



Strojové učení

10 Shlukování (Clustering)

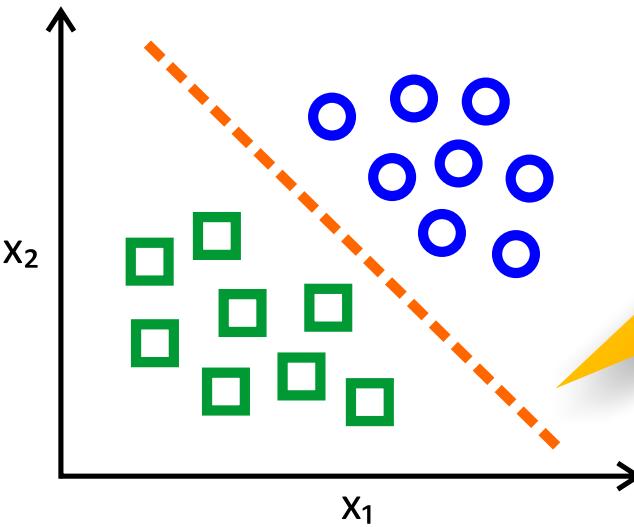
- Učení bez učitele
- Shlukování a jeho aplikace
- Metoda K-means
- Optimalizační kritérium K-means
- Výběr centroidů
- Volba počtu shluků





Úvod do učení bez učitele

Učení s učitelem (paradigma pro připomenutí)



hypotéza získaná učením na trénovací množině definiuje rozhodovací hranici, oddělující vzorky třídy od vzorků třídy .

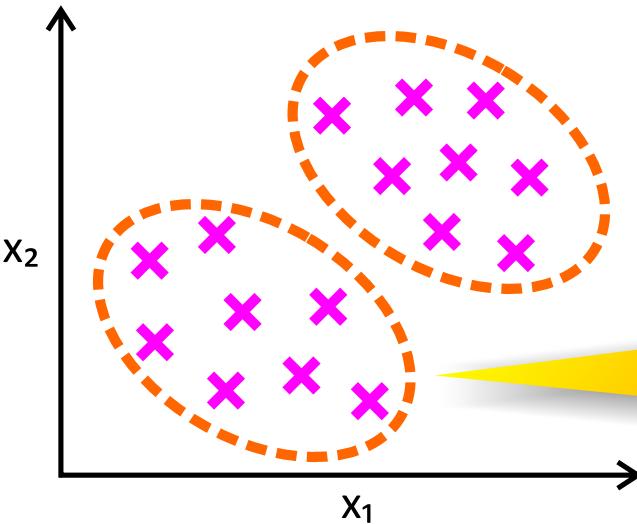
Trénovací množina:

$$\{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), (\mathbf{x}^{(3)}, y^{(3)}), \dots, (\mathbf{x}^{(m)}, y^{(m)})\}$$



Úvod do učení bez učitele

Paradigma učení bez učitele



shlukovací algoritmus
(Clustering Algorithm)

algoritmus se pokouší nalézt v datech nějakou organizaci/strukturu – např. **shluky**

Trénovací množina:

$$\{ \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \dots, \mathbf{x}^{(m)} \}$$

odpovědi učitele nejsou k dispozici



Úvod do učení bez učitele

Typické aplikace shlukování – komerce

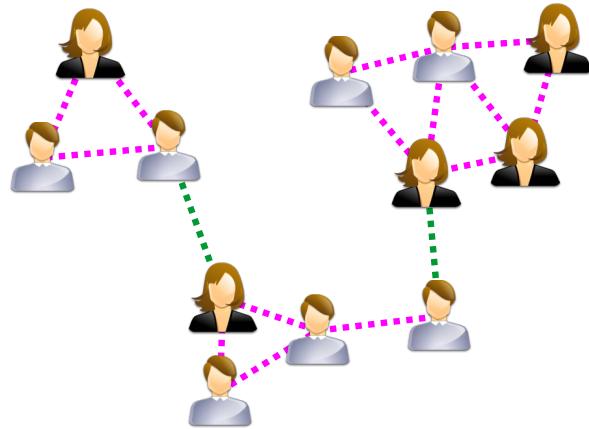
(Image © <http://marketsegmentation.blogspot.cz/>)

