



EVROPSKÁ UNIE  
Evropské strukturální a investiční fondy  
Operační program Výzkum, vývoj a vzdělávání



MINISTERSTVO ŠKOLSTVÍ,  
MLÁDEŽE A TĚLOVÝCHOVY

Název projektu	Rozvoj vzdělávání na Slezské univerzitě v Opavě
Registrační číslo projektu	CZ.02.2.69/0.0./0.0/16_015/0002400

## Dolování dat

### Statistika v kontextu dolování dat

Jan Górecki



**SLEZSKÁ  
UNIVERZITA**  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ

# Obsah přednášky

---

- Typy statistických metod
  - Kontingenční tabulky
  - Regresní analýza
  - Diskriminační analýza
  - Shluková analýza
- 





- A formal science that deals with collection, analysis, interpretation, explanation and presentation of (usually numerical) data.
-



- **Deskripční** – cílem je popsat základní charakteristiky daných dat
  - **Konfirmační** – cílem je potvrdit resp. vyvrátit zkoumanou hypotézu
  - **Explorační** – cílem je “objevit” možnou hypotézu, která je podporovaná daty
-

# Kontingenční tabulky



- zjišťování vztahu mezi dvěma kategoriálními veličinami

čtyřpolní tabulka

	Úvěr ano	Úvěr ne	$\Sigma$
Vysoký příjem	$a_{11}$	$a_{12}$	$r_1$
Nízký příjem	$a_{21}$	$a_{22}$	$r_2$
$\Sigma$	$s_1$	$s_2$	$n$

# Kontingenční tabulky

---



SLEZSKÁ  
UNIVERZITA  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ

$\chi^2$  test:

$$\chi^2 = \sum_{i=1}^R \sum_{j=1}^S \frac{(a_{ij} - o_{ij})^2}{o_{ij}} = n \times \sum_{i=1}^R \sum_{j=1}^S \frac{\left( a_{ij} - \frac{r_i \cdot s_j}{n} \right)^2}{r_i \cdot s_j}$$

pro  $\chi^2 \geq \chi^2_{(R-1)(S-1)}(\alpha)$  předpokládáme  
závislost mezi X a Y

---

# Regresní analýza



SLEZSKÁ  
UNIVERZITA  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ

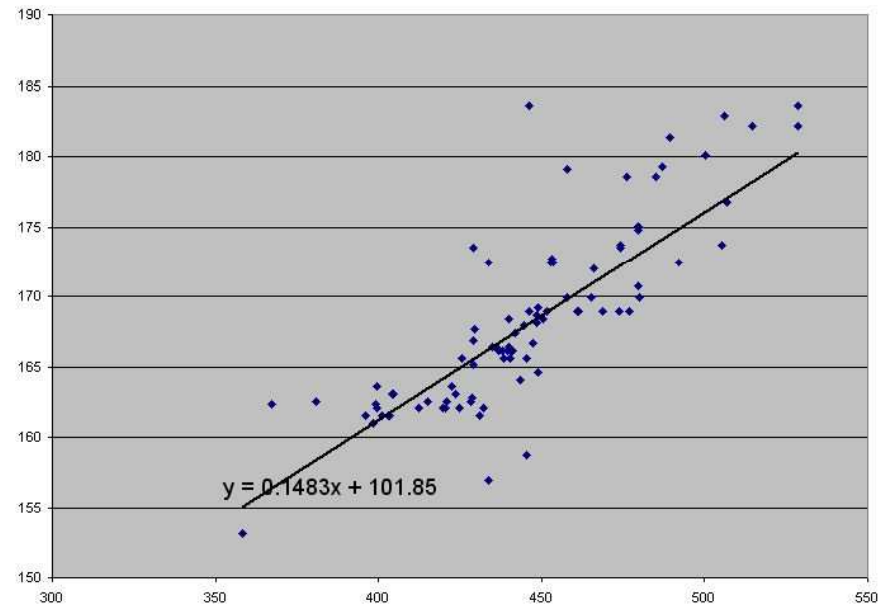
- zjišťování funkční závislosti jedné numerické (spojité) veličiny na jiných numerických veličinách

lineární regrese pro dvě veličiny  $x$  a  $y$ :

$$y = \beta_1 x + \beta_0 + \varepsilon.$$

Hodnoty koeficientů ( $\beta_1$  a  $\beta_0$ ) se zjišťují pomocí:

