

DOMÁCÍ ÚKOL

METODY S LATENTNÍMI PROMĚNNÝMI

Přednášející: Doc. Ing. O. Pytela, DrSc.

Příklad 1:

Vypočtete algoritmem NIPALS 1. latentní proměnnou z matice $A = \begin{bmatrix} 1 & 1 & 6 \\ 2 & 2 & 4 \\ 3 & 0 & 2 \end{bmatrix}$.

Nejprve vypočteme vektor aritmetických průměrů

$$\bar{x}^T = [2 \quad 1 \quad 4]$$

a vektor směrodatných odchylek

$$s^T = [1 \quad 1 \quad 2].$$

Standartizací obdržíme matici

$$A = \begin{bmatrix} -1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & -1 & -1 \end{bmatrix}.$$

Protože všechny sloupce mají stejnou variabilitu, vezmeme za odhad hlavní komponenty t_1 např. první sloupec:

$$t_1^T = [-1 \quad 0 \quad 1].$$

Po dosazení do vztahu

$$p_1^T = (t_1^T t_1)^{-1} t_1^T A$$

a znormováním podle

$$p_1^N = (p_1^T p_1)^{-\frac{1}{2}} p_1$$

získáme počáteční odhad vektoru míry příspěvků odhadu vektoru hlavní komponenty t_1 . Dosazením pak do vztahu

$$t_1^T = (p_1^T p_1)^{-1} p_1^T A^T$$

získáme odhad hlavní komponenty t_1 . Iteračním opakováním tohoto postupu dostaneme stabilní rozklad vektorů t_1 a p_1 . Konvergenční kritérium má tvar

$$d = (t_{\text{nové}} - t_{\text{staré}})^T (t_{\text{nové}} - t_{\text{staré}}) (t_{\text{nové}}^T t_{\text{nové}})^{-1}.$$

Výpočet se ukončí, je-li $d/N < 10^{-10}$.

V našem případě se obdrží stabilní rozklad vektorů t_1 a p_1 po 9 krocích s hodnotami

$$p_1^T = [0.627968 \quad -0.459689 \quad -0.627968]$$

$$t_1^T = [-1.255935 \quad -0.459689 \quad 1.715624].$$

Příklad 2.

S použitím vhodných kritérií určete nezbytný počet latentních proměnných, bylo-li z dat určeno: $PRESS(0)=S(0)=100$, $PRESS(1)=20$, $S(1)=10$, $PRESS(2)=3.5$, $S(2)=3.4$, $PRESS(3)=3.45$, $S(3)=3.39$.

K určení nezbytného počtu latentních proměnných použijeme Woldovo kritérium založené na poměru hodnoty $PRESS(P)$ a hodnoty $S_R(P-1)$

$$PRESS(P) / S_R(P-1).$$

Zařazení další $(P+1)$ latentní proměnné je nevhodné, je-li hodnota kritéria větší než 0.95. Dostaneme:

$$\begin{aligned} P=1: & \quad PRESS(1) / S_R(0) = 0.2 \\ P=2: & \quad PRESS(2) / S_R(1) = 0.35 \\ P=3: & \quad PRESS(3) / S_R(2) = 1.01 \end{aligned}$$

Protože pro $P=3$ je hodnota kritéria větší než 0.95, není čtvrtá latentní proměnná již významná.

Příklad 3.

Odhadněte hodnotu chybějícího prvku $A[2,2]$, jestliže výpočtem z nekompletní matice byly určeny vektory $p^T = [0.541 \quad 0.423 \quad 0.514 \quad 0.514]$ a $t^T = [-1.340 \quad -0.735 \quad 2.076]$.

Prvky zdrojové matice odpovídající p -té latentní proměnné lze rekonstruovat podle vzorce

$$A_p^{\text{pred}} = t_p p_p^T.$$

Odtud odhad hodnoty prvku $A[2,2]$ je

$$A[2,2] = t[2] p[2] = -0.311.$$

Příklad 4.

Výpočtem metodou PCA byly určeny vektory $p_1^T = [0.012 \ 0.458 \ -0.352 \ 0.987]$ a $p_2^T = [0.926 \ -0.238 \ 0.872 \ -0.115]$. Vyberte reprezentativní sloupce charakterizující nejlépe zdrojovou matici.

Vektory p_1^T a p_2^T představují míru příspěvku příslušné latentní proměnné k popisu variability sloupců zdrojové matice. Vektor p_1^T nám říká, že reprezentativním sloupcem zdrojové matice je sloupec čtvrtý $p_1^T[4] = 0.987$ a vektor p_2^T , že reprezentativním sloupcem je sloupec první $p_2^T[1] = 0.926$ a třetí $p_2^T[3] = 0.872$.

Příklad 5.

Vysvětlete, proč vysvětlená variabilita je při výpočtu metodou FA vždy nižší, než při výpočtu metodou PCA.

Je to způsobeno tím, že z hlediska faktorové analýzy je metoda PCA považována za úplnou komponentní analýzu (neboť pomocí hlavních komponent lze přesně reprodukovat variabilitu zdrojové matice) a metoda FA za neúplnou komponentní analýzu (neboť připouští existenci matice jedinečností, která představuje část variability nevysvětlitelné společnými faktory).

Příklad 6.

Výpočtem metodou kanonických korelací bylo zjištěno:

$$0.297 X_1 + 0.298 X_2 + 0.050 X_3 + 0.256 X_4 = 0.493 Y_1 - 0.213 Y_2 \quad r_1 = 0.830$$

$$0.006 X_1 - 0.115 X_2 + 0.950 X_3 + 0.056 X_4 = 0.493 Y_1 + 0.213 Y_2 \quad r_2 = 0.512$$

Vypočtěte skupinový korelační koeficient a interpretujte výsledky.

Skupinový korelační koeficient R vypočteme ze vztahu

$$R^2 = 1 - (1 - r_1^2)(1 - r_2^2)$$

tj.

$$R^2 = 0.770, R = 0.878.$$

tj. výsledek popisuje 77% variability dat. Dále vidíme, že na součet parametrů Y_1 a Y_2 má parametr X_3 zanedbatelný vliv vzhledem k ostatním parametrům zanedbatelný, ostatní parametry mají vliv stejný. Na rozdíl parametrů Y_1 a Y_2 má výrazný vliv parametr X_3 , vliv ostatních parametrů je nevýrazný.

Příklad 7.

Uveďte typický konkrétní příklad vhodný pro zpracování metodou PLS.

Pomocí metody PLS ověřte, zda lze popsat tepelné vlastnosti materiálů pomocí konstant tepelné vodivosti λ a tepelné kapacity c_p .

materiál	λ	c_p
:	:	:
:	:	:

Příklad 8.

Jeden objekt je charakterizován metrickými znaky (2,10), druhý (3,8), třetí (4,9), čtvrtý (10,4) a pátý (11,5). Vypočítejte matici vzdáleností v Euklidově metrice a proveďte shlukování metodou průměrné vazby. Výsledky interpretujte graficky.

Matice vzdáleností znaků má tvar

1	0.00				
2	2.24	0.00			
3	2.24	1.41	0.00		
4	10.00	9.75	7.81	0.00	
5	10.30	8.54	8.06	1.41	0.00

Vidíme, že nejmenší vzdálenost mají prvky 3-2 a 5-4. Vytvoříme tak první shluky, spočteme těžiště nových shluků a vypočteme matici vzdáleností:

1	0.00		
2-3	2.12	0.00	
4-5	5.12	8.06	0.00

Vidíme, že nejmenší vzdálenost mají shluk 2-3 a znak 1. Vytvoříme z nich shluk, spočteme těžiště nového shluku a vypočteme matici vzdáleností:

1-2-3	3.00	
4-5	58.75	0.00

Výsledný dendrogram pak vypadá následovně:

