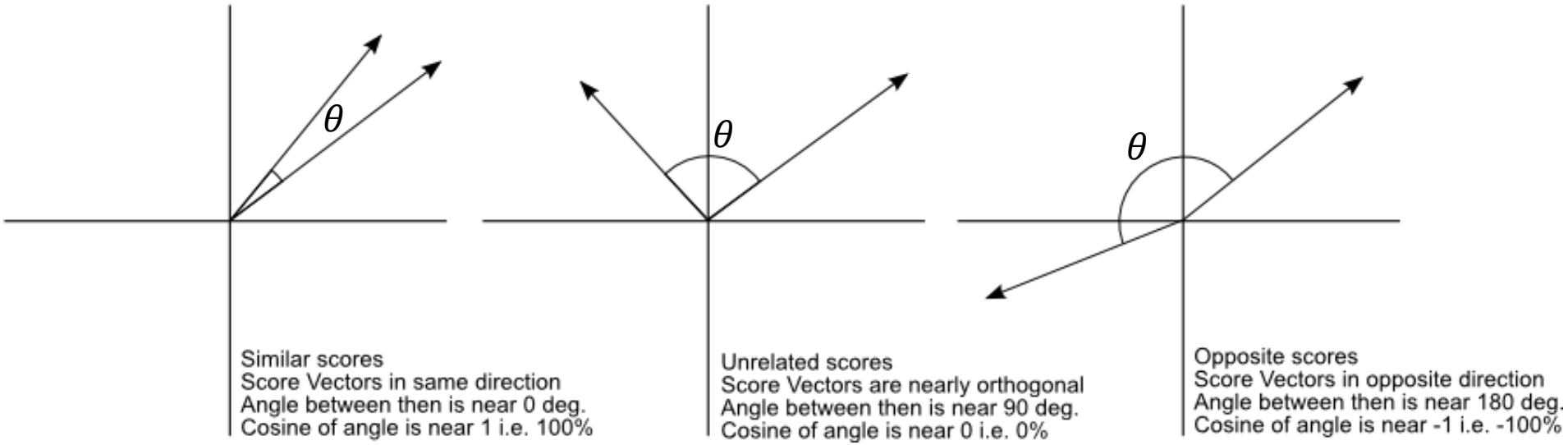
- Nekada nam je opseg bitan i dobro je da dužina vektora utiče na sličnost
- Nekada ovo ne želimo

Kosinusna udaljenost

$$similarity\left(x^{(i)}, x^{(q)}\right) = \frac{\sum_{d=1}^{D} x_d^{(i)} x_d^{(q)}}{\sqrt{\sum_{d=1}^{D} (x^{(i)})^2} \sqrt{\sum_{d=1}^{D} (x^{(q)})^2}} = \frac{{x^{(i)}}^T x^{(q)}}{\|x^{(i)}\| \|x^{(q)}\|}$$

$$= \cos(\theta)$$

Dužina vektora (normalizacija)



Kosinusna udaljenost

- Veoma je efikasna u slučaju *sparse* vektora

- U opštem slučaju važi $-1 < similarity < 1$
  - Za pozitivne vrednosti obeležja (npr. gledamo tekstualni dokument i obeležja su nam frekvencije pojave reči) važi $0 < similarity < 1$

- Udaljenost možemo definisati kao 1 - *similarity*

Kosinusna udaljenost

Originalni dokumenti:

1	0	0	0	5	3	0	0	1	0	0	0	0
3	1	0	0	2	0	0	1	0	1	0	0	0

$similarity = 0.54$

Isti dokumenti, pri čemu je svakom dupliran broj reči:

2	0	0	0	10	6	0	0	2	0	0	0	0
6	2	0	0	4	0	0	2	0	2	0	0	0

$similarity = 0.54$

- Kosinusna sličnost je invarijantna u odnosu na dužinu dokumenta, fokusira se isključivo na sadržaj dokumenata

Normalizovati ili ne?

- Da li želimo da dokumenti budu sličniji što su duži ili nam je važan samo sadržaj?

long document

short tweet

Normalizacija može da rezultuje time da različiti objekti izgledaju mnogo sličniji

long document

long document

Čest kompromis: ograničiti maksimalnu i minimalnu dužinu dokumenta

Mere sličnosti/udaljenosti

- Pored Euklidske i kosinusne mere postoje i mnoge druge mere sličnosti/udaljenosti
  - Manhattan, Jaccard, Hamming, Correlation-based, Rank-based, Mahalanobis,…
- Mere se mogu i kombinovati
- Na primer, jedan dokument možemo reprezentovati tekstom i obeležjem koje nam govori koliko puta je dokument pročitan. Kada upoređujemo dva dokumenta:
  - Koristićemo kosinusnu sličnost na obeležjima vezanim za tekst
  - Koristićemo Euklidsku udaljenost za broj čitanja

Zaključak

- Kako definisati metriku udaljenosti/sličnosti?

- Nema jednog odgovora primenljivog na sve. Kada biramo meru, to veoma zavisi od rešavanog problema