**Metoda nejmenších čtverců**

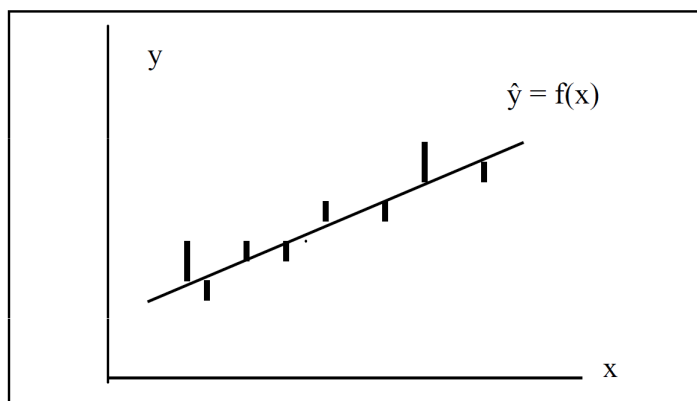


# Metoda nejmenších čtverců



SLEZSKÁ  
UNIVERZITA  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ

- Tato metoda minimalizuje rozdíly mezi pozorovanou hodnotou  $y$  a očekávanou hodnotou  $\hat{y}=f(x)$  spočítanou v tomto případě na základě funkce  $\beta_1x + \beta_0$



- uvažujeme druhou mocninu (kvadrát, čtverec) těchto rozdílů:

$$(y - f(x))^2$$

Úlohu pro  $n$  pozorování lze tedy formálně zapsat jako hledání

$$\min \sum_{i=1}^n (y_i - f(x_i))^2$$



# Diskriminační analýza

---



SLEZSKÁ  
UNIVERZITA  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ

- pro odlišení příkladů patřících do různých tříd
- ke každé třídě (hodnotě nominální veličiny)  $c_j$ ,  $j=1, \dots, R$  existuje (diskriminační) funkce  $f_j$  taková, že

$$f_j(\mathbf{x}) = \max_i f_i(\mathbf{x})$$

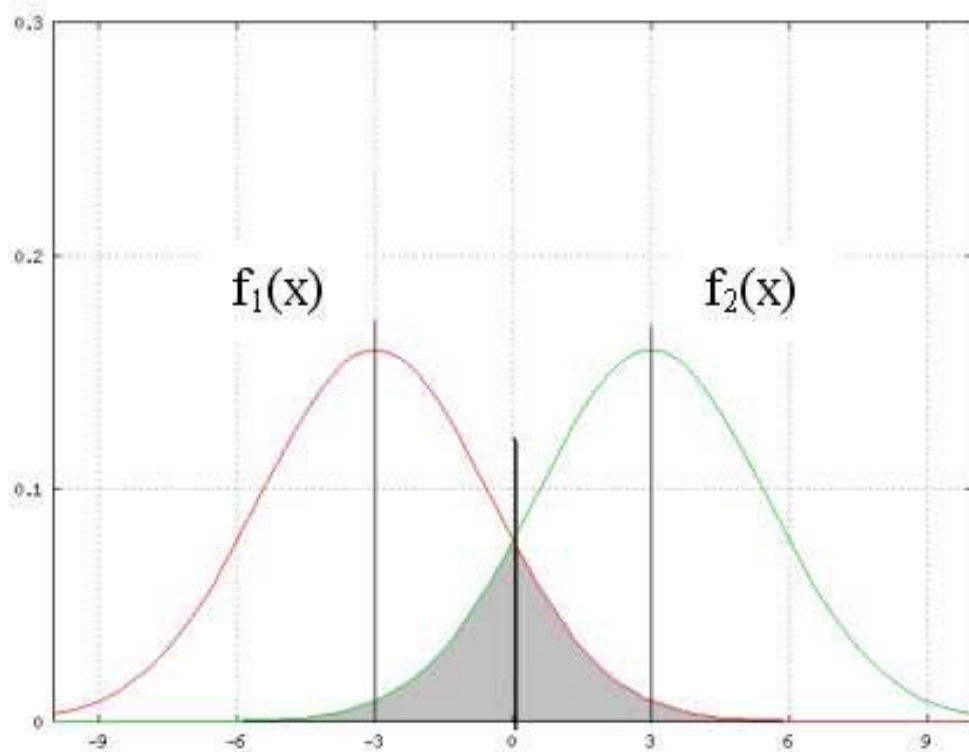
právě když příklad  $\mathbf{x}=[x_1, x_2, \dots, x_v]$  patří do třídy  $c_j$ .