Analýza/segmentace trhu



- reklama
- marketing
- výzkum příležitostí

Analýza sociálních sítí



- recommender systems
- reklama/marketing
- kustomizace
- bezpečnost
- výzkum (sociologie)



Úvod do učení bez učitele

Typické aplikace shlukování – věda

IT – organizace výpočetních
clusterů/cloudů



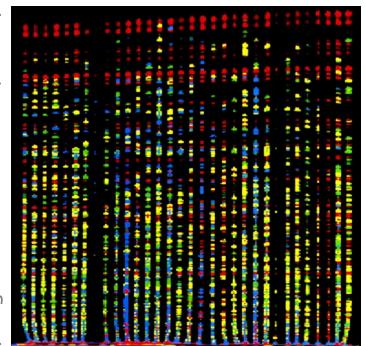
(Image © <http://science.psu.edu>)

Astronomie



© 1986 Jerry Lodriguss and John Martinez

Genetika



(Image © Genome Research / CSHL)

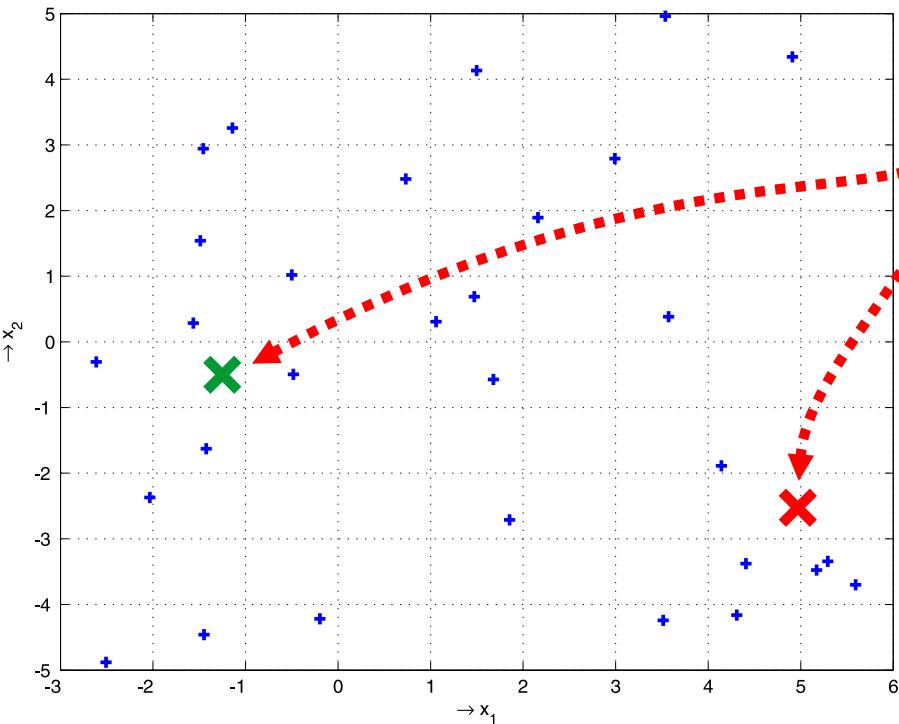


Shlukování metodou K-středních (*K-means*)

Centroidy (počáteční výběr)

K-means je 2-krokový iteracní algoritmus:

| : (1) výpočet shluků, (2) přesun centroidů :|



centroidy shluků*
(*Cluster Centroids*)

*budoucích
(počáteční volba je náhodná)

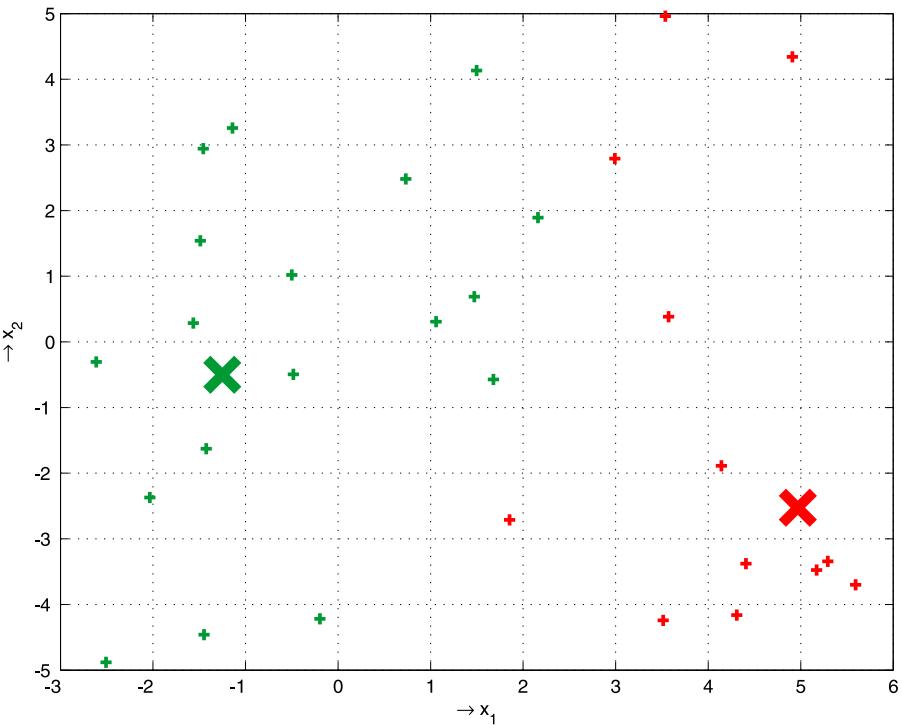
} cheme-li mít 2
shluky, volíme
2 centroidy...



Shlukování metodou K-středních (*K-means*)

Krok 1 – Výpočet shluků

Každý vzorek přiřadíme k jednomu z centroidů – **nejbližšímu...**

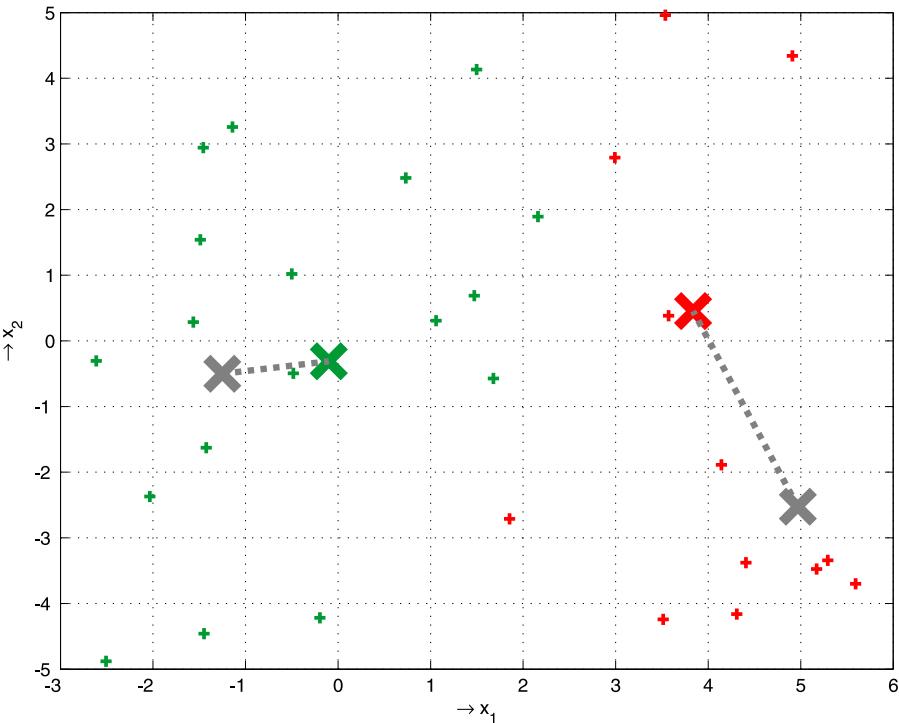




Shlukování metodou K-středních (*K-means*)

Krok 2 – Přesun centroidů

Centroidy přesuneme do pozice střední hodnoty (*Mean*) souřadnic přiřazených projekcí vzorků...



Po přesunutí opět provedeme krok 1 – přiřazení projekcí vzorků nejbližšímu centroidu...



Shlukování metodou K-středních (*K-means*)

Algoritmus

Vstup:

- parametr K (počet shluků)
- trénovací množina $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \dots, \mathbf{x}^{(m)}\}$, $x_0^{(i)} = 1$

Inicializace:

- náhodně zinicializujeme K centroidů $\mu_1, \mu_2, \dots, \mu_K \in \mathbb{R}^n$

Iterace:

do {

for $i \leftarrow 1 \dots m$

 ① $c^{(i)} \leftarrow$ index k centroidu nejblíže $\mathbf{x}^{(i)}$, tj. $\min_k \|\mathbf{x}^{(i)} - \mu_k\|^2$

for $k \leftarrow 1 \dots K$

 ② $\mu_k \leftarrow$ průměr průmětů vzorků $\mathbf{x}^{(i)}$, kde $c^{(i)} = k$

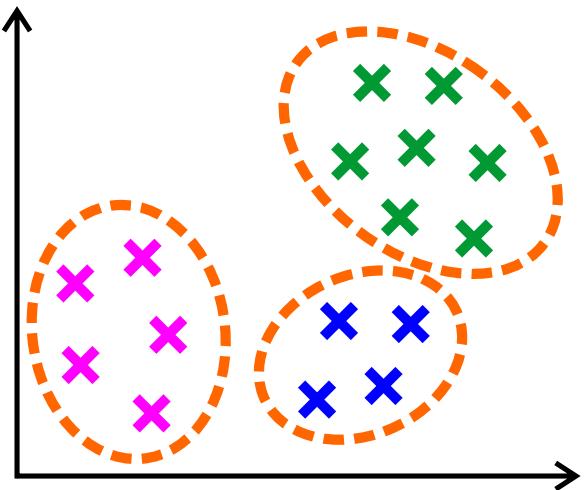
}



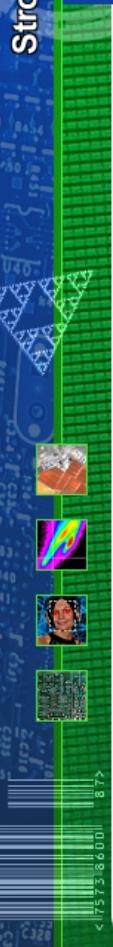
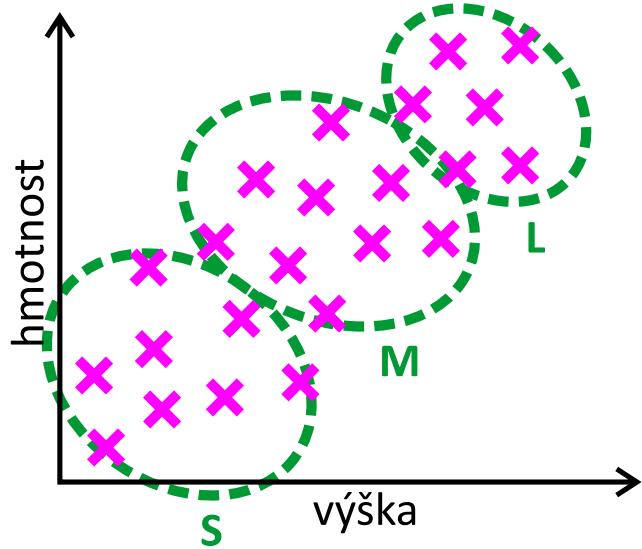
K-means pro obtížně separovatelné shluky

Segmentace trhu

ideálně separabilní třídy



realita: **velikost oblečení**





K-means: Cíl optimalizace

Proč analyticky vyjadřovat cenovou funkci?

Proč vyjadřovat matematicky **optimalizační kritérium**, když jej algoritmus (zjevně) nepotřebuje?

- jako každý učící se algoritmus, i **K-means Clustering** má vyjádřitelné optimalizační kritérium, tj. cenovou funkci, kterou je třeba minimalizovat – je vhodné jí dokázat zapsat
- známá (a vypočítatelná) cenová funkce umožňuje algoritmus **ladit** – ukazuje, zda běží korektně a konverguje
- na základě optimalizačního kritéria lze stanovit optimální počet shluků K



K-means: Cíl optimalizace

Optimalizační kritérium (cenová funkce)

Označme:

$c^{(i)}$ – index shluku ($1, 2, \dots, K$), ke kterému je přiřazen vzorek $\mathbf{x}^{(i)}$

$\boldsymbol{\mu}_k$ – centroid shluku k ($\boldsymbol{\mu}_k \in \mathbb{R}^n$)

$\boldsymbol{\mu}_{c^{(i)}}$ – centroid shluku, ke kterému je přiřazen vzorek $\mathbf{x}^{(i)}$

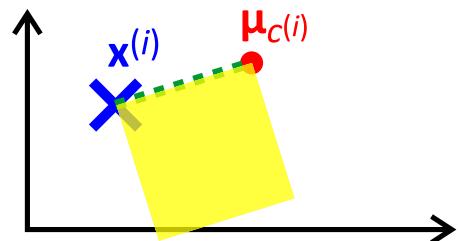
Cenová funkce:

$$J(c^{(1)}, c^{(2)}, \dots, c^{(m)}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \dots, \boldsymbol{\mu}_K) = \frac{1}{m} \sum_{i=1}^m \|\mathbf{x}^{(i)} - \boldsymbol{\mu}_{c^{(i)}}\|^2$$

Optimalizační kritérium:

$$\min_{c^{(1)}, \dots, c^{(m)}, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K} J(c^{(1)}, \dots, c^{(m)}, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K)$$

zkreslení (*Distortion*)
algoritmu K-means





K-means: Cíl optimalizace

Souvislost algoritmu a optimalizačního kritéria

Inicializace:

- náhodně zinicializujeme K centroidů $\mu_1, \mu_2, \dots, \mu_K \in \mathbb{R}^n$

Iterace:

do { minimalizace $J(\dots)$ vzhledem k $c^{(1)}, c^{(2)}, \dots, c^{(m)}$
 $(\mu_1, \mu_2, \dots, \mu_K$ se nemění)

1 **for** $i \leftarrow 1..m$
 $c^{(i)} \leftarrow$ index k centroidu nejblíže $\mathbf{x}^{(i)}$, tj. $\min_k \|\mathbf{x}^{(i)} - \mu_k\|^2$

for k ← 1 .. K

$\mu_k \leftarrow$ průměr průmětů vzorků $\mathbf{x}^{(i)}$, kde $c^{(i)} = k$

minimalizace $J(\dots)$ vzhledem k $\mu_1, \mu_2, \dots, \mu_K$

- pohled na algoritmus jako na minimalizaci cenové funkce dovoluje provádět „debugging“, tj. sledovat, zda algoritmus konverguje



Výběr centroidů

Inicializace algoritmu



Inicializace:

- náhodně zinicializujeme K centroidů $\mu_1, \mu_2, \dots, \mu_K \in \mathbb{R}^n$

JAK?

Iterace:

do {

 for $i \leftarrow 1 \dots m$

 ① $c^{(i)} \leftarrow$ index k centroidu nejblíže $\mathbf{x}^{(i)}$, tj. $\min_k \|\mathbf{x}^{(i)} - \mu_k\|^2$

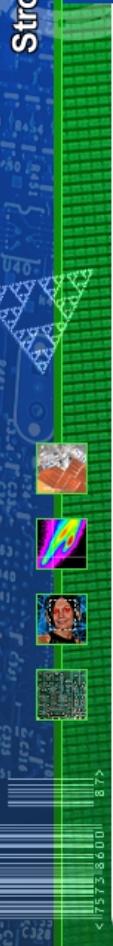
 ② for $k \leftarrow 1 \dots K$

$\mu_k \leftarrow$ průměr průmětů vzorků $\mathbf{x}^{(i)}$, kde $c^{(i)} = k$

}

Vhodný počáteční výběr centroidů může konvergenci algoritmu dramaticky urychlit... **optimální volba zřejmě není možná**

→ běžně používaný způsob výběru centroidů





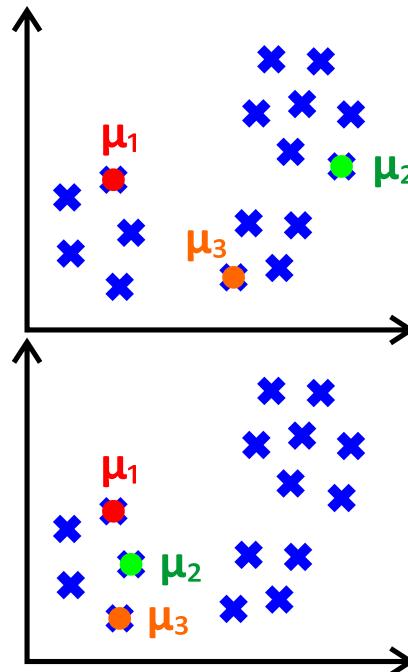
Výběr centroidů

„Náhodná“ inicializace

„Náhodnost“ nespočívá v libovolném, tj. **náhodném, umístění centroidu do příznakového prostoru**, ale ve výběru náhodného vzorku z trénovací množiny jako centroidu...

- počet shluků $K \ll m$
- **náhodně vybereme K vzorků z TM** (klidně funkci `rand()` s rovnoměrnou distribucí)
- centroidy nastavíme shodně s těmito K vybranými vzorky

$$\mu_1 = \mathbf{x}^{(i)}, \mu_2 = \mathbf{x}^{(j)}, i \neq j$$



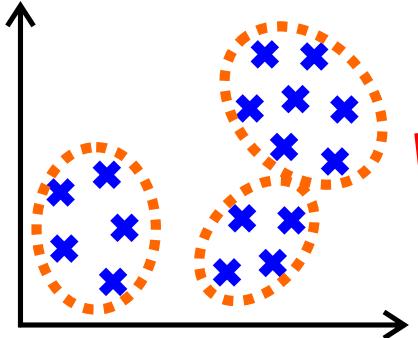
v závislosti na inicializaci může algoritmus poskytnout různé výsledky



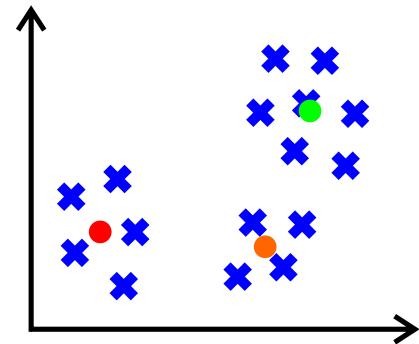
Výběr centroidů

Uvíznutí v lokálním optimu

Trénovací množina:

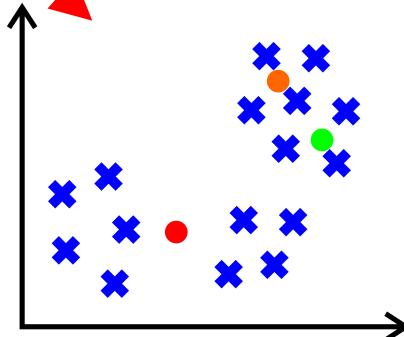
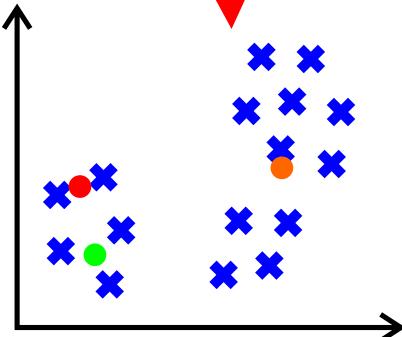


globální optimum



Lokální optima:

- lokální extrémy funkce zkreslení $J(\dots)$



Jak najít globální optimum?



Výběr centroidů

Nalezení globálního optima

```
for i ← 1 .. 100 {  
    náhodná inicializace  
     $c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K \leftarrow \text{K-means}(\dots)$   
     $j^{(i)} \leftarrow J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K)$   
}
```

vybereme to shlukování, pro které je hodnota zkreslení $j^{(i)}$ **nejmenší**

- tento postup funguje nejlépe při relativně malém K , tj. počet shluků je zhruba mezi 2 a 10 (v takovém případě poskytuje tento postup nejmarkantnější zlepšení oproti variantě s jednou náhodnou inicializací)
- počet iterací se obvykle pohybuje od 50 do 1000

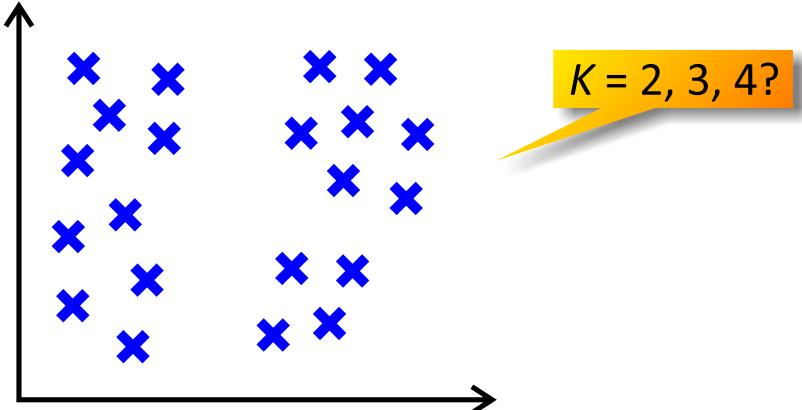




Volba počtu shluků

Strategie

- počet shluků K lze vybrat automaticky (viz dále)
- tradiční a v praxi velmi často používaný postup je ovšem **manuální volba** počtu shluků – na základě zejména nějaké analýzy dat, vizualizace, atp.
- počet shluků bohužel často není zřejmý (ani při expertní analýze)

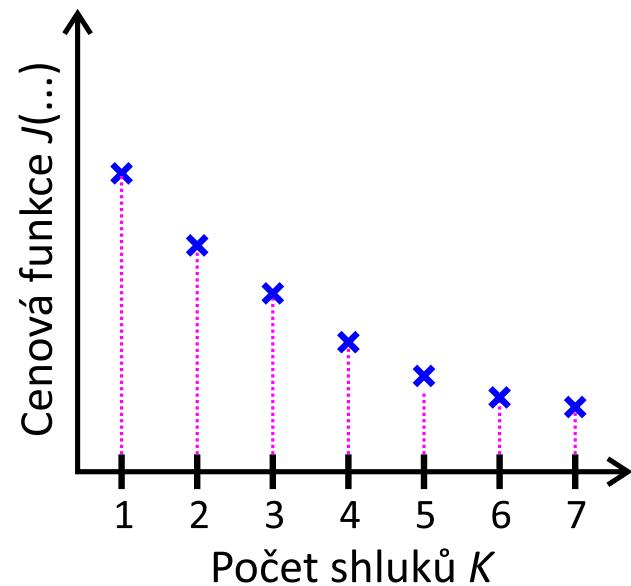
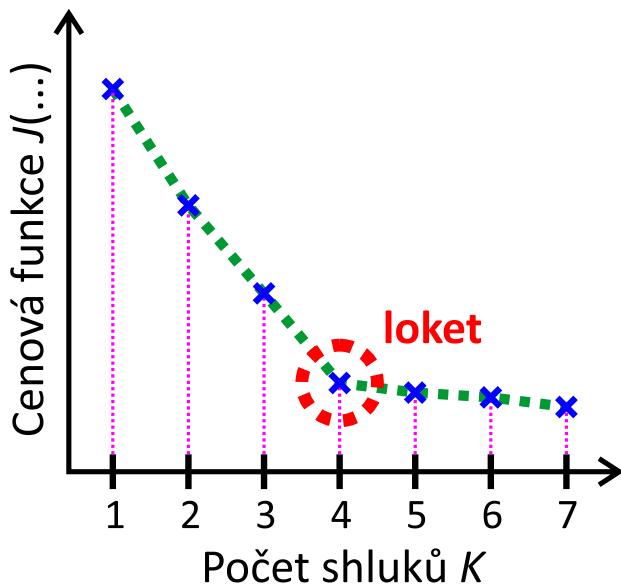




Volba počtu shluků

Metoda lokte (*Elbow Method*)

Vyjádření zkreslení $J(\dots)$ v závislosti na počtu shluků – pak se vybere místo „zalomení“ (loket) této závislosti...



Někdy ovšem tato závislost žádný loket nemá...



Volba počtu shluků

Metoda vyhovění účelu (*Downstream Purpose*)

Algoritmus K-means je obvykle použit k nějakému konkrétnímu účelu (např. segmentace trhu) – počet shluků K pak může vycházet/být jednoznačně určen charakterem tohoto účelu...

segmentace trhu k určení velikosti oblečení

