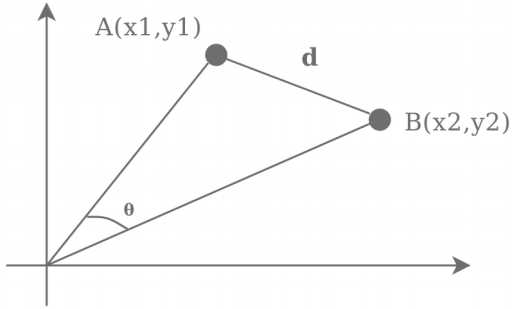
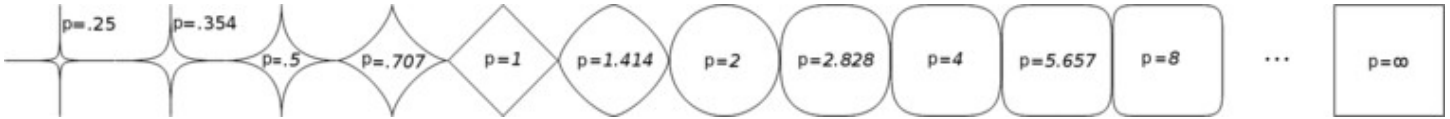
Zadanie 1 – metriky a zhlukovacie algoritmy

4. októbra 2019
I-SUNS cv. 2

Metriky

- Minkowski vzdialenosť: $d = \left(\sum_{i=1}^{n} |A_i - B_i|^p \right)^{1/p}$
  - Manhattanská: $p = 1$
  - Euklidovská: $p = 2$





- Kosínusová: $\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$
- Mahalanobis, ...

Zdroje obrázkov:
https://lylelin317.wordpress.com/2014/03/11/distance-and-similarity-measuring-methods-in-machine-learning/
https://cmry.github.io/notes/euclidean-v-cosine

Normalizovanie dát

	HDP (mil. $) 18 036 648 – 1 559	Pop. cez 60 rokov 0.024 – 0.334	Podiel žien v pracovnom pomere 0.124 – 0.861
Saudská Arábia	653 219	0.056	0.2
Švajčiarsko	670 790	0.241	0.62
Portugalsko	199 122	0.279	0.53
Jordánsko	37 517	0.057	0.145

Normalizovanie dát

	Saudská Arábia	Švajčiarsko	Portugalsko	Jordánsko
Saudská Arábia				
Švajčiarsko	17 570 mil.			
Portugalsko	454 097 mil.	471 668 mil.		
Jordánsko	615 701 mil.	633 273 mil.	161 605 mil.	

Normalizovanie dát

- Do rozmedzia <0,1> alebo <-1,1>
- Základné typy:
  - Min-max (rescaling)

$$x_{norm} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

  - Standardization

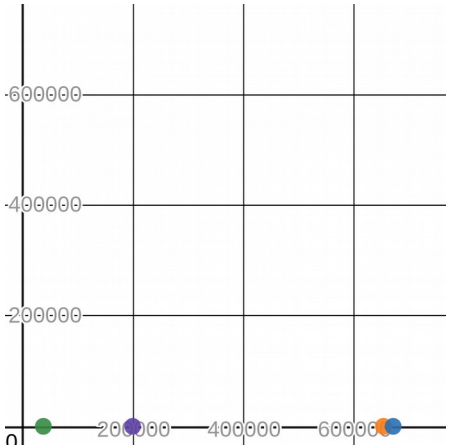
$$x_{norm} = \frac{x - \mu}{\sigma}$$

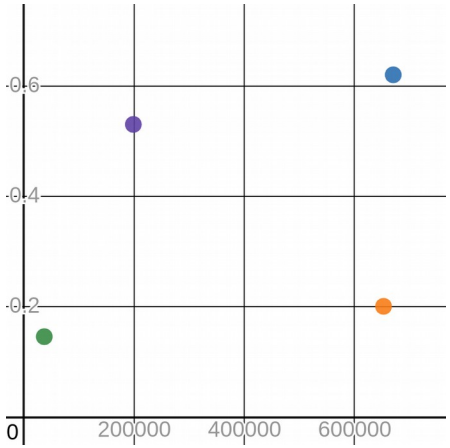
  - L2 normalizácia

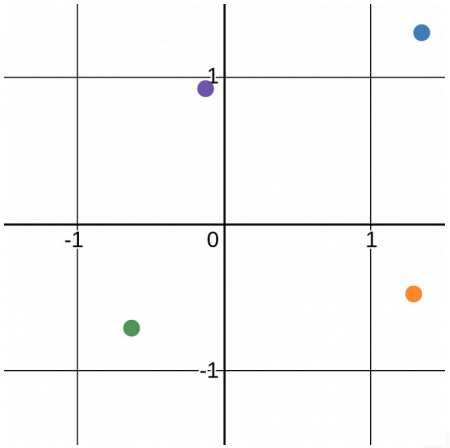
$$x_{norm} = \frac{x}{\|x\|}$$

Normalizovanie dát

Saudská Arábia
Švajčiarsko
Portugalsko
Jordánsko







Zhlukovacie algoritmy

Kmeans, DBSCAN, Chinese whispers

Zhlukovanie

- Učenie bez učiteľa
- Input: body v priestore (bez známej triedy, spoločných čŕt ...)
- Output: skupiny (potenciálne triedy) z bodov, kde body v rámci skupiny sú si navzájom viac podobné ako body z rozdielnych skupín
- Obvykle sú body vo viacrozmernom priestore a ich podobnosť určujeme pomocou vzájomných vzdialeností (Euklid, Mahalanobis, L1 ...)

How many clusters?

Six Clusters

Two Clusters

Four Clusters

Zdroje obrázkov:
https://apandre.wordpress.com/visible-data/cluster-analysis/

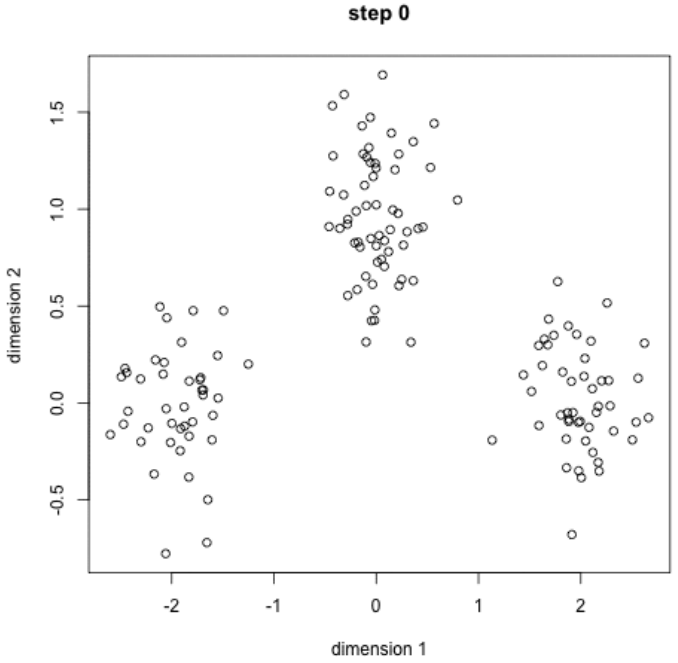
Typy zhlukovaní

- Connectivity models
  - Hierarchické modely
  - Nevhodná pre veľké datasety
- Centroid models
  - K-means
  - Počet zhlukov určený predom
- Distribution models
  - Založené na pravdepodobnosti
  - Náchylné na pretrénovanie
- Density models
  - Hľadajú “husto obsadené” podpriestory
  - DBSCAN

Zdroje:
https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/

K-means

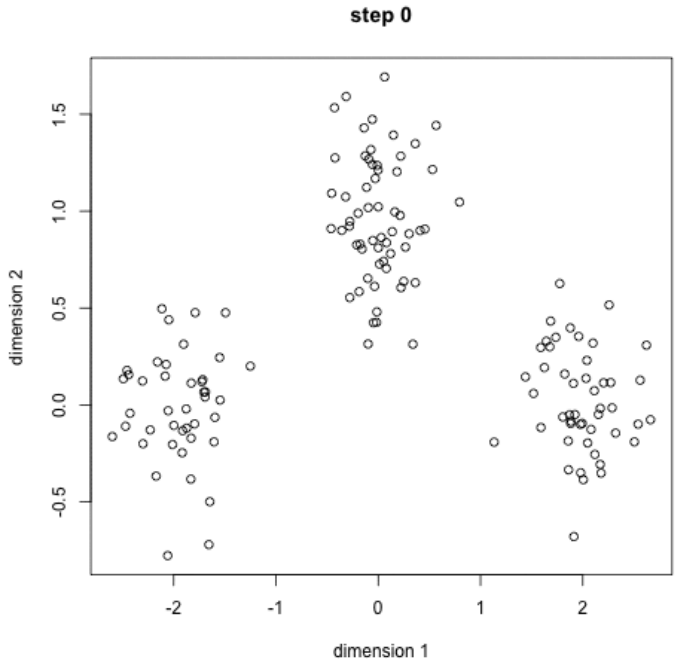
1. Zvoľme počet clusterov, každému prislúcha jeden centroid zvolený náhodne
2. Každý bod priradíme ku tomu zhluku, ktorého centroid je mu najbližšie
3. Každému clustru vypočítame nový centroid ako priemer bodov priradených clustru
4. Opakujeme bod 2-3 do konvergencie alebo zastavujúcej podmienky



Zdroje:
https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68
https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

K-means

- Lineárna zložitosť O(n)
- Nekonzistentný, možno neopakovateľný
- Nie vždy poznáme počet zhlukov dopredu



Zdroje:
https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68
https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

DBSCAN

1. Vyberme bod, ktorý ešte nebol navštívený
2. Ak má v okolí (d<hyperparameter ε) dostatočný počet bodov (hyperparameter minPoints), stáva sa z neho počiatok clustra. Body z okolia pridáme do clustra. Označíme ho ako navštívený.
3. Pre každý nenavštívený bod v clustri, pridáme všetky body z jeho okolia do clustra a označíme ho ako navštívený.
4. Opakujeme bod 3, kým je čo pridať.
5. Opakujeme bod 2-4 do konvergencie alebo zastavujúcej podmienky.

Zdroje:
https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68
https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

DBSCAN

- Netreba dopredu určený počet clusterov
- Vie identifikovať noise
- Nevhodný, ak majú zhluky rozdielnu hustotu

Zdroje:
https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/

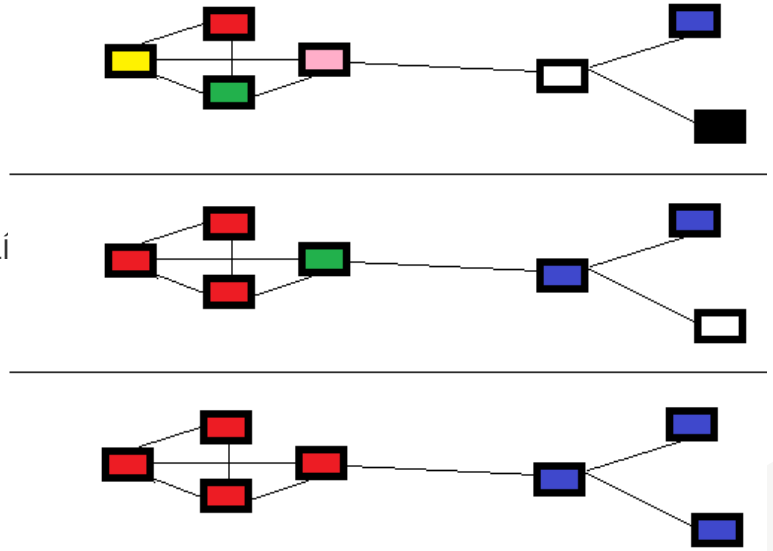
Chinese Whispers

1) Každý bod je svojím vlastným clusterom.

2) Vyberme náhodne bod, pozrime, do akých zhlukov patria body v jeho okolí (hyperparameter ε) a priraďme ho do najpočetnejšieho zhluku (pri remíze vyberieme náhodne).

3) Opakujme bod 2. pre každý bod v priestore.

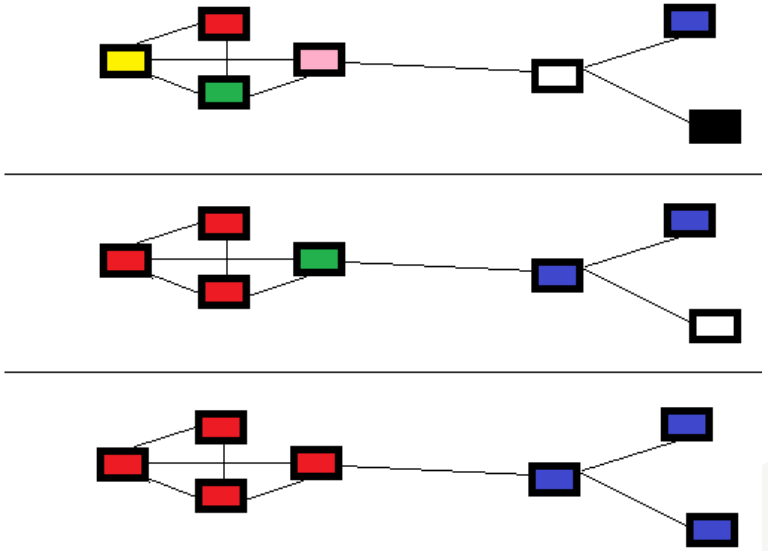
4) Opakujeme bod 2-3 do konvergencie alebo zastavujúcej podmienky.



Zdroje:
https://en.wikipedia.org/wiki/Chinese_Whispers_(clustering_method)

Chinese Whispers

- Lineárna zložitosť O(n)
- Netreba dopredu určený počet clusterov
- Nekonzistentný, možno neopakovateľný



Zdroje:
https://en.wikipedia.org/wiki/Chinese_Whispers_(clustering_method)

Priestor na otázky