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 KARVINÉ

Obsah přednášky

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

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

Statistika a její metody

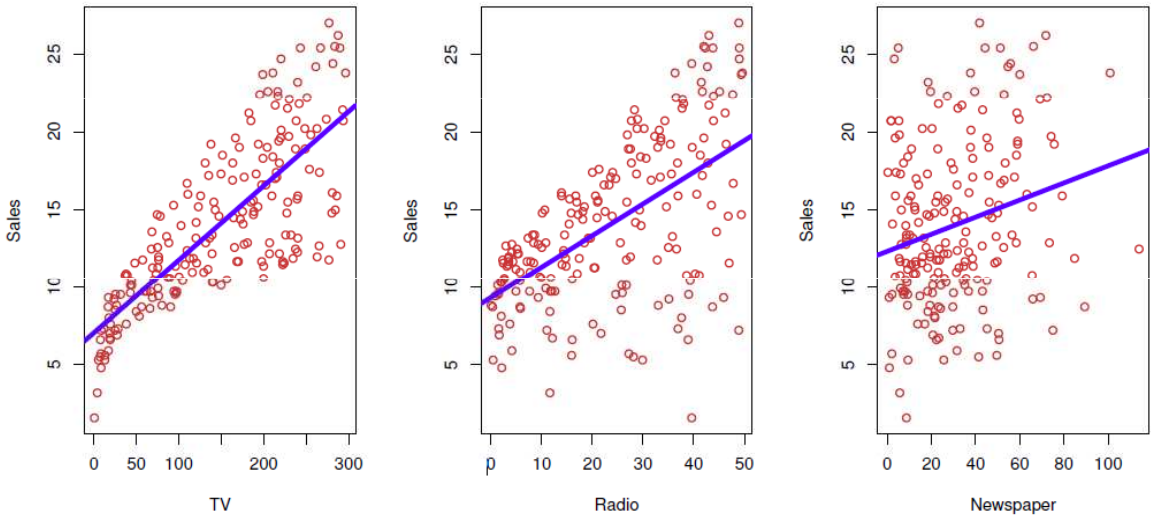
SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

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

Metody:

- **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

Motivace

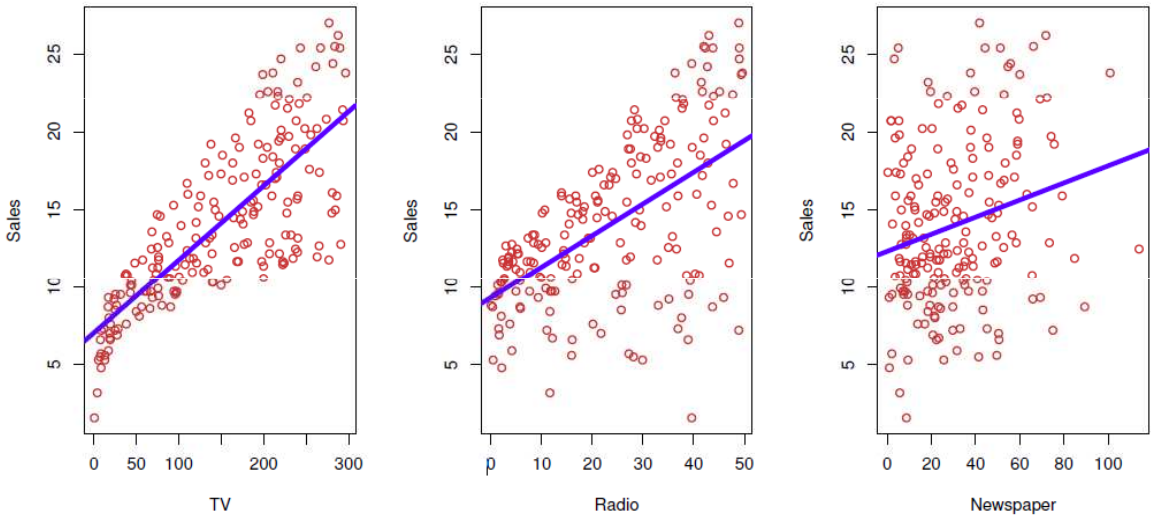


SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Sales – prodej produktu v tisících
TV, Radio, Newspaper - rozpočet na reklamu v daném médiu

Cíl: Na základě těchto údajů navrhněte marketingový plán na příští rok, který povede k vysokým prodejům produktů.
Otázka: Jaké informace by byly užitečné pro poskytnutí takového doporučení?

Motivace



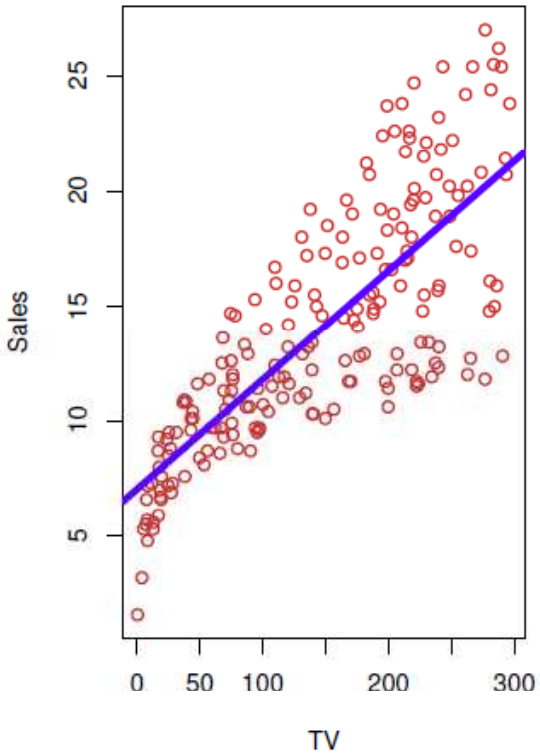
1. Existuje vztah mezi rozpočtem na reklamu a prodejem?
2. Jak silný je vztah mezi rozpočtem na reklamu a prodejem?
3. Která média jsou spojena s prodejem?
4. Jaký je vztah mezi každým médiem a prodejem?
5. Jak přesně můžeme předpovídat budoucí prodeje?
6. Je vztah lineární?
7. Existuje synergie mezi reklamními médii?

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Lineární regrese

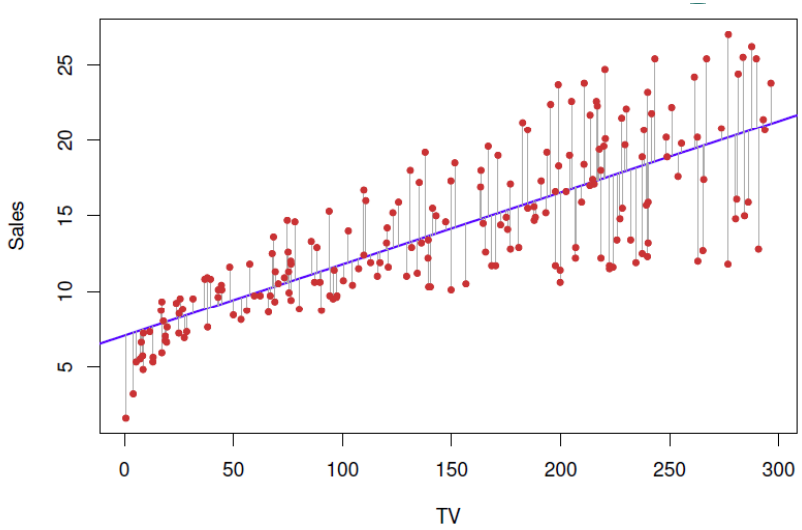
SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

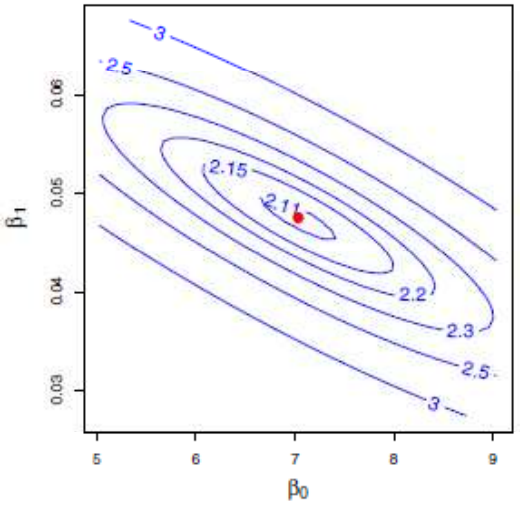
1. **Definice:**
   - Jednoduchá lineární regrese je statistická metoda, která zkoumá lineární vztah mezi dvěma kvantitativními proměnnými.
2. **Matematický Model:**
   - $y = \beta_0 + \beta_1 x + \epsilon$
   - kde:
     - y je závislá proměnná
     - x je nezávislá proměnná
     - β_0 je y-intercept (parameter modelu)
     - β_1 je sklon regresní čáry (parameter modelu)
     - ϵ je chyba
3. **Příklad:**
   - Analýza vztahu mezi rozpočtem na reklamu (x) a prodejem produktu (y).
4. **Význam:**
   - Pomáhá předpovídat hodnotu závislé proměnné na základě hodnoty nezávislé proměnné.
   - Umožňuje odhadnout, jak se změna v hodnotě nezávislé proměnné ovlivní závislou proměnnou.

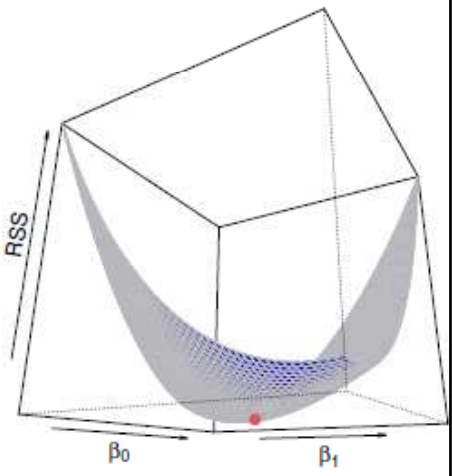


Odhad parametrů modelu

1. **Cíl:**
   - Odhad hodnot β_0 a β_1 tak, aby byla minimalizována chyba mezi skutečnými a předpovězenými hodnotami y.
2. **Metoda Nejmenších Čtverců:**
   - Minimalizuje součet čtverců reziduí RSS = $e_1^2 + e_2^2 + \ldots + e_n^2$, kde $e_i = y_i - (\beta_0 + \beta_1 x_i)$ je rozdíl mezi skutečnou (y_i) a předpovězenou hodnotou $(\beta_0 + \beta_1 x_i)$ pro i-tý bod. (RSS = Residual sum of squares)
   - Výpočet:
     - $\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$
     - $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
3. **Příklad s Daty:**
   - Použití dat o prodeji a reklamním rozpočtu k odhadu parametrů modelu lineární regrese.
4. **Interpretace:**
   - Jak interpretovat odhadnuté parametry a jaký mají význam pro předpovědění prodeje na základě rozpočtu na reklamu.



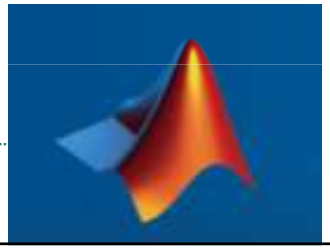


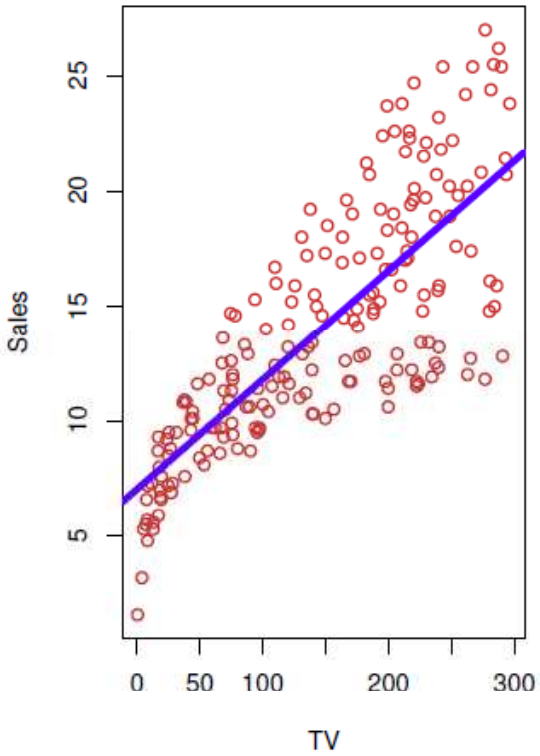


Odpovědi na 7 otázek

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

1. **Existuje vztah mezi rozpočtem na reklamu a prodejem?**
   - Analýza korelace a regresní analýza k identifikaci vztahu.
2. **Jak silný je vztah mezi rozpočtem na reklamu a prodejem?**
   - Hodnocení koeficientu determinace (R^2) k měření síly vztahu (R^2 odpovídá kvadrátu korelace X a Y).
3. **Která média jsou spojena s prodejem?**
   - Výsledky analýzy pro TV, Radio a Newspaper.
4. **Jaký je vztah mezi každým médiem a prodejem?**
   - Hodnoty koeficientů regrese pro každé médium.
5. **Jak přesně můžeme předpovídat budoucí prodeje?**
   - Model predikce s odhadnutými parametry a intervaly spolehlivosti.
6. **Je vztah lineární?**
   - Grafická a statistická analýza pro ověření linearity vztahu.
7. **Existuje synergie mezi reklamními médii?**
   - Vícerozměrná analýza interakcí mezi různými médii.





Kontingenční tabulky

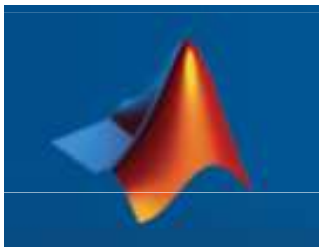
SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- zjišťování vztahu mezi dvěma **kategoriálními** veličinami

čtyřpolní tabulka

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

příjem	úvěr
vysoký	ano
vysoký	ano
nízký	ne
nízký	ano
nízký	ano
nízký	ne
vysoký	ano
vysoký	ano
nízký	ne
vysoký	ano
nízký	ne
nízký	ano



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

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

o_{ij} …očekávané množství při platnosti hypotézy o nezávislosti veličin

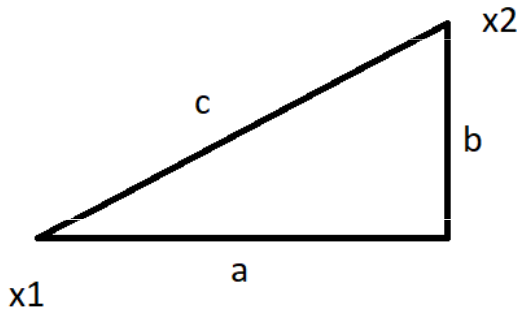
Shluková analýza

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- slouží pro nalezení skupin (shluků) navzájem si podobných příkladů
- Např. dva příklady $\boldsymbol{x_1} = [x_{11},\ldots,x_{1m}]$ a $\boldsymbol{x_2} = [x_{21},\ldots,x_{2m}]$

Eukleidovská vzdálenost

$$d_E(\boldsymbol{x_1},\boldsymbol{x_2}) = \sqrt{\sum_{j=1}^{m}(x_{1j} - x_{2j})^2}$$



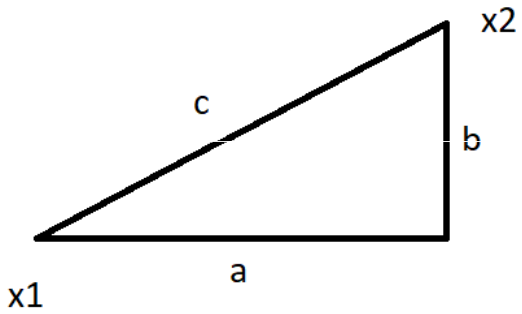
$d_E(\boldsymbol{x_1},\boldsymbol{x_2}) = c$

Jaká je tato vzdálenost pro body x1 = [1, 2] a x2 = [4,6]?

Minkovského (Manhattan) vzdálenost

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

$$d_H(\mathbf{x_1},\mathbf{x_2}) = \sum_{j=1}^{m} |x_{1j} - x_{2j}|$$



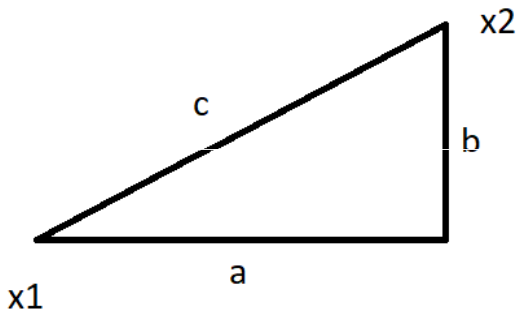
$$d_E(\boldsymbol{x_1},\boldsymbol{x_2}) = a+b$$

Jaká je tato vzdálenost pro body x1 = [1, 2] a x2 = [4,6]?

Čebyševova vzdálenost

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

$$d_C(\mathbf{x_1},\mathbf{x_2}) = \max_j |x_{1j} - x_{2j}|$$

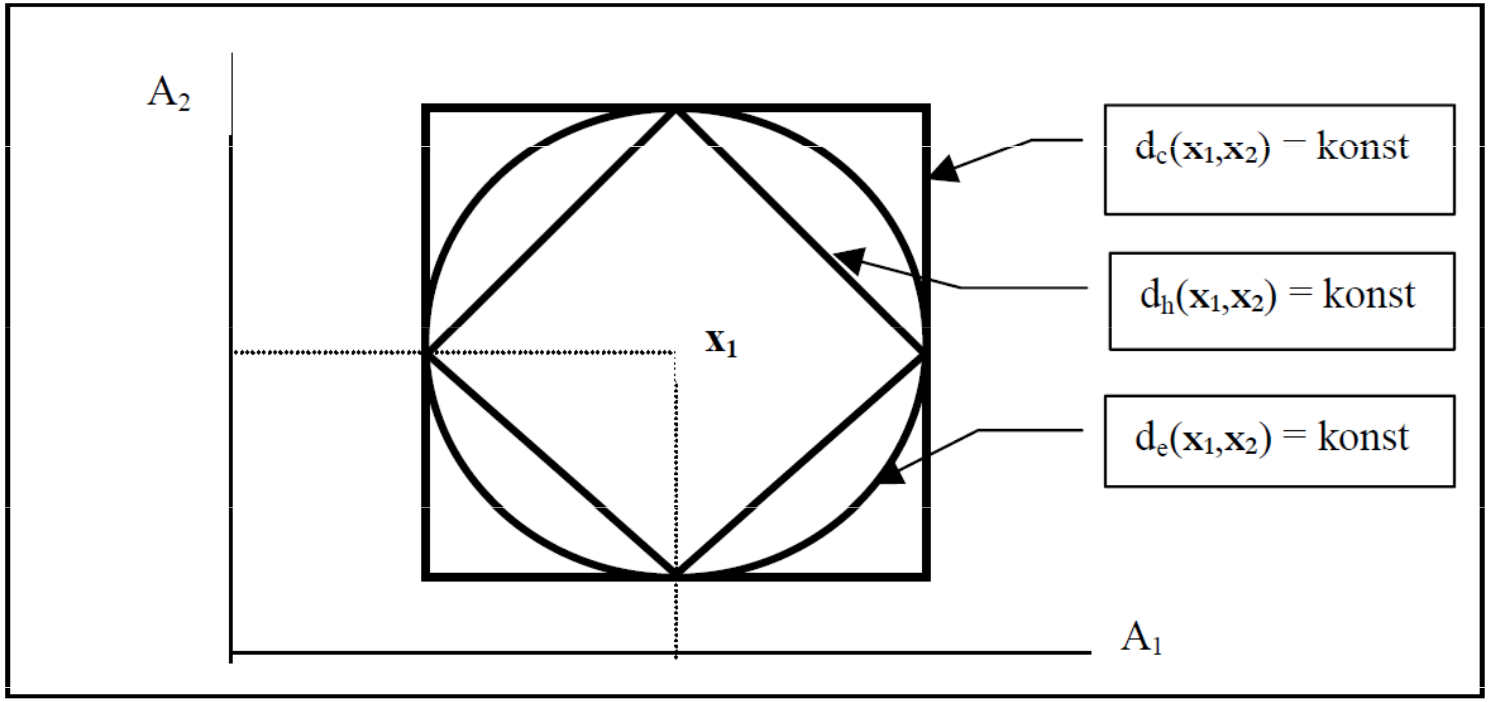


$d_E(\boldsymbol{x_1},\boldsymbol{x_2}) = \max(a,b) = a$

Jaká je tato vzdálenost pro body x1 = [1, 2] a x2 = [4,6]?

Rozdíl mezi $d_M(x_1,x_2)$, $d_E(x_1,x_2)$ a $d_C(x_1,x_2)$ ve 2D

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ



Pozor: pro 1D všechny vzdálenosti splývají (dávají stejný výsledek)

Normování

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

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

Metody shlukové analýzy

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

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

Hierarchické shlukování

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

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 přiklad do samostatného shluku**

hlavní cyklus

1. dokud je vice než jeden shluk

1.1. najdi dva navzájem nejbližší shluky a spoj je

1.2. spočítej pro tento nový shluk jeho vzdálenost od ostatnich shluků

Vzdálenost mezi shluky

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

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

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

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

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

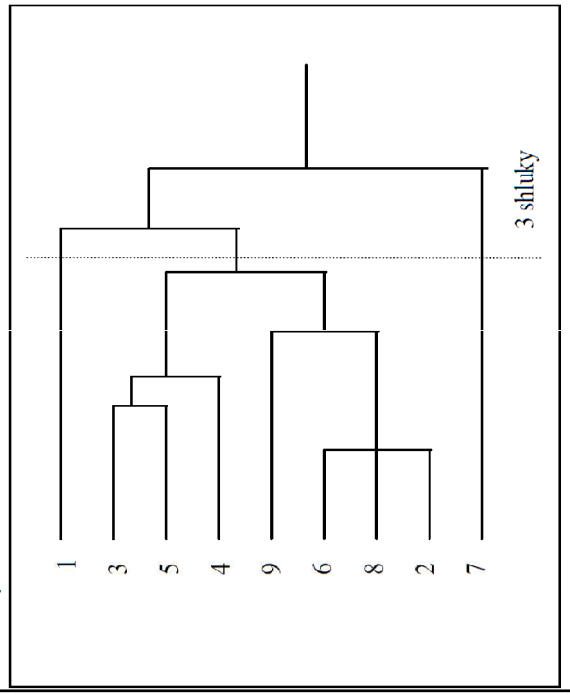
- *metoda průměrné vzdálenosti* - vzdálenost mezi shluky U a V je dána průměrem ze vzdálenosti mezi jejich příklady (n_U je počet příkladů ve shluku **U** a n_V je počet příkladů ve shluku **V**)

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

Dendrogram

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

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

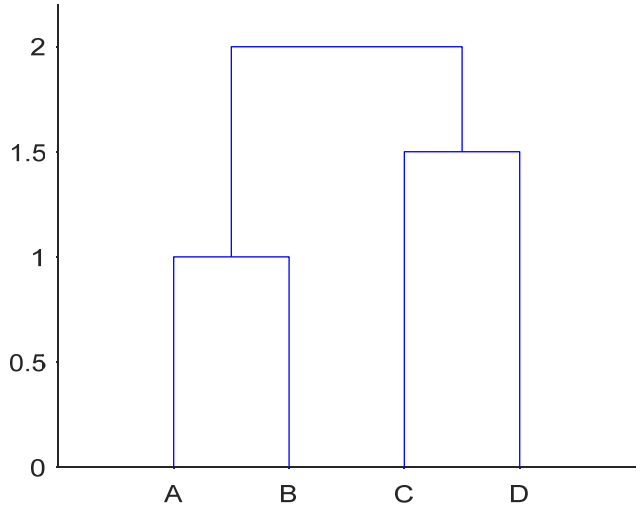


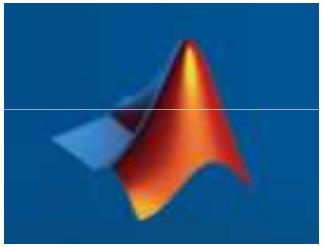
Příklad

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Jak proběhne hierarchické shlukování pro 4 jednorozměrné body
A = [0], B = [1], C = [3] a D = [4,5]
pro *eukleidovskou* vzdálenost a metodu *nejbližšího* souseda?

	A	B	C	D
A				
B				
C				
D				





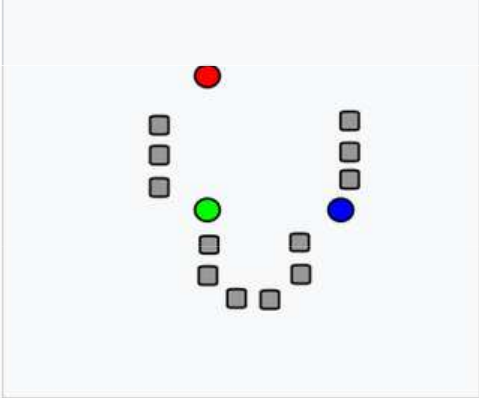
Metoda *K* –středů - Algoritmus

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

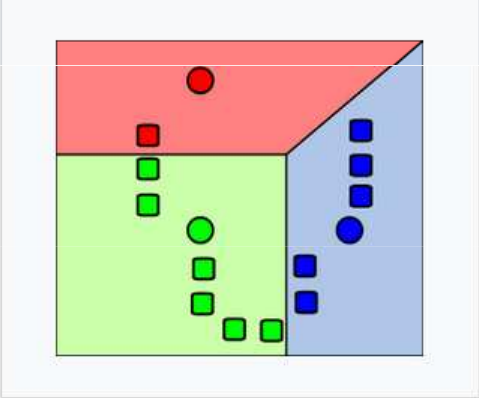
1. urči centroidy pro všechny shluky v aktuálním rozkladu (v prvním opakování zcela náhodně)
2. pro každý příklad **x**
   - 2.1. urči vzdálenosti $d(\mathbf{x},\mathbf{c}_k)$, k=1,…,K kde $\mathbf{c}_k$ je centroid *k*-tého shluku
   - 2.2. urči centroid $\mathbf{c}_l$ tak, že $d(\mathbf{x},\mathbf{c}_l) = \min_k d(\mathbf{x},\mathbf{c}_k)$
   - 2.3. není-li **x** součástí shluku *l* (k jehož centroidu $\mathbf{c}_l$ má nejblíže) přesuň **x** do shluku *l*
3. došlo-li k nějakému přesunu potom jdi na 1, jinak konec

Ukázka algoritmu K-středů

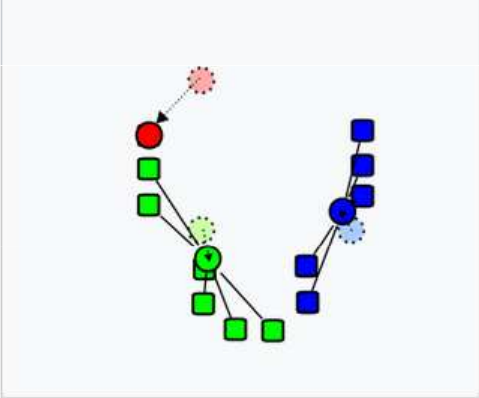
SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ



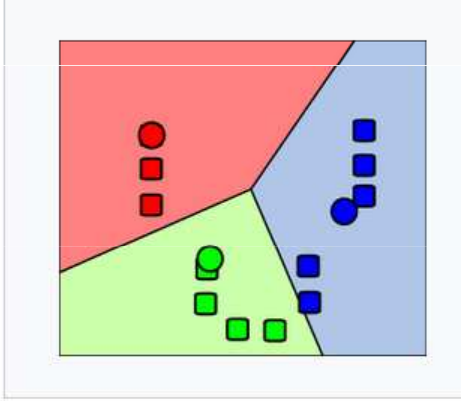
1. *k* výchozích centroidů (zde je *k*=3) se náhodně umístí v prostoru dat (shlukované objekty šedé, centroidy barevné)



2. Objekty se přiradí 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