DĚLAT
DOBRÝ SOFTWARE
NÁS BAVÍ

PROFINIT

Clustering velkých dat

Jan Hučín

12. dubna 2019

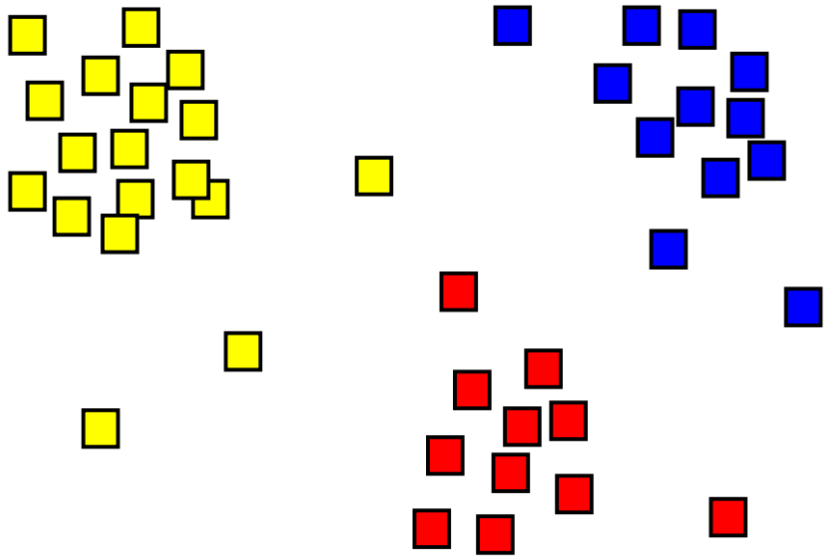
Osnova

1. Účel a typy clusteringu
2. Metriky a podobnosti
3. Aglomerativní metody
4. Přiřazovací metody
5. Clustering a Spark ML

PROFINIT

Účel clusteringu

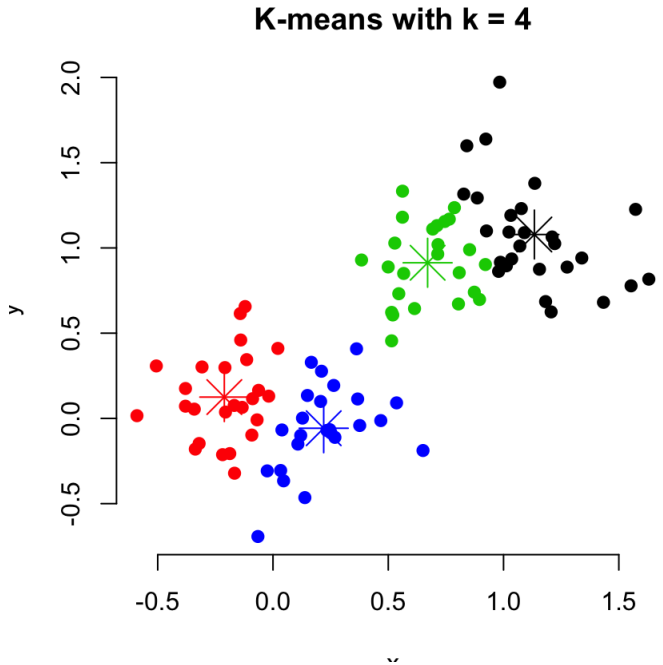
- unsupervised learning
- sdružení jednotek do logických shluků (clusterů)
  - blízké jednotky v jednom clusteru
  - vzdálené jednotky v různých clusterech
  - vyžaduje metriku vzdálenosti / podobnosti

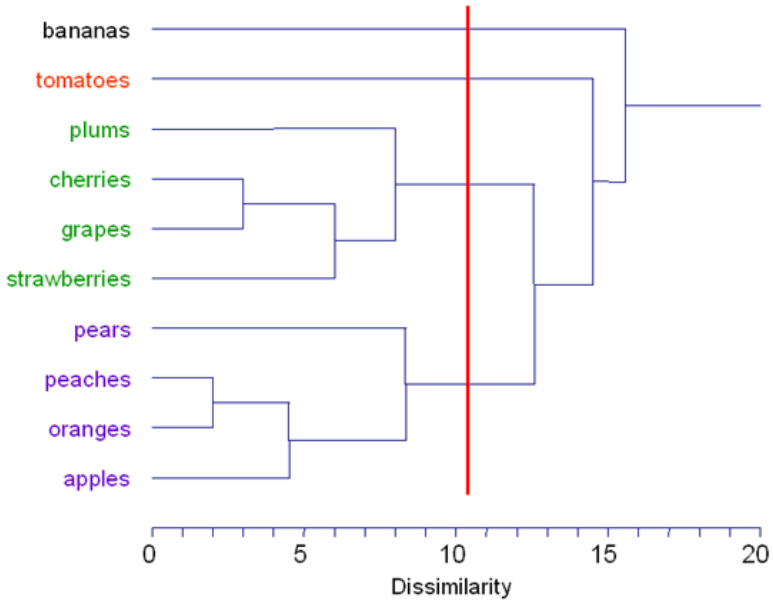


PROFINIT

Typy clusteringu

- aglomerativní
  - jednotky se sdružují postupně
  - počet clusterů není předem dán
- přiřazovací
  - počet clusterů dán předem
  - jednotky se přiřazují definovaným clusterům





PROFINIT

Metriky a podobnosti

- Určuje vzdálenost mezi dvěma prvky
  - čím podobnější, tím bližší
- Vzdálenost
  - jednotky jako body v nějakém prostoru – délka cesty
  - norma vektoru: Eukleidovská, Manhattan, maximální (L_∞)
  - problém vysoké dimenzionality
  - editační (Levensteinova) vzdálenost mezi řetězci
- Podobnost
  - jednotky jako množiny vlastností
  - číselné vyjádření shody množin nebo shody míry vlastností
  - <0; 1>

PROFINIT

Podobnosti množin

A, *B* – množiny

Jaccardova podobnost

$$\frac{|A \cap B|}{|A \cup B|}$$

Cosinová podobnost

$$\frac{|A \cap B|}{\sqrt{|A||B|}}$$

Další: Dice-Sorensen, overlap atd.

PROFINIT

Podobnosti vektorů

A, *B* – vektory v n-rozměrném prostoru

vážená Jaccardova podobnost

$$\frac{\sum_i \min(a_i, b_i)}{\sum_i \max(a_i, b_i)}$$

Cosinová podobnost

$$\frac{\sum_i a_i b_i}{\|A\|\|B\|}$$

Jiné podobnosti:

› např. podobnost řetězců – délka společného podřetězce

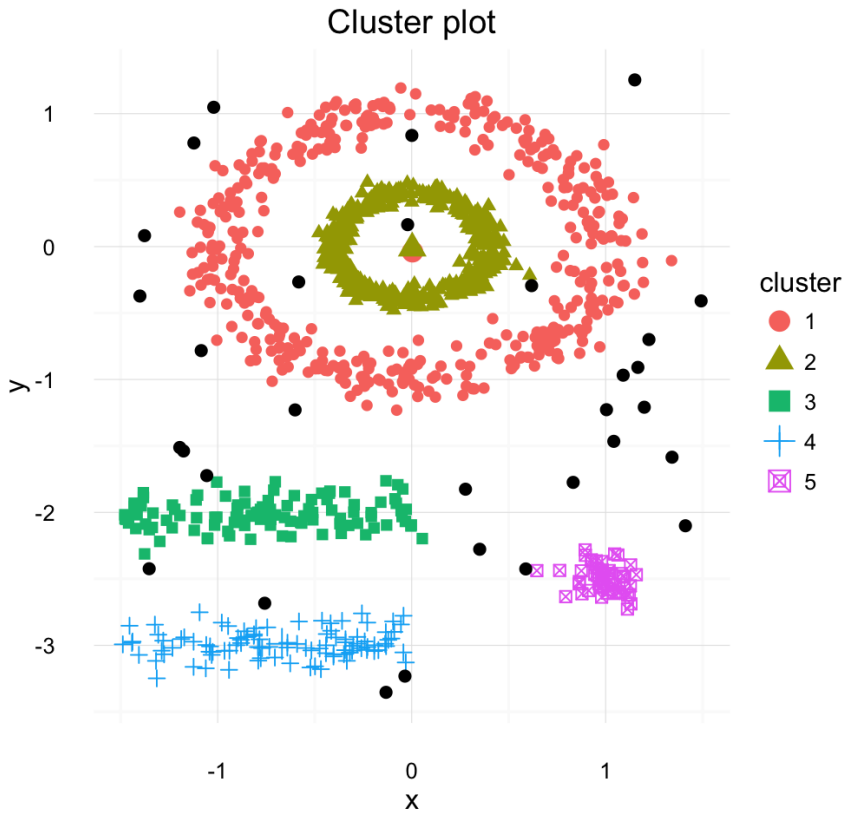
PROFINIT

Aglomerativní metody

- postupné sdružování nejbližších jednotek/clusterů
- mohou vznikat i složité clustery
- finální clusterování – lze dynamicky

Příklady:

- hclust
- dbscan



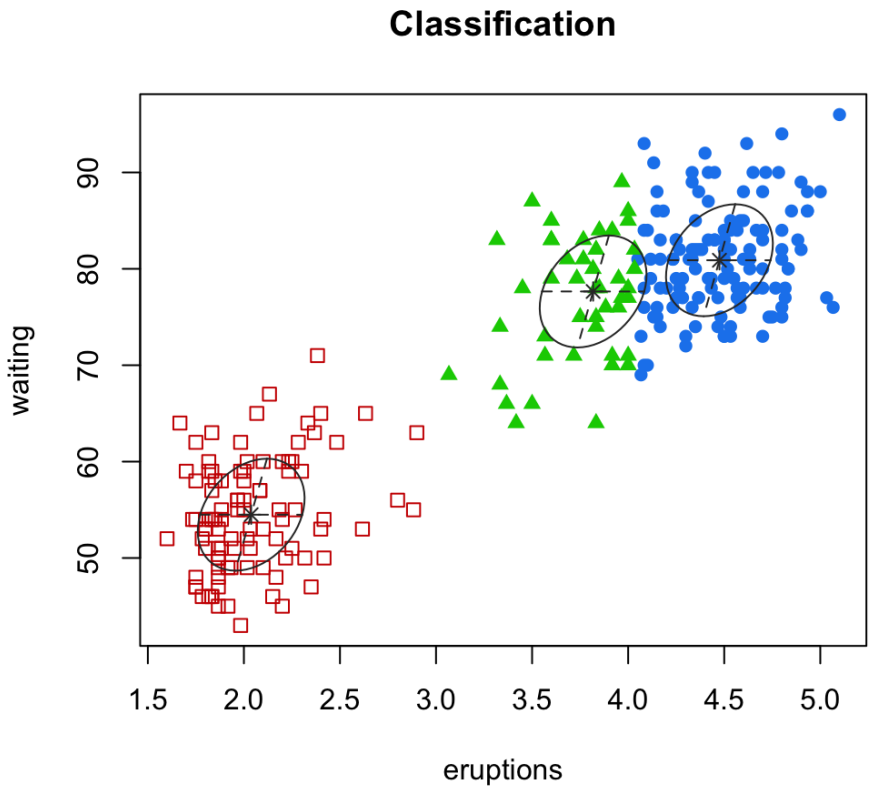
PROFINIT

Přiřazovací metody

- stanoví se počet clusterů
- výchozí reprezentanti clusterů (centroid, clusteroid)
- body postupně přiřazovány
- opakování s jinou výchozí reprezentací clusterů

Příklady:

- k-means
- Gaussian mixture
- Power Iteration (PIC)



PROFINIT

Spark ML

Co je Spark ML

- nadstavba nad Sparkem
- pro RDD i pro DataFrame
- popisné statistiky
- lineární algebra
- modely (regrese, Bayes, stromy)
- redukce dimenzionality (hlavní komponenty)
- clustering
- a další (viz spark.apache.org)

Clustering v Spark ML 1.6

- aglomerativní metody ne – mj. příliš náročné, více než $O(N^2)$
- **K-means**
- Gaussian mixture
- **Power Iteration Clustering (PIC)**
- **Latent Dirichlet Association (LDA)**
- a další

PROFINIT

K-means

Princip:

- jednotka náleží do clusteru, k jehož středu je nejblíže (L_2)

Vstup:

- RDD s elementy array
- K
- parametry pro běh (mj. počet opakovaných běhů)

Výstup:

- centroidy
- metoda pro zatřídění obecného bodu

PROFINIT

Power Iteration Clustering

Princip:

- detekce komunit v grafu
- embedding grafu do 1D prostoru + k-means

Vstup:

- RDD jako řídká trojúhelníková matice afinit (podobností)
- K

Výstup:

- přiřazená ID clusterů (komunity)

Problémy:

- Ilustrační příklad nekonverguje.
- Na Metacentru padá při vyšším limitu počtu iterací.
- Řešeno vlastní implementací.

PROFINIT

Latent Dirichlet Association

Princip:

- stanoví témata (topics) podle frekvence slov v dokumentech
- řeší problém vysoké dimenzionality

Vstup:

- korpus – RDD vektorů (řádky=dokumenty, sloupce=slova, hodnoty=četnosti)
- K

Výstup:

- popisy témat pomocí nejtypičtějších slov

PROFINIT

Díky za pozornost

PROFINIT

Profinit, s.r.o.
Tychonova 2, 160 00 Praha 6

Telefon
+ 420 224 316 016

Web
www.profinit.eu

LinkedIn
linkedin.com/company/profinit

Twitter
twitter.com/Profinit_EU