---

# $f_1$ i $f_2$ stejný rozptyl



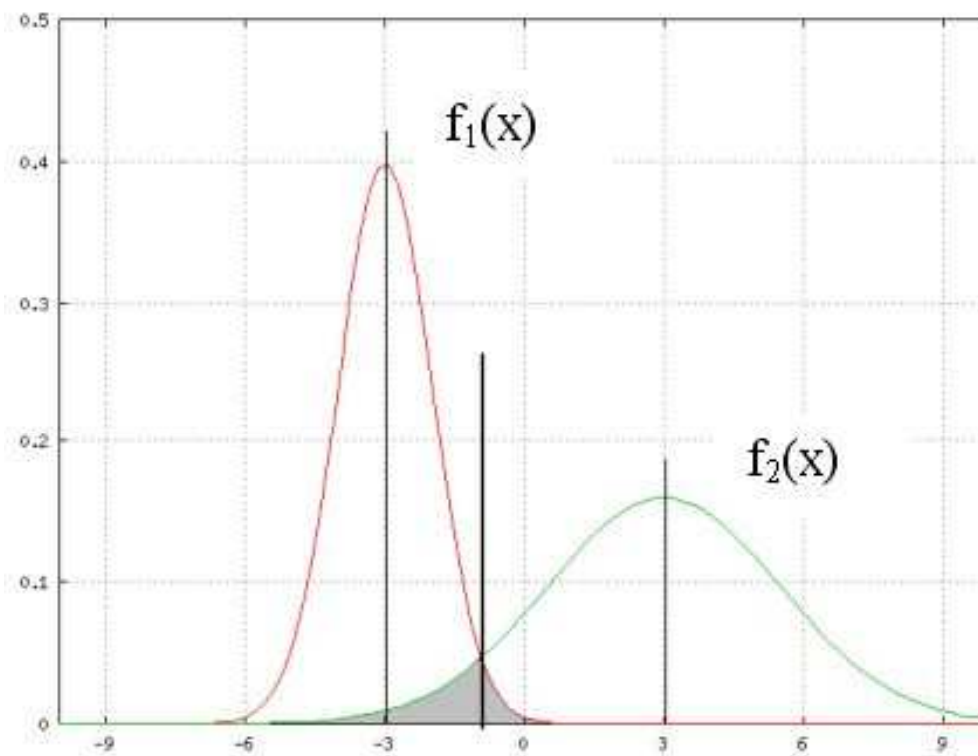
SLEZSKÁ  
UNIVERZITA  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ



# $f_1$ a $f_2$ různý rozptyl



SLEZSKÁ  
UNIVERZITA  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ



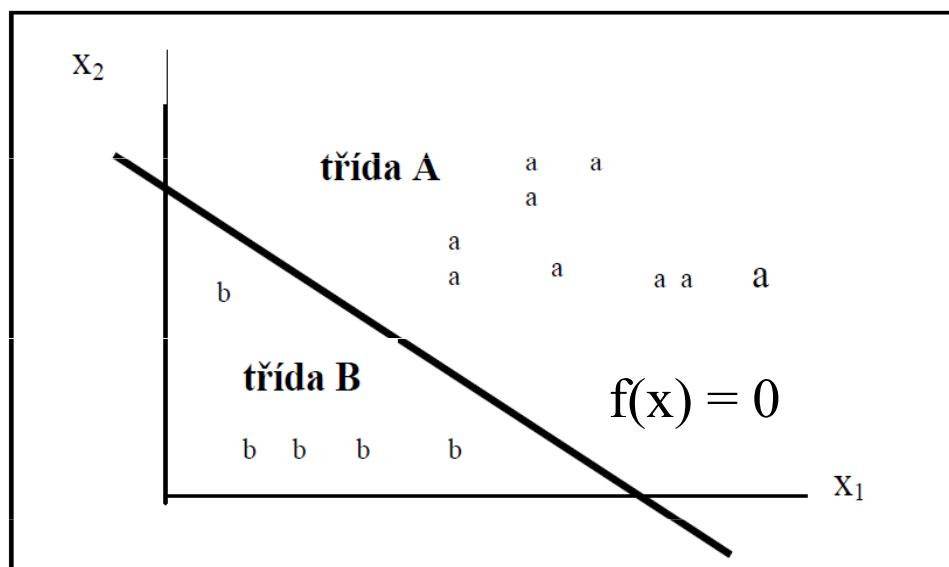
# Lineární diskriminace do dvou tříd



$$f_j = q_{0j} + q_{1j}x_1 + q_{2j}x_2 + \dots + q_{vj}x_v$$

$j=1,2$

$$f(x) = f_1(x) - f_2(x).$$



# Lineární a kvadratická diskriminační analýza

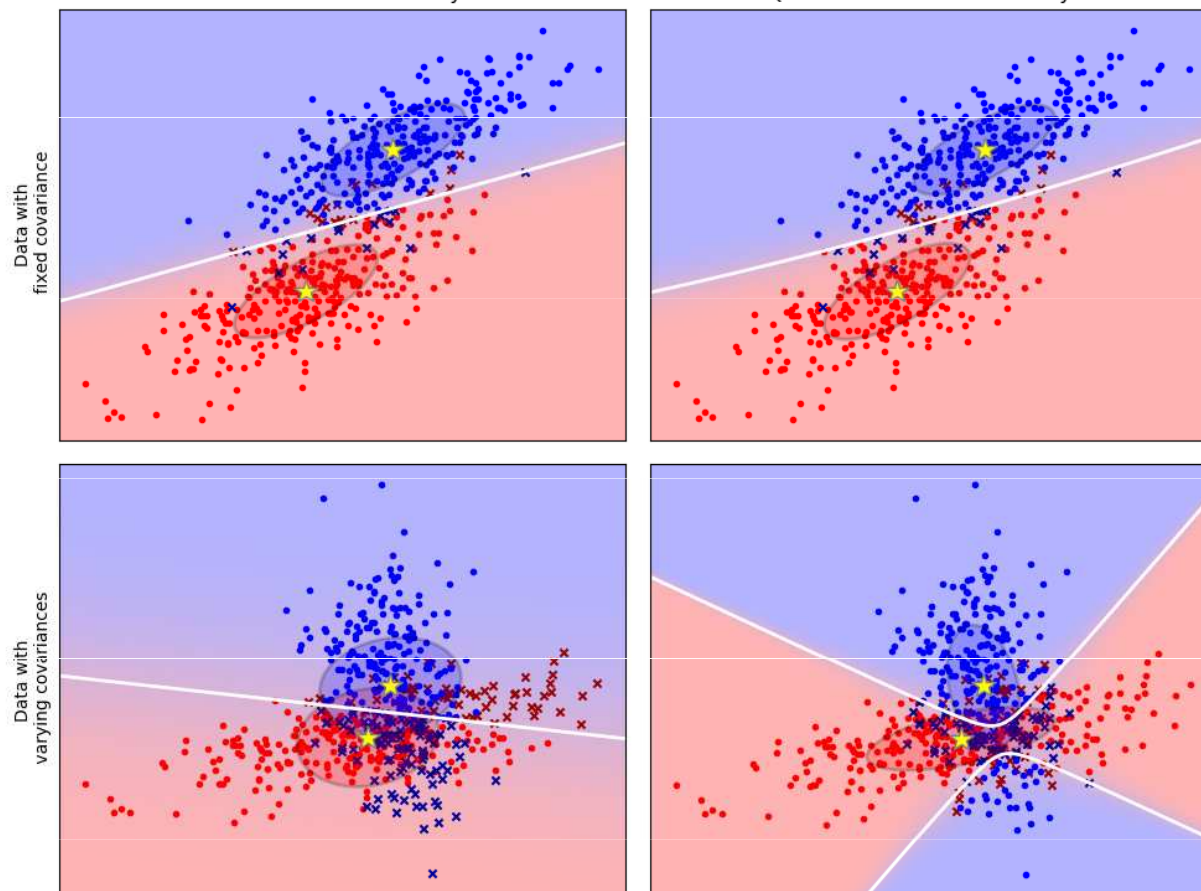


**SLEZSKÁ  
UNIVERZITA**  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ

## Linear Discriminant Analysis vs Quadratic Discriminant Analysis

Linear Discriminant Analysis

Quadratic Discriminant Analysis



# Shluková analýza

---



- slouží pro nalezení skupin (shluků) navzájem si podobných příkladů
- Např. dva příklady  $\mathbf{x}_1 = [x_{11}, \dots, x_{1m}]$  a  $\mathbf{x}_2 = [x_{21}, \dots, x_{2m}]$

Eukleidovská vzdálenost

$$d_E(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^m (x_{1j} - x_{2j})^2}$$

