SEMESTRÁLNÍ PRÁCE

Metody s latentními proměnnými a klasifikační metody

Ing. Pavel Bouchalík

1. ZADÁNÍ

Tato semestrální práce je písemným vypracováním zkouškových otázek z okruhu **Metody s latentními proměnnými a klasifikační metody**.

Otázka 1: Vypočtěte algoritmem NIPALS 1. latentní proměnnou z matice A [řádek,sloupec]: A[1,1]=1, A[2,1]=2, A[3,1]=3, A[1,2]=1, A[2,2]=2, A[3,2]=0, A[1,3]=6, A[2,3]=4, A[3,3]=2. Matici před zpracováním standardizujte.

Otázka 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.

Otázka 3: Odhadněte hodnotu chybějícího prvku A[2,2], jestliže výpočtem z nekompletní matice byly určeny vektory p: 0,541 0,423 0,514 0,514 t: -1,340 -0,735 2,076

Otázka 4: Výpočtem metodou PCA byly určeny vektory p1: 0,012 0,458 -0,352 0,987 p2: 0,926 -0,238 0,872 -0,115. Vypočtěte komunality a vyberte sloupec, který nejlépe charakterizuje celou matici.

Otázka 5: Vysvětlete, proč vysvětlená variabilita je při výpočtu metodou FA vždy nižší, než při výpočtu metodou PCA.

Otázka 6: Výpočtem metodou kanonických korelací bylo zjištěno:
0,297 X1 + 0,298 X2 + 0,050 X3 + 0,256 X4 = 0,493 Y1 - 0,213 Y2 r1 = 0,830
0,006 X1 - 0,115 X2 + 0,950 X3 + 0,056 X4 = 0,493 Y1 + 0,213 Y2 r1 = 0,512
Vypočtěte skupinový korelační koeficient a interpretujte výsledky.

Otázka 7: Uveďte nějaký konkrétní příklad vhodný pro zpracování metodou PLS.

Otázka 8: Jeden objekt je charakterizován metrickými znaky (2,10), druhý (3,8), třetí (4,9), čtvrtý (10,4) a pátý (11,5). Vypočtěte matici vzdáleností v Euklidově metrice a dokumentujte výpočet shlukování některou z používaných metod. Výsledky interpretujte graficky.

Otázka 9: Popište slovně postup aplikace metod s latentními proměnnými nebo klasifikačních metod na nějakém konkrétním příkladu ze své praxe.

2. VYPRACOVANÉ OTÁZKY

2.1 Výpočet pomocí algoritmu NIPALS

Zdrojová matice má dle zadání tvar: $A=\begin{Vmatrix}1 & 1 & 6\\ 2 & 2 & 4\\ 3 & 0 & 2\end{Vmatrix}$ V další fázi provedeme její standardizaci. To znamená, že od každého prvku matice odečteme příslušný sloupcový průměr a získaný rozdíl vydělíme sloupcovou směrodatnou odchylkou. Vypočtený vektor sloupcových průměrů je $\bar{\bar{X}}^T=\|2 \quad 1 \quad 4\|$ a vektor sloupcových směrodatných odchylek je $s^T=\|1 \quad 1 \quad 2\|$. Standardizovaná matice vypočtená dle výše uvedeného návodu má tvar: $A=\begin{Vmatrix}-1 & 0 & 1\\ 0 & 1 & 0\\ 1 & -1 & -1\end{Vmatrix}$

Po standardizaci zdrojové matice můžeme přikročit k určení první latentní proměnné. Postup spočívá v tom, že se zvolí vektor matice T a na základě níže uvedeného vztahu (1) se provede výpočet vektoru faktorových zátěží p_1.

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

Jako vektor "t" volíme obvykle vektor zdrojové matice s největší variabilitou. V našem případě zvolíme: $t=\begin{Vmatrix}-1\\0\\1\end{Vmatrix}$.

Dosadíme do vztahu (1): $(t^Tt)=\begin{Vmatrix}-1 & 0 & 1\end{Vmatrix}\begin{Vmatrix}-1\\0\\1\end{Vmatrix}=\begin{Vmatrix}2\end{Vmatrix}$, $(t^Tt)^{-1}t^TA=\begin{Vmatrix}0,5\end{Vmatrix}\begin{Vmatrix}-1 & 0 & 1\end{Vmatrix}\begin{Vmatrix}-1 & 0 & 1\\0 & 1 & 0\\1 & -1 & -1\end{Vmatrix}$. Výsledek výpočtu je $p_1=\begin{Vmatrix}1 & -0,5 & 1\end{Vmatrix}$.

Další fází výpočtu je normování vektoru faktorových zátěží dle vztahu (2).

$$p_1^N=(p_1^Tp_1)^{-1/2} \qquad (2)$$

$$p_1^N=\left(\begin{Vmatrix}1\\-0,5\\1\end{Vmatrix}\begin{Vmatrix}1 & -0,5 & 1\end{Vmatrix}\right)^{-0,5}\begin{Vmatrix}1 & -0,5 & 1\end{Vmatrix}=\begin{Vmatrix}0,6667 & -0,333 & -0,6667\end{Vmatrix}$$

Takto získaný normovaný vektor dosadíme do vztahu (3).

$$t_1=(p_1^Tp_1)^{-1}p_1^TA \qquad (3)$$

$$t_1=\begin{Vmatrix}-1,3333 & 0,3333 & 1,6667\end{Vmatrix}$$

Výše uvedený postup se opakuje do té doby než získáme stabilní odhady. Výpočet končí, když je podíl d/N menší jak 10^{-10}. Index N je počet normování a d je konvergenční kritérium, které se vypočte dle vztahu (4).

$$d=(t_{nové}-t_{staré})^T(t_{nové}-t_{staré})(t_{nové}t_{staré})^{-1} \qquad (4)$$

K dosažení stabilního odhadu stačilo provést 8 opakování. Výsledek je uvedený v tabulce č.1

tabulka č.1: Vypočtené vektory latentních proměnných a vektorových zátěží

Opakování	Vektor zátěží p_1	Normovaný vektor zátěží p_1^N	Vektor latentních hodnot
1	1 -0,5 -1	0,6667 -0,3333 -0,6667	-1,3333 -0,3334 1,6667
2	-0,6750 -0,4500 0,6750	0,6396 -0,4264 0,6396	-1,2792 -0,4264 1,7056
3	0,6314 0,4510 -0,6314	0,6312 -0,4508 0,6312	-1,2623 -0,4508 1,7232
4	0,6289 -0,4574 -0,6289	0,6288 -0,4573 -0,6288	-1,2577 -0,4573 1,7150
5	0,6282 -0,4591 -0,6282	0,6282 -0,4591 -0,6282	-1,2564 -0,4591 1,7154
6	0,6280 -0,4695 0,6280	0,6280 -0,4695 -0,6280	-1,2561 -0,4596 1,7157
7	0,6279 -0,4596 -0,6279	0,6280 -0,4597 -0,6280	-1,2559 -0,4597 1,7156
8	0,6280 -0,4597 -0,6280	0,6280 -0,4597 -0,6280	-1,2559 -0,4597 1,7156

2.2 Určení minimálního počtu latentních proměnných

K určení počtu latentních proměnných se používá celá řada numerických postupů s různou mírou nespolehlivosti, resp. s různou mírou pravděpodobnosti jejich selhání. Numerické postupy jsou založeny na odhadu experimentální chyby pomocí reziduí. Rezidua mohou být klasická nebo odhadnutá na základě JackKnife techniky. V tomto případě hovoříme o tzv. predikované sumě čtverců nebo-li *Prediction sums of squares*, čili *PRESS*. Hodnoty PRESS máme zadány:

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.

Použijeme tzv. Woldovo kritérium, které je definováno následujícím vztahem:

$$W=\frac{PRESS(P)}{S(P-1)} \qquad (5)$$

Pokud hodnota W ve (5) překročí hodnotu 0,95 je následující latentní proměnná nevýznamná. Provedeme dosazení dle uvedeného vztahu:

W_1=20/100=0,2
W_2=3,5/100=0,35
W_3=3,45/3,4=1,01

V tomto případě je W_3>0,95 a platí, že čtvrtá latentní proměnná není významná. Signifikantní počet proměnných je tedy 3.

2.3 Odhad hodnoty chybějícího prvku

Pro určení chybějících hodnot ve zdrojové matici se používá tzv. metoda *Krátkého cyklu*. Jde o rekonstrukci zdrojové matice A s použitím prvních p latentních proměnných popisujících podstatnou část variability zdrojové matice bez zahrnutí experimentální chyby. Prvky zdrojové matice odpovídající p-té latentní proměnné je možné získat ze vzorce:

$$A_p^{pred}=t_p p_p^T \quad (6)$$

Pro odhad hodnoty prvku A[2,2] takto platí:

A[2,2] = t[2] p[2] = -0,3109

Hodnota prvku zdrojové matice ve druhém řádku a druhém sloupci je -0,3109.

2.4 Výpočet komunalit

Komunalista charakterizuje jaké procento z celkového rozptylu charakterizuje příslušná latentní proměnná. Výpočet se provádí pomocí vzorce:

$$h_1^2 = \sum_{p=1}^{P} \left(P_{ip}^N \right)^2 \quad (7)$$

$$P_{ip}^N = \frac{P_{ip}}{\sum_{i=1}^{M} P_{ip}^2} \quad (8)$$

h_1^2 - je komunalita
P_{ip}^N - normované zátěže
P_{ip} - zátěže
P – číslo latentní proměnné
M – počet sloupců zdrojové matice

Vektory zátěží jsou: p1: 0,012 0,458 -0,352 0,987 p2: 0,926 -0,238 0,872 -0,115
Podle vzorce (8) jsou hodnoty normovaných zátěží:

$P_{11}^N = 0,012/(0,012^2+0,458^2-0,352^2+0,987^2)=0,0917$
$P_{21}^N = 0,458/(0,012^2+0,458^2-0,352^2+0,987^2)=0,3501$
$P_{31}^N = -0,352/(0,012^2+0,458^2-0,352^2+0,987^2)=-0,2691$
$P_{41}^N = 0,987/(0,012^2+0,458^2-0,352^2+0,987^2)=0,7547$

$P_{12}^N = 0,926/(0,926^2-0,238^2+0,872^2-0,115^2)=0,5487$
$P_{22}^N = -0,238/(0,926^2-0,238^2+0,872^2-0,115^2)=-0,1410$
$P_{32}^N = 0,872/(0,926^2-0,238^2+0,872^2-0,115^2)=0,5167$
$P_{42}^N = -0,115/(0,926^2-0,238^2+0,872^2-0,115^2)=-0,0681$

Komunality vypočteme takto:

$h_1^2 = 0{,}092^2 + 0{,}549^2 = 0{,}309$
$h_2^2 = 0{,}3501^2 + (-0{,}1410)^2 = 0{,}142$
$h_3^2 = -0{,}2691^2 + 0{,}5167^2 = 0{,}339$
$h_4^2 = 0{,}7546^2 + 0{,}0681^2 = 0{,}5742$

Čtvrtý sloupec nejlépe charakterizuje danou matici.

2.5 Vysvětlená variabilita u FA

Metoda PCA je z hlediska faktorové analýzy považována za úplnou komponentní analýzu. Pomocí hlavních komponent je možné přesně reprodukovat variabilitu zdrojové matice. Metoda FA je z tohoto pohledu neúplnou komponentní analýzou, protože připouští existenci matice jedinečností, která zahrnuje část variability, která se nedá vysvětlit společnými faktory.

2.6 Kanonická korelační analýza

Výpočet skupinového korelačního koeficientu provedeme pomocí vztahu (9):

$$R_{XY}^2 = 1-(1-r_1^2)(1-r_2^2)\ldots(1-r_p^2) \qquad (9)$$

Kde $\underline{r}$ jsou kanonické korelační koeficienty. Tyto jsou součástí zadání, takže výpočet provedeme dosazením do vztahu (9).

$r_1 = 0{,}830$, $r_2 = 0{,}512$
$R_{XY}^2 = 1-(1-0{,}830^2)(1-0{,}512^2) = 0{,}7704$
$R_{XY} = 0{,}878$

Hodnota korelačního koeficientu je 77% to znamená, že jsme 77% variability vysvětlili pomocí určeného modelu.

Z první ze dvou níže uvedených rovnic je patrné, že nárůst hodnot parametrů X bude mít za následek nárůst parametrů Y. Dále je patrné, že parametr X_3 není významný vzhledem k nízké hodnotě příslušného koeficientu. Z druhé rovnice je vidět, že s nárůstem parametrů X_1, X_3 a X_4 budou růst parametry Y a naopak s nárůstem parametru X_2 budou klesat. Parametr X_1 má zanedbatelný význam ve srovnání se zbývajícími parametry X.

$0{,}050X_3 + 0{,}256X_4 = 0{,}493Y_1 - 0{,}213Y_2$
$0{,}006X_1 - 0{,}115X_2 + 0{,}950X_3 + 0{,}056X_4 = 0{,}493Y_1 + 0{,}213Y_2$

2.7 Příklad vhodný pro metodu PLS

Metoda projekce latentních proměnných struktur PLS (partial lest square) je zobecněným postupem pro získání popisu vztahu mezi závislým náhodným vektorem $\underline{y}$ a vysvětlujícím náhodným vektorem $\underline{x}$ prostřednictvím latentních proměnných. Tímto postupem je možné opakováním kroků až do konvergence získat vektor latentních proměnných a zátěží. Možným příkladem z praxe, na který by se metoda dala aplikovat je např. následující tabulka, která popisuje úpravu povrchu titanové běloby organickými povrchovými úpravami a vyhodnocení výsledného stavu pomocí snášivosti různě polárními sloučeninami a měřením kontaktního úhlu:

Činidlo	spotřeba oleje	spotřeba vody	spotřeba DOP	kontaktní úhel
polyethylenglykol 200	******	******	******	******
polyethylenglykol 400	******	******	******	******
polyethylenglykol 600	******	******	******	******
trimethylolpropan	******	******	******	******
polyol TP 15	******	******	******	******
polyol TP 32	******	******	******	******
α - olefiny Neraten	******	******	******	******

2.8 Určení eukleidovské vzdálenosti

Eukleidovská vzdálenost se používá k určení míry podobnosti metrických objektů. Výpočet Eukleidovské vzdálenosti e_d je dán vztahem (10).

$$e_d = \left(\sum_{P=1}^{P} \left(x_{kp} - x_{iP} \right)^2 \right) \qquad (10)$$

Dosazením do vzorce (10) získáme míry podobnosti jednotlivých objektů.

ed(1,2)= 2,236068
ed(1,3)= 2,236068
ed(1,4)= 10
ed(1,5)= 10,29563
ed(2,3)= 1,414214
ed(2,4)= 8,062258
ed(2,5)= 8,544004
ed(3,4)= 7,81025
ed(3,5)= 8,062258
ed(4,5)= 1,414214

Abychom mohli použít některou metodu shlukování objektů musíme získané výsledky nejprve uspořádat do matice vzdáleností:

0	2,24	2,24	10	10,3
2,24	0	1,41	8,06	8,54
2,24	1,41	0	7,81	8,06
10	8,06	7,81	0	1,41
10,3	8,54	8,06	1,41	0

Použijeme například metodu nejbližšího souseda. Tato metoda je založena na principu, že se objekty shlukují podle nejmenší vzdálenosti mezi dvěma nejbližšími sousedy. Postup je následující:

1. krok
Nejmenší vzdálenost mají objekty 2-3 a 4-5 a to 1,41. Vytvoříme první tři shluky -1, 2-3 a 4-5.

2. krok
Vypočteme novou matici vzdáleností. Nejkratší vzdálenost mezi bodem 1 a shluky 2-3 je 2,24 (oba body jsou stejně vzdáleny).
Vzdálenost mezi bodem 1 shlukem 4-5 je 10, protože nejbližší soused bodu 1 ze shluku 4-5 je bod 4
Pro výpočