Učení bez učitele

Matematické metody pro ITS (11MAMY)

Jan Přikryl

s využitím díla G. James et al., *An Introduction to Statistical Learning*

11. přednáška 11MAMY

středa 13. dubna 2022

verze: 2022-04-10 23:48

Ústav aplikované matematiky
ČVUT v Praze, Fakulta dopravní

Obsah přednášky

Učení bez učitele

- Shlukování

Učení bez učitele

Nesupervizované versus supervizované učení:

- Většina tohoto kurzu je zaměřena na metody učení s učitelem (supervizovaného učení), jako je regrese a klasifikace.
- V takové situaci pozorujeme jak soubor vlastností $X_1, X_2, \ldots, X_p$ každého objektu, tak rovněž odpověď nebo odezvu Y. Cílem pak je předpovídat Y pomocí $X_1, X_2, \ldots, X_p$.
- V této předášce se místo toho soustředíme na nesupervizované učení (učení bez učitele), kde pozorujeme pouze vlastnosti $X_1, X_2, \ldots, X_p$. Nezajímá nás předpovídání, protože nemáme přidruženou proměnnou odpovědi Y.

Cíle učení bez učitele

- Cílem je objevit zajímavé věci o měřeních: existuje nějaký informativní způsob, jak vizualizovat daná data? Můžeme mezi proměnnými nebo mezi pozorováními odhalit nějaké podskupiny?
- Probereme dvě metody:
  - *analýzu hlavních komponent*, nástroj používaný pro vizualizaci dat nebo předběžné zpracování dat před tím, než použijeme supervizované postupy, a
  - *shlukování*, širokou třídu metod k objevování neznámých podskupin v datech.

Učení bez učitele je obtížné

- Učení bez učitele je subjektivnější než učení s učitelem, protože zde analýza nemá jednoduchý cíl jako je předpověď odpovědi.
- Ale techniky učení bez učitele mají rostoucí význam v řadě oborů:
  - podskupiny pacientek s rakovinou prsu seskupené na základě měření jejich genové exprese,
  - skupiny kupujících charakterizované historií jejich prohlížení zboží a nákupů,
  - filmy seskupené podle hodnocení uděleného jejich diváky.

Další výhoda

- Je často snazší získat neoznačená data – z laboratorního přístroje nebo počítače – než označená data, která mohou vyžadovat lidský zásah.
- Tak je například obtížné automaticky posoudit celkové vyznění recenze filmu: je příznivá nebo ne?

Analýza hlavních komponent (PCA)

Analýza hlavních komponent

- **Analýza hlavních komponent** (PCA, angl. *Principal Component Analysis*) poskytuje nízkorozměrnou reprezentaci souboru dat. Stanovuje posloupnost lineárních kombinací proměnných, které mají maximální rozptyl a jsou navzájem nekorelovány.
- Kromě toho, že poskytuje odvozené proměnné k použití v úlohách supervizovaného učení, slouží PCA také jako nástroj pro vizualizaci dat.

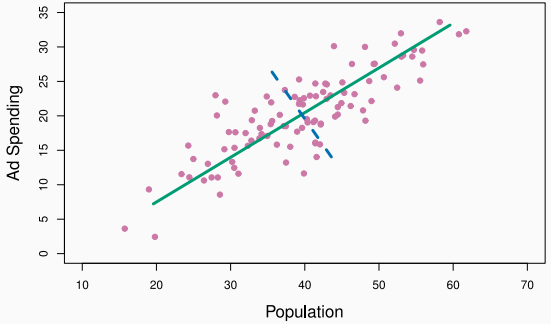
Analýza hlavních komponent: podrobnosti

- První hlavní komponenta souboru vlastností $X_1, X_2, \ldots, X_p$ je normalizovaná lineární kombinace těchto vlastností

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \cdots + \phi_{p1} X_p,$$

  která má největší rozptyl. Slovem normalizovaná rozumíme, že $\sum_{j=1}^{p} \phi_{j1}^2 = 1$.
- Na prvky $\phi_{11}, \ldots, \phi_{p1}$ odkazujeme jako na zátěže první hlavní komponenty; dohromady zátěže vytvářejí vektor zátěže dané hlavní komponenty, $\phi_1 = (\phi_{11}, \phi_{21}, \ldots, \phi_{p1})^{\mathsf{T}}$.
- Omezujeme zátěže tak, že jejich součet druhých mocnin je roven jedné, neboť pokud bychom jinak připustili, aby tyto prvky byly v absolutní hodnotě libovolně velké, mohlo by to vést k libovolně velkému rozptylu.

PCA: příklad



Velikost populace (pop) a náklady na reklamu (ad) pro 100 různých měst jsou zobrazeny jako purpurové kroužky. Zelená plná přímka označuje směr první hlavní komponenty, modrá čárkovaná přímka označuje směr druhé hlavní komponenty.

Výpočet hlavních komponent

- Předpokládejme, že máme soubor dat $\mathbf{X}$ o rozměrech $n \times p$. Protože nás zajímá pouze rozptyl, předpokládáme, že každá z proměnných v $\mathbf{X}$ byla vycentrována tak, že má střední hodnotu rovnou nule (to znamená, že sloupcové střední hodnoty $\mathbf{X}$ jsou nulové).
- Hledáme pak lineární kombinaci hodnot vlastností ve vzorku ve tvaru

$$z_{i1} = \phi_{11} x_{i1} + \phi_{21} x_{i2} + \cdots + \phi_{p1} x_{ip}$$

pro $i = 1, \ldots, n$, která má největší rozptyl ve vzorku za omezující podmínky, že $\sum_{j=1}^{p} \phi_{j1}^2 = 1$.

- Jelikož každé z x_{ij} má střední hodnotu nula, je tomu tak i pro z_{i1} (při libovolných hodnotách ϕ_{j1}). Rozptyl ve vzorku z_{i1} lze tudíž zapsat jako $1/n \sum_{i=1}^{n} z_{i1}^2$.

Výpočet: pokračování

- Vektor zátěže první hlavní komponenty dosazený do předchozí rovnice řeší úlohu optimalizace

$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximalizuj}} \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{j1}x_{ij}\right)^2 \text{ za podmínky } \sum_{j=1}^{p}\phi_{j1}^2 = 1.$$

- Tato úloha se dá řešit pomocí singulárního rozkladu matice $\mathbf{X}$, standardního postupu v lineární algebře.
- Na Z_1 se pak odkazujeme jako na *první hlavní komponentu* se získanými hodnotami $z_{11}, \ldots, z_{n1}$.

- Vektor zátěže ϕ_1 se složkami $\phi_{11}, \phi_{21}, \ldots, \phi_{p1}$ definuje v prostoru vlastností směr, ve kterém se data nejvíce mění.
- Jestliže promítneme n bodů dat $x_1, \ldots, x_n$ na tento směr, jsou hodnoty projekcí samotná skóre $z_{11}, \ldots, z_{n1}$ hlavní komponenty.

Další hlavní komponenty

- Druhá hlavní komponenta je lineární kombinace $X_1, \ldots, X_p$, která má maximální rozptyl mezi všemi lineárními kombinacemi, jež jsou nekorelované se Z_1.
- Skóre $z_{12}, z_{22}, \ldots, z_{n2}$ druhé hlavní komponenty mají tvar

$$z_{i2} = \phi_{12} x_{i1} + \phi_{22} x_{i2} + \cdots + \phi_{p2} x_{ip},$$

kde ϕ_2 je vektor zátěže druhé hlavní komponenty o složkách $\phi_{12}, \phi_{22}, \ldots, \phi_{p2}$.

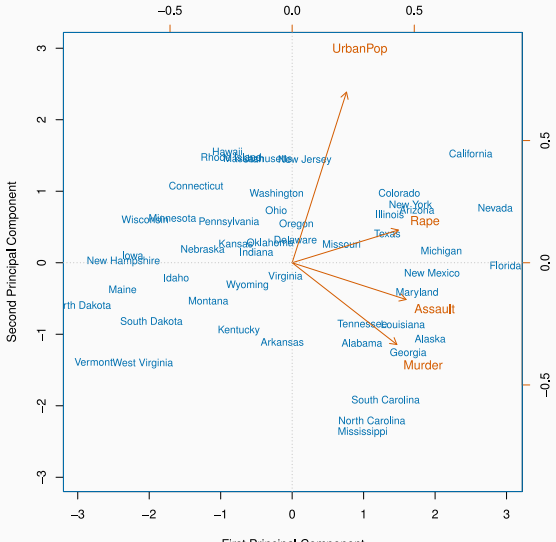
Další hlavní komponenty: pokračování

- Ukazuje se, že omezení na Z_2, aby bylo nekorelované se Z_1, je ekvivalentní omezení směru ϕ_2 tak, aby byl ortogonální (kolmý) ke směru ϕ_1. A tak dále.
- Směry hlavních komponent $\phi_1, \phi_2, \phi_3, \ldots$ jsou uspořádaná posloupnost pravých singulárních vektorů matice $\mathbf{X}$ a rozptyly složek jsou $1/n$ násobky kvadrátů singulárních čísel.
- Nejvýše může být $\min(n-1, p)$ hlavních komponent.

Ilustrace

- Data `USAarrests`: Pro každý z padesáti států ve Spojených státech soubor dat obsahuje počet zatčení na 100 000 obyvatel za každý ze tří zločinů: `Assault`, `Murder` a `Rape` (přepadení, vražda a znásilnění). Zaznamenáváme rovněž hodnoty `UrbanPop` (procento obyvatel každého státu žijících v městských oblastech).
- Vektory skóre hlavních komponent mají délku $n = 50$ a vektory zátěže hlavních komponent mají délku $p = 4$.
- PCA byla provedena po normalizaci každé proměnné tak, aby měla střední hodnotu nula a směrodatnou odchylku jedna.

Data USAarrests: graf PCA



Podrobnosti k obrázku

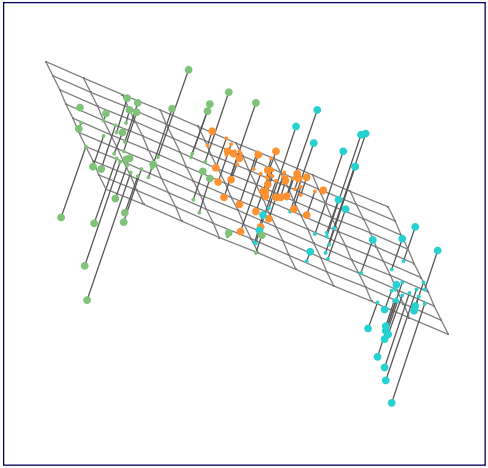
První dvě hlavní komponenty pro data `USAarrests`.

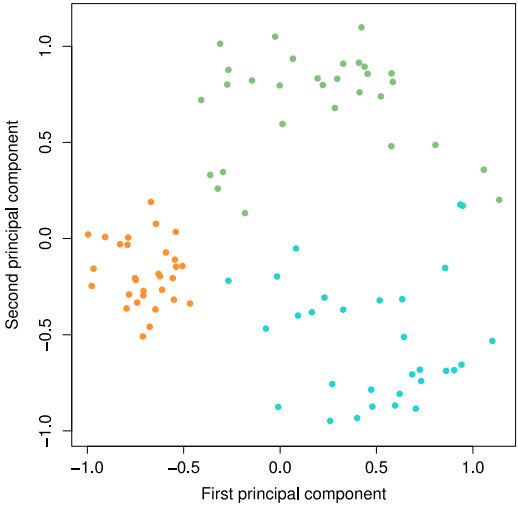
- Modré názvy států reprezentují skóre pro první dvě hlavní komponenty.
- Oranžové šipky označují vektory zátěže prvních dvou hlavních komponent (s osami souřadnic nahoře a vpravo). Tak například zátěž pro `Rape` je u první hlavní komponenty 0,54 a u druhé hlavní komponenty je to 0,17 (slovo `Rape` je v grafu centrováno kolem bodu $[0{,}54, 0{,}17]$).
- Tento graf je známý jako **biplot**, protože zobrazuje jak skóre, tak zátěže hlavních komponent.

Zátěže PCA

	PC1	PC2
Murder	0,5358995	-0,4181809
Assault	0,5831836	-0,1879856
UrbanPop	0,2781909	0,8728062
Rape	0,5434321	0,1673186

Jiná interpretace hlavních komponent



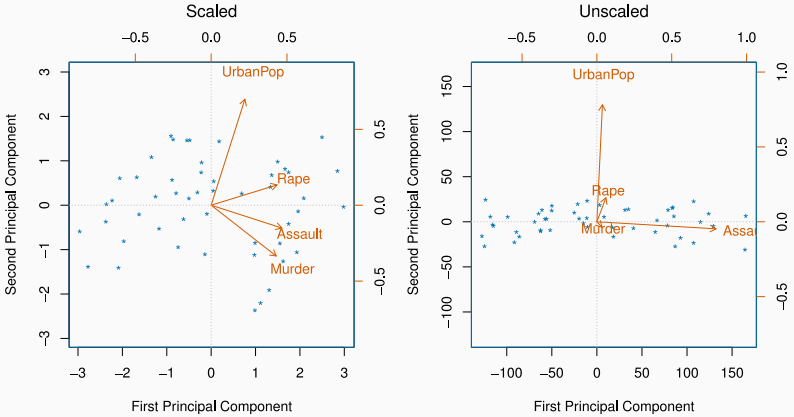


PCA nachází nadrovinu nejbližší k pozorováním

- Vektor zátěže první hlavní komponenty má jednu velmi speciální vlastnost: definuje v p-rozměrném prostoru přímku, která je nejbližší k těm n pozorováním (mírou blízkosti je zde průměrná druhá mocnina euklidovské vzdálenosti).
- Pojem hlavních komponent jako dimenzí, které jsou nejblíže k těm n pozorováním, se rozšiřuje i za pouhou první komponentu.
- Například první dvě hlavní komponenty souboru dat generují rovinu, která je nejblíže těm n pozorováním ve smyslu průměrné druhé mocniny euklidovské vzdálenosti.

Škálování proměnných hraje roli

- Jsou-li proměnné v odlišných jednotkách, doporučuje se přeškálovat je tak, aby každá měla směrodatnou odchylku rovnu jedné.
- Pokud jsou proměnné ve stejných jednotkách, můžete je škálovat nebo ne.



Podíl vysvětleného rozptylu

- Abychom pochopili sílu každé komponenty, zajímá nás znalost **podílu rozptylu**, který každá z nich vysvětluje (PVE, **Proportion Variance Explained**).
- *Celkový rozptyl* přítomný v souboru dat (za předpokladu, že proměnné byly vycentrovány tak, že mají střední hodnotu nula) je definován jako

$$\sum_{j=1}^{p} \text{var}(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$$

  a rozptyl, který vysvětluje m-tá hlavní komponenta, je

$$\text{var}(Z_m) = \frac{1}{n} \sum_{i=1}^{n} z_{im}^2.$$

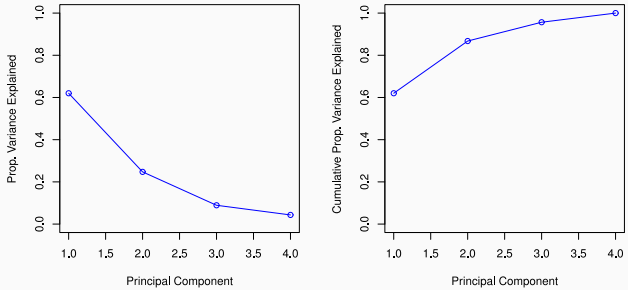
- Dá se ukázat, že $\sum_{j=1}^{p} \text{var}(X_j) = \sum_{m=1}^{M} \text{var}(Z_m)$, kde $M = \min(n-1, p)$.

Podíl vysvětleného rozptylu: pokračování

- PVE m-té hlavní komponenty je tudíž dán následující kladnou veličinou mezi 0 a 1:

$$\frac{\sum_{i=1}^{n} z_{im}^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2}.$$

- Hodnoty PVE jsou v součtu rovny rovny jedné. Někdy zobrazujeme kumulativní hodnoty PVE.



Kolik hlavních komponent bychom měli užít?

Používáme-li hlavní komponenty jako shrnutí našich dat, kolik komponent k tomu stačí?

- Na tuto otázku není jednoduchá odpověď, protože křížová validace není k tomuto účelu k dispozici.
  - **Q:** Proč ne?
  - Kdy bychom mohli použít křížovou validaci k volbě počtu komponent?
- Jako vodítko se dá použít „sutinový graf“ z předchozího slajdu: hledáme nějaký „zlom“.

Shlukování

Shlukování

- *Shlukování* představuje velmi širokou třídu technik pro nalézání *podskupin* neboli *shluků* (také klastrů) v souboru dat.
- Hledáme rozdělení dat do rozdílných skupin takových, že pozorování uvnitř každé skupiny jsou si navzájem zcela podobná.
- Abychom to učinili konkrétním, musíme definovat, co pro dvě nebo více pozorování znamená, že jsou *podobná* nebo *odlišná*.
- Toto je ovšem často úvaha, která je specifická pro daný obor a musí se dělat na základě znalostí o studovaných datech.

PCA versus shlukování

- PCA vyhledává nízkorozměrnou reprezentaci daných pozorování, která vysvětluje značný podíl rozptylu.
- Shlukování vyhledává mezi danými pozorováními homogenní podskupiny.

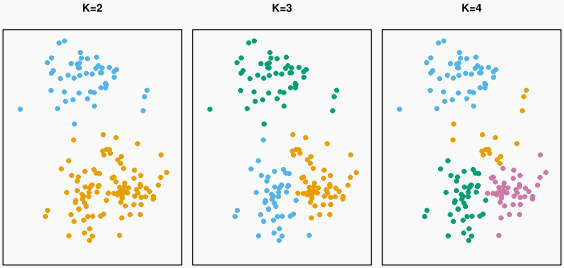
Shlukování pro segmentaci trhu

- Předpokládejme, že máme přístup k velkému počtu měření (např. medián příjmu domácnosti, zaměstnání, vzdálenost od nejbližší městské oblasti atd.) pro velký počet osob.
- Naším cílem je provést segmentaci trhu tím, že stanovíme podskupiny osob, které by mohly být vnímavější k určitým způsobům reklamy nebo které by pravděpodobněji zakoupily konkrétní produkt.
- Úloha provést segmentaci trhu vede ke shlukování osob v daném souboru dat.

Dvě metody shlukování

- V *metodě K-průměrů* se snažíme rozdělit pozorování do předem specifikovaného počtu shluků.
- U *hierarchického shlukování* nevíme předem, kolik shluků chceme; ve skutečnosti končíme se stromovitou vizuální reprezentací daných pozorování, tak zvaným *dendrogramem*, který nám umožňuje vidět najednou shluky získané pro každý jejich možný počet, od 1 do n.

Shlukování metodou K-průměrů



Simulovaný soubor dat se 150 pozorováními ve dvourozměrném prostoru. Panely ukazují výsledky použití metody K-průměrů s různými hodnotami K, počtu shluků. Barva každého pozorování označuje shluk, k němuž bylo přiřazeno algoritmem metody K-průměrů. Poznamenáváme, že zde není žádné uspořádání shluků, takže obarvení shluků je libovolné. Tyto štítky shluků nebyly při shlukování použity; jsou to spíše výstupy procedury shlukování.

Podrobnosti metody K-průměrů

Nechť $\mathcal{C}_1, \ldots, \mathcal{C}_K$ označují množiny obsahující indexy pozorování v každém shluku. Tyto množiny mají dvě vlastnosti:

a) $\mathcal{C}_1 \cup \mathcal{C}_2 \cup \cdots \cup \mathcal{C}_K = \{1, \ldots, n\}$. Jinými slovy, každé pozorování patří alespoň do jednoho z K shluků.

b) $\mathcal{C}_k \cap \mathcal{C}_{k'} = \emptyset$ pro všechna $k \neq k'$. Jinými slovy, shluky přes sebe nepřesahují: žádné pozorování nepatří do více než jednoho shluku.

Například je-li i-té pozorování v k-tém shluku, je $i \in \mathcal{C}_k$.

Podrobnosti metody K-průměrů: pokračování

- Myšlenka za metodou K-průměrů je, že *dobré* shlukování je to, pro něž je vnitroshluková variabilita co nejmenší.
- **Vnitroshluková variabilita** pro shluk $\mathcal{C}_k$ je míra $\mathrm{WCV}(\mathcal{C}_k)$ kvantity, kterou se pozorování uvnitř shluku navzájem liší.
- Chceme tudíž řešit úlohu

$$\underset{\mathcal{C}_1,\ldots,\mathcal{C}_K}{\text{minimalizuj}} \left\{ \sum_{k=1}^{K} \mathrm{WCV}(\mathcal{C}_k) \right\}.$$

- Ve slovech tento vzorec říká, že chceme rozdělit pozorování do K shluků tak, aby celková vnitroshluková variabilita, sečtená přes všechny K shluky, byla co nejmenší.

Jak definovat vnitroshlukovou variabilitu?

- Typicky používáme euklidovskou vzdálenost

$$\mathrm{WCV}(\mathcal{C}_k) = \frac{1}{|\mathcal{C}_k|} \sum_{i,i' \in \mathcal{C}_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2,$$

  kde $|\mathcal{C}_k|$ označuje počet pozorování v k-tém shluku.

- Kombinace p vede na optimalizační úlohu, která definuje shlukování metodou K-průměrů:

$$\underset{\mathcal{C}_1,\ldots,\mathcal{C}_K}{\text{minimalizuj}} \left\{ \sum_{k=1}^{K} \frac{1}{|\mathcal{C}_k|} \sum_{i,i' \in \mathcal{C}_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}.$$

Algoritmus shlukování metodou K-průměrů

a) Každému z pozorování náhodně přiřaď číslo od 1 do K. Tato čísla slouží jako počáteční přiřazení pozorování do shluků.

b) Iteruj do té doby, až se přiřazení do shluků přestane měnit:

   2a. Pro každý z K shluků vypočítej *centroid* shluku. Centroid k-tého shluku je vektor p středních hodnot vlastností pro pozorování v k-tém shluku.

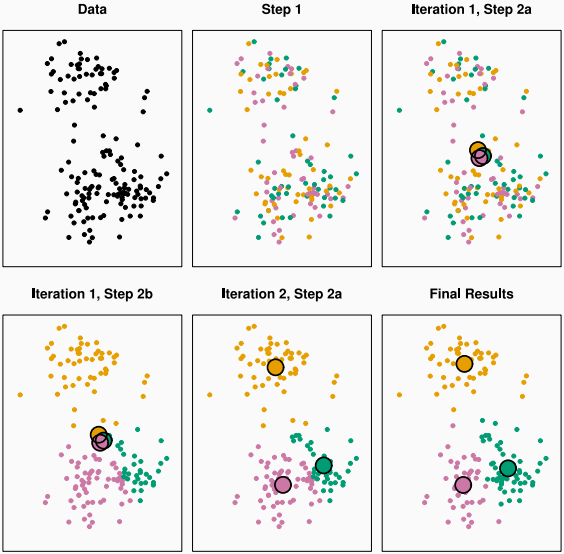
   2b. Přiřaď každé pozorování do shluku, jehož centroid je nejblíže (kde *nejblíže* je definováno pomocí euklidovské vzdálenosti).

- Tento algoritmus zaručeně v každém kroku snižuje hodnotu cílové funkce (4). **Q:** Proč? Všimněte si, že

$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2,$$

  kde $\bar{x}_{kj} = \frac{1}{|C_k|}\sum_{i\in C_k} x_{ij}$ je střední hodnota vlastnosti j ve shluku C_k.
- Není však zaručeno, že to bude dávat globální minimum. **Q:** Proč ne?

Příklad


Data
Step 1
Iteration 1, Step 2a
Iteration 1, Step 2b
Iteration 2, Step 2a
Final Results

Podrobnosti k předchozímu obrázku

Postup algoritmu K-průměrů s $K = 3$:

- *Nahoře vlevo:* Znázorněna jsou pozorování.
- *Nahoře střed:* V Kroku 1 algoritmu se každé pozorování náhodně přiřadí některému shluku.
- *Nahoře vpravo:* V Kroku 2a se vypočítají centroidy shluků. Ty jsou znázorněny jako velké obarvené disky. Na začátku jsou centroidy umístěny téměř úplně přes sebe, protože počáteční přiřazení do shluků bylo provedeno náhodně.
- *Dole vlevo:* V Kroku 2b se každé pozorování přiřadí k nejbližšímu centroidu.
- *Dole střed:* Ještě jednou se provede Krok 2a, což vede k novým centroidům shluků,
- *Dole vpravo:* Výsledky získané po 10 iteracích.

Příklad: rozdílné startovací hodnoty

Podrobnosti k předchozímu obrázku

Shlukování metodou K-průměrů provedené šestkrát na datech z předchozího obrázku s $K = 3$. pokaždé s rozdílným náhodným přiřazením pozorování v Kroku 1 algoritmu metody.

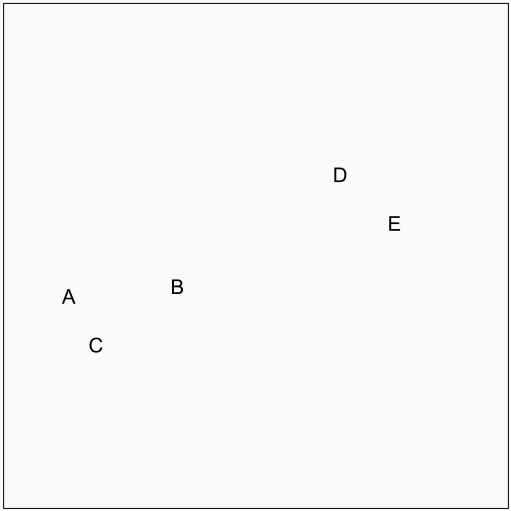
Nad každým grafem je hodnota cílové funkce (4).

Byla získána tři rozdílná lokální optima, z nichž jedno vedlo k menší hodnotě cílové funkce a poskytuje lepší oddělení shluků.

Příklady označené červeně všechny dosáhly téhož nejlepšího řešení s hodnotou cílové funkce 235,8.

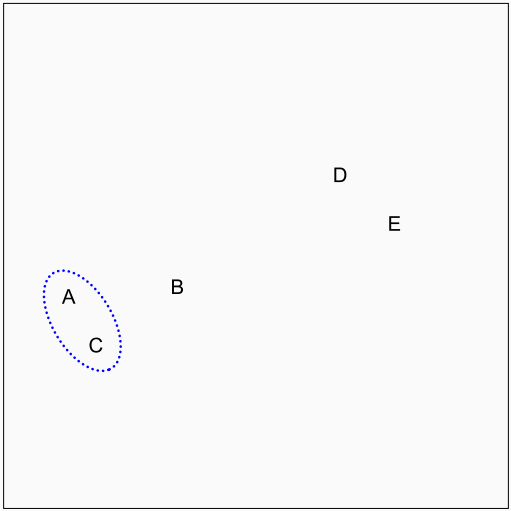
- Shlukování metodou K-průměrů od nás vyžaduje, abychom předem specifikovali počet shluků K. To může znamenat nevýhodu (později probereme strategie pro volbu K).
- *Hierarchické shlukování* je alternativní přístup, který nevyžaduje, abychom se vázali na konkrétní volbu K.
- V této sekci popíšeme shlukování *zdola nahoru* neboli *aglomerativní*. Je to nejběžnější typ hierarchického shlukování a název odkazuje na skutečnost, že se buduje dendrogram počínaje listy a shluky se kombinují směrem ke kmeni.

Tvoří se hierarchie způsobem „zdola nahoru“ ...

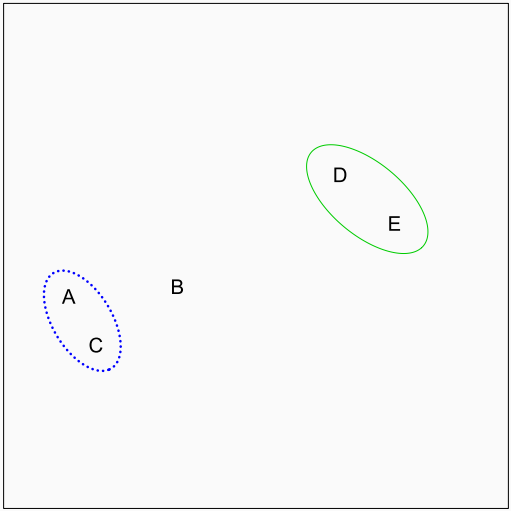


Hierarchické shlukování: myšlenka

Tvoří se hierarchie způsobem „zdola nahoru“ ...

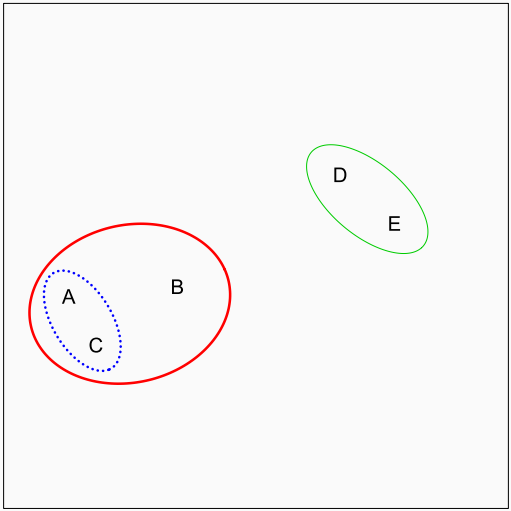


Tvoří se hierarchie způsobem „zdola nahoru“ ...



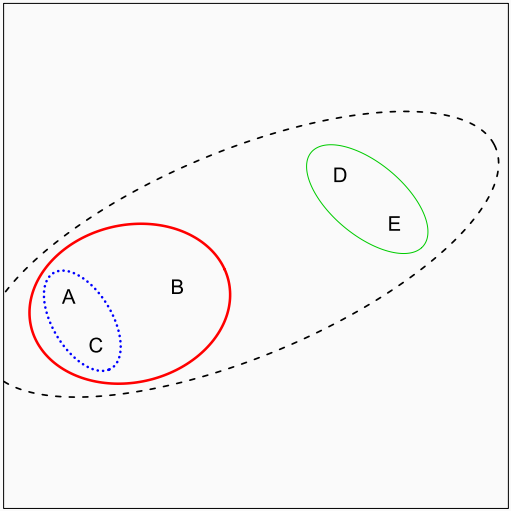
Hierarchické shlukování: myšlenka

Tvoří se hierarchie způsobem „zdola nahoru“ ...



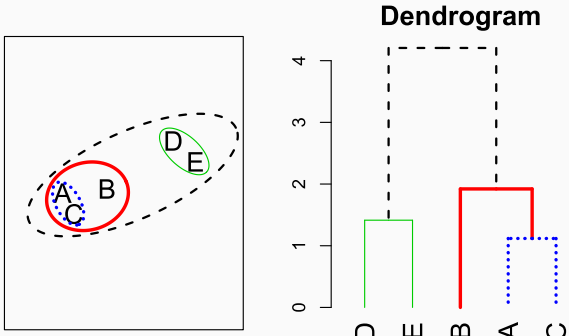
Hierarchické shlukování: myšlenka

Tvoří se hierarchie způsobem „zdola nahoru“ ...



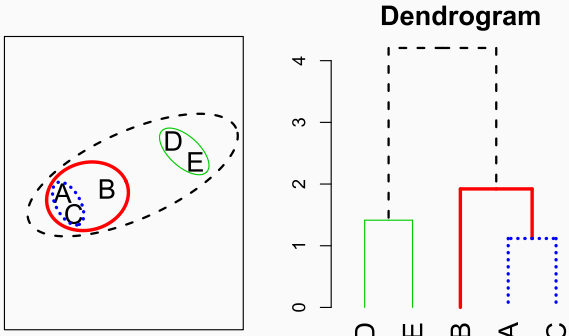
Postup slovy:

- Začni a každým bodem jako vlastním shlukem.
- Urči nejbližší dva shluky a spoj je v jeden.
- Opakuj.
- Konči, když jsou všechny body v jediném shluku.

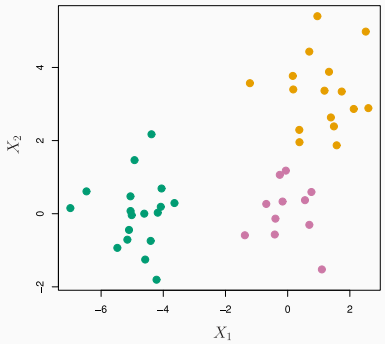


Postup slovy:

- Začni a každým bodem jako vlastním shlukem.
- Urči nejbližší dva shluky a spoj je v jeden.
- Opakuj.
- Konči, když jsou všechny body v jediném shluku.

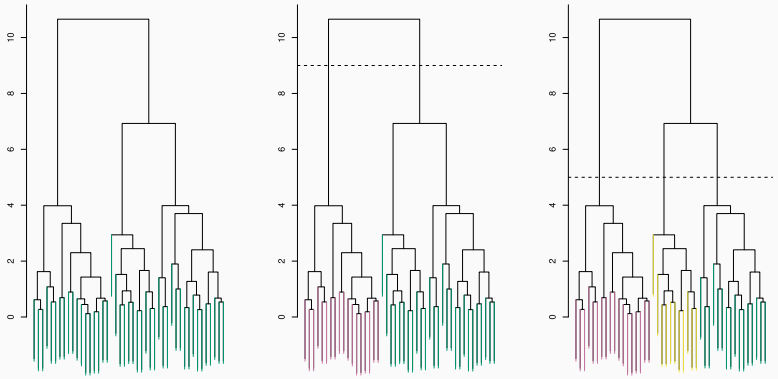


Příklad



45 pozorování vygenerovaných ve dvourozměrném prostoru. Ve skutečnosti jsou zde tři rozličné třídy, označené různými barvami. My však budeme považovat tyto štítky tříd za neznámé a budeme se snažit ta pozorování shluknout, abychom z dat ty třídy odhalili.

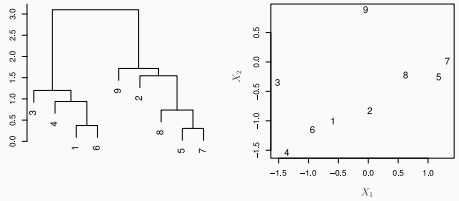
Aplikace hierarchického shlukování



Podrobnosti k předchozímu obrázku

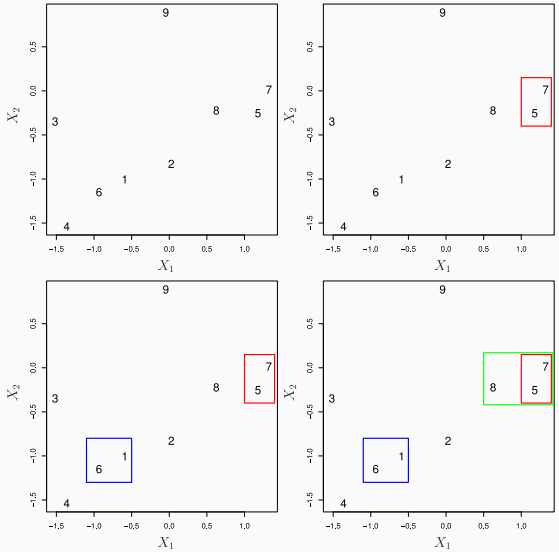
- *Vlevo:* Dendrogram získaný hierarchickým shlukováním dat z předchozího obrázku metodou nejvzdálenějšího souseda a užitím euklidovské vzdálenosti.
- *Střed:* Dendrogram z levého panelu uříznutý ve výšce 9 (označeno čárkovanou přímkou). Tento řez vede na dva rozličné shluky označené rozdílnými barvami.
- *Vpravo:* Dendrogram z levého panelu, tentokrát uříznutý ve výšce 5. Tento řez vede na tři rozličné shluky označené rozdílnými barvami. Poznamenáváme, že ty barvy se nepoužívaly při shlukování, jsou zde prostě použity pro zobrazovací účely.

Další příklad



- Ukázka toho, jak správně interpretovat dendrogram s devíti pozorováními ve dvourozměrném prostoru. K vytvoření dendrogramu vlevo byla použita surová data vpravo.
- Pozorování 5 a 7 jsou si navzájem docela podobná, stejně tak jako pozorování 1 a 6.
- Avšak pozorování 9 *už není podobnější* pozorování 2 než pozorováním 8, 5 a 7, i když jsou si pozorování 9 a 2 blízká co do vodorovné vzdálenosti.
- Je to proto, že pozorování 2, 8, 5 a 7 se všechna spojí s pozorováním 9 ve stejné výšce, přibližně 1.8.

Spojování v předchozím příkladu

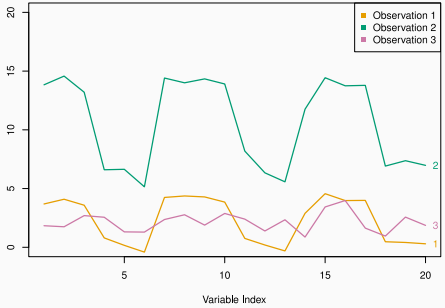


Typy vazby

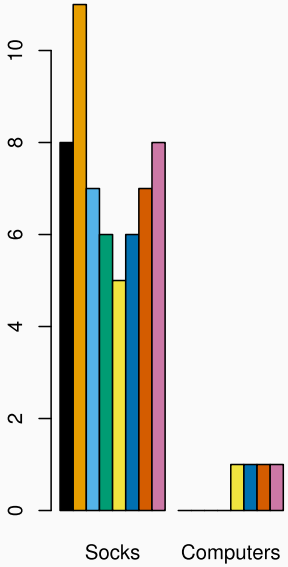
Vazba	*Popis*
Nejvzdálenější soused	Maximální odlišnost mezi shluky. Vypočítej po dvou všechny odlišnosti mezi pozorováními ve shluku A a pozorováními ve shluku B a zaznamenej *největší* z těchto odlišností.
Nejbližší soused	Minimální odlišnost mezi shluky. Vypočítej po dvou všechny odlišnosti mezi pozorováními ve shluku A a pozorováními ve shluku B a zaznamenej *nejmenší* z těchto odlišností.
Průměr	Průměrná odlišnost mezi shluky. Vypočítej po dvou všechny odlišnosti mezi pozorováními ve shluku A a pozorováními ve shluku B a zaznamenej *průměr* z těchto odlišností.
Centroid	Odlišnost mezi centroidem pro shluk A (střední vektor délky p) a centroidem pro shluk B. Centroidní vazba může vést k nežádoucím *inverzím*.

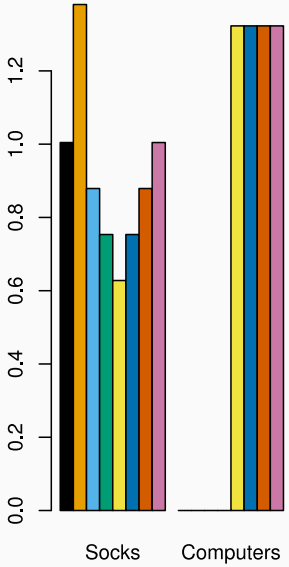
Volba míry rozdílnosti

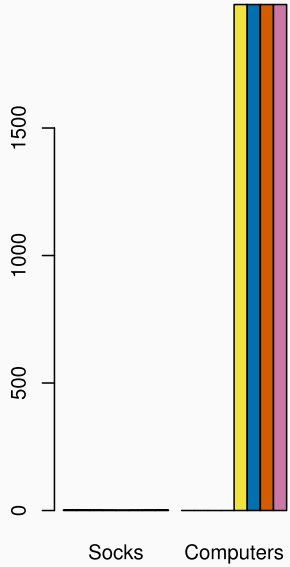
- Dosud používána euklidovská vzdálenost.
- Alternativou je *vzdálenost založená na korelaci*, která považuje dvě pozorování za podobná, jsou-li jejich vlastnosti vysoce korelovány.
- Toto je neobvyklé použití korelace, která se normálně počítá mezi proměnnými; zde se počítá mezi profily pozorování pro každou dvojici pozorování.



Škálování proměnných hraje roli





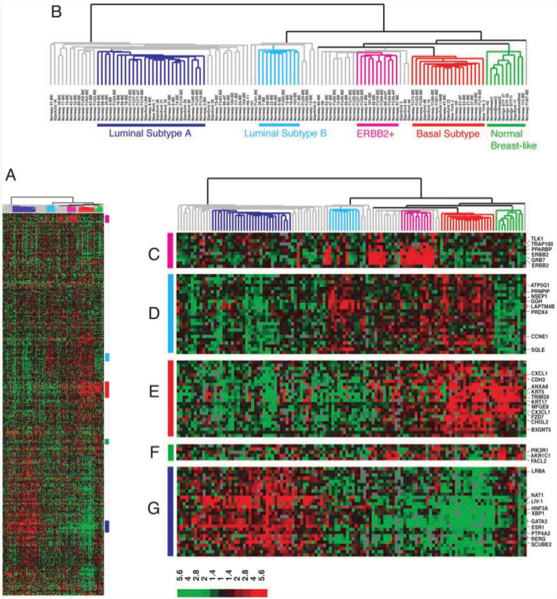


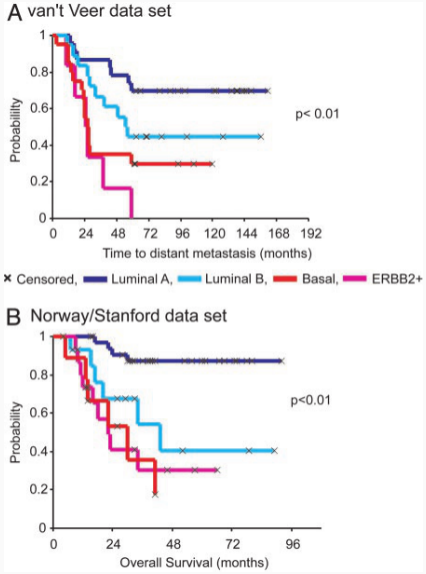
Praktické otázky

- Měla by být pozorování nebo vlastnosti nejprve nějakým způsobem normalizovány? Například by proměnné možná měly být nejdříve vycentrovány tak, aby měly střední hodnotu nula, a přeškálovány tak, aby měly směrodatnou odchylku rovnou jedné.
- V případě hierarchického shlukování:
  - Jaká míra rozdílnosti by měla být použita?
  - Jaký typ vazby by se měl použít?
- Kolik shluků zvolit (jak v metodě K-průměrů, tak při hierarchickém shlukování)? Obtížný problém. Není shoda na metodě. Viz Elements of Statistical Learning, kap. 13 pro více podrobností.

Příklad: studium rakoviny prsu technikou microarray

- „Repeated observation of breast tumor subtypes in independent gene expression data sets“, Sorlie et al., PNAS 2003.
- Průměrná vazba, korelační metrika.
- Shlukování vzorků na základě 500 *vlastních genů:* každá žena byla měřena před chemoterapií a po ní. Vlastní geny mají nejmenší variabilitu uvnitř/mezi.


B
Luminal Subtype A
Luminal Subtype B
ERBB2+
Basal Subtype
Normal
Breast-like
A
C
D
E
F
G
TLK1
TRAP100
PPARBP
ERBB2
GRB7
ERBB2
ATP5G1
PRNPIP
NSEP1
GGH
LAPTM4B
PRDX4
CCNE1
SQLE
CXCL1
CDH3
ANXA8
KRT5
TRIM29
KRT17
MFGE8
CX3CL1
FZD7
CHI3L2
B3GNT5
PIK3R1
AKR1C1
FACL2
LRBA
NAT1
LIV-1
HNF3A
XBP1
GATA3
ESR1
PTP4A2
RERG
SCUBE2
5.6
4
2.8
2
1.4
1
1.4
2
2.8
4
5.6


A van't Veer data set
Probability
1
0.8
0.6
0.4
0.2
0
p< 0.01
0 24 48 72 96 120 144 168 192
Time to distant metastasis (months)
× Censored, Luminal A, Luminal B, Basal, ERBB2+
B Norway/Stanford data set
Probability
1
0.8
0.6
0.4
0.2
0
p<0.01
0 24 48 72 96
Overall Survival (months)

Závěry

- *Učení bez učitele* je důležité pro pochopení variability a struktury seskupování u neoznačených souborů dat a může být užitečným předběžným procesem pro učení s učitelem.
- Je vnitřně obtížnější než *učení s učitelem,* protože zde není žádný zlatý standard (jako nějaká výstupní proměnná) a žádný jednotlivý cíl (jako přesnost na testovacím souboru).
- Je to aktivní oblast výzkumu s mnoha nástroji vyvinutými v poslední době, jako jsou *samoorganizační mapy, analýza nezávislých komponent* a *spektrální shlukování.*
  Viz *The Elements of Statistical Learning,* kapitola 14.