# Hammingova vzdálenost

---



SLEZSKÁ  
UNIVERZITA  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ

$$d_H(\mathbf{x}_1, \mathbf{x}_2) = \sum_{j=1}^m |x_{1j} - x_{2j}|$$

# Čebyševova vzdálenost

---



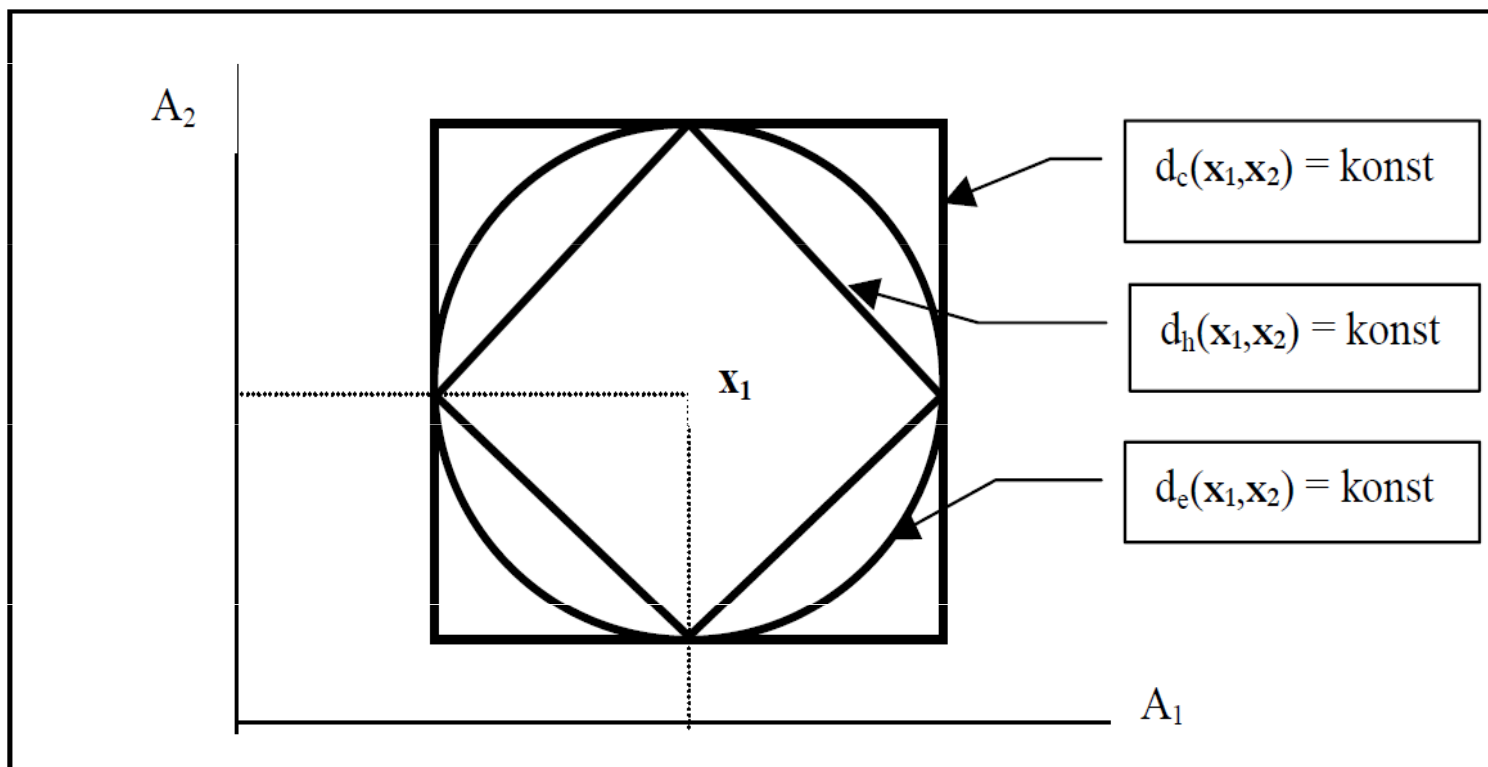
SLEZSKÁ  
UNIVERZITA  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ

$$d_C(\mathbf{x}_1, \mathbf{x}_2) = \max_j |x_{1j} - x_{2j}|$$

---



# Rozdíl mezi $d_H(x_1, x_2)$ , $d_E(x_1, x_2)$ a $d_C(x_1, x_2)$





- Výše uvedené míry vzdálenosti závisí na měřítku veličin. Proto je třeba veličiny normovat
  - Konkrétní hodnota se obvykle dělí nějakou jinou hodnotou:
    - směrodatnou odchylkou
    - rozpětím (max-min).
-



- hierarchické shlukování,
  - metoda *K*-středů (*K*-means clustering).
-

# Hierarchické shlukování



Při hierarchickém shlukování se obvykle postupuje metodou „zdola nahoru“. Začíná se tedy v situaci, kdy každý příklad tvoří jeden samostatný shluk. Postupně se pak jednotlivé shluky spojují, až skončíme s jedním shlukem obsahujícím všechny příklady

## Algoritmus hierarchického shlukování

### Inicializace

1. urči vzájemné vzdálenosti mezi všemi příklady
2. zařaď každý příklad do samostatného shluku

### Hlavní cyklus

1. dokud je více než jeden shluk
  - 1.1. najdi dva navzájem nejbližší shluky a spoj je
  - 1.2. spočítej pro tento nový shluk jeho vzdálenost od ostatních shluků

# Vzdálenost mezi shluky



- *metoda nejbližšího souseda* - vzdálenost mezi shluky  $U$  a  $V$  je dána minimem ze vzdálenosti mezi jejich příklady

$$D(U, V) = \min_{k,l} d(\mathbf{x}_k, \mathbf{x}_l), \mathbf{x}_k \in U, \mathbf{x}_l \in V$$

- *metoda nejvzdálenějšího souseda* - vzdálenost mezi shluky  $U$  a  $V$  je dána maximem ze vzdálenosti mezi jejich příklady

$$D(U, V) = \max_{k,l} d(\mathbf{x}_k, \mathbf{x}_l), \mathbf{x}_k \in U, \mathbf{x}_l \in V$$

- *metoda průměrné vzdálenosti* - vzdálenost mezi shluky  $U$  a  $V$  je dána průměrem ze vzdálenosti mezi jejich příklady ( $n_U$  je počet příkladů ve shluku  $U$  a  $n_V$  je počet příkladů ve shluku  $V$ )

$$D(U, V) = \frac{1}{n_U n_V} \sum_{k=1}^{n_U} \sum_{l=1}^{n_V} d(\mathbf{x}_k, \mathbf{x}_l)$$

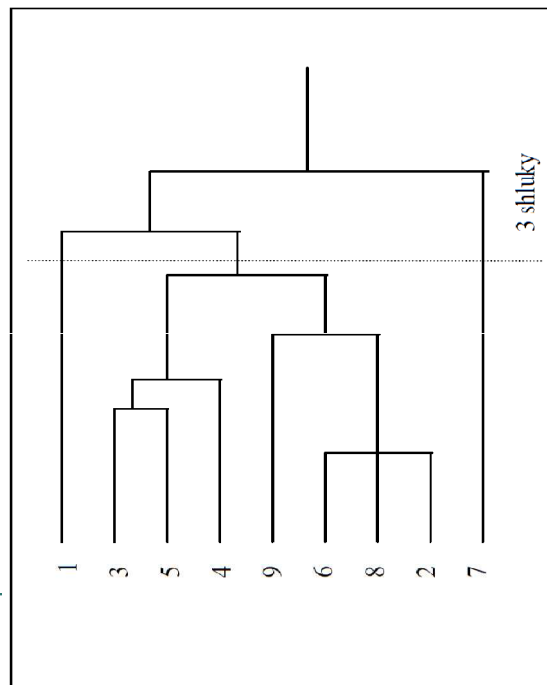
- *centroidní metoda* - vzdálenost mezi shluky  $U$  a  $V$  je dána vzdáleností mezi středy shluků.

$$D(U, V) = d(\mathbf{u}, \mathbf{v}), \mathbf{u} \text{ je střed shluku } U \text{ a } \mathbf{v} \text{ je střed shluku } V$$

# Dendrogram



- Proces hierarchického shlukování bývá zachycen v podobě tzv. dendrogramu. Ten ukazuje (odspoda nahoru) postupné spojování shluků počínaje očíslovanými příklady. Optimální počet shluků zde není předem znám, odvodíme ho až rozbořením výsledků – tak, že někde dendrogram „rozřízneme“



# Metoda $K$ -středů

---



SLEZSKÁ  
UNIVERZITA  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ

1. urči centroidy pro všechny shluky v aktuálním rozkladu (v prvním opakování zcela náhodně)

2. pro každý příklad  $\mathbf{x}$

2.1. urči vzdálenosti  $d(\mathbf{x}, \mathbf{c}_k)$ ,  $k=1, \dots, K$  kde  $\mathbf{c}_k$  je centroid  $k$ -tého shluku

2.2. urči centroid  $\mathbf{c}_l$  tak, že  $d(\mathbf{x}, \mathbf{c}_l) = \min_k d(\mathbf{x}, \mathbf{c}_k)$

2.3. není-li  $\mathbf{x}$  součástí shluku  $l$  (k jehož centroidu  $\mathbf{c}_l$  má nejbližší) přesuň  $\mathbf{x}$  do shluku  $l$

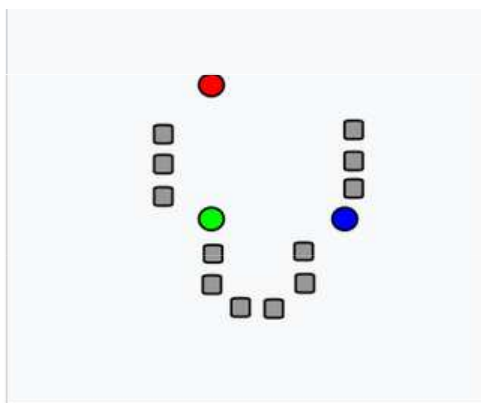
3. došlo-li k nějakému přesunu potom jdi na 1, jinak konec

---

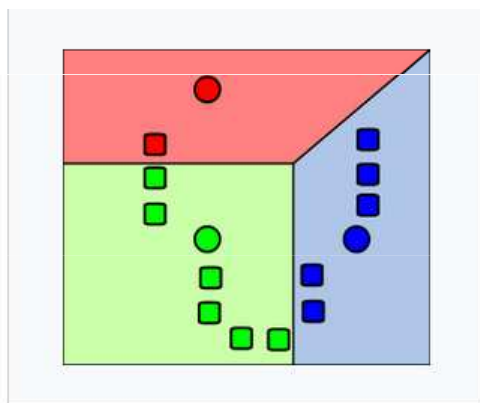
# Ukázka algoritmu K-středů



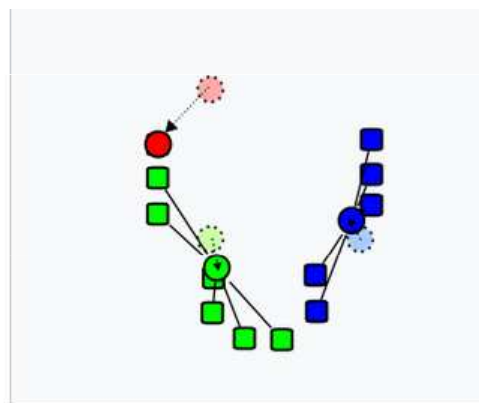
SLEZSKÁ  
UNIVERZITA  
OBCHODNĚ PODNIKATELSKÁ  
FAKULTA V KARVINĚ



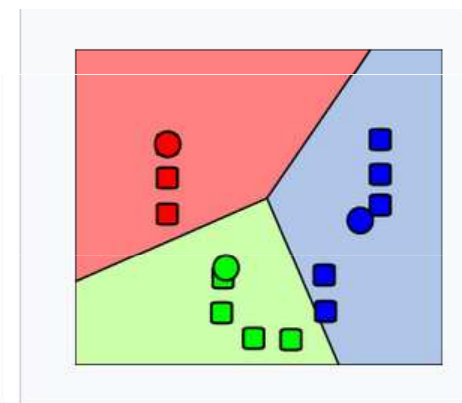
1.  $k$  výchozích centroidů (zde je  $k=3$ ) se náhodně umístí v prostoru dat (shlukované objekty šedé, centroidy barevné)



2. Objekty se přiřadí nejbližším centroidům, čímž vznikne  $k$  shluků. Centroidy tak definují [Voroného teselaci](#) prostoru.



3. Přepočtou se centroidy shluků tak, aby šlo o těžiště objektů, jež patří do těchto shluků.



4. Kroky 2 a 3 se opakují, dokud nedojde k ustálení ([konvergence](#)).



# Děkuji za pozornost

Některé snímky převzaty od:  
prof. Ing. Petr Berka, CSc. [berka@vse.cz](mailto:berka@vse.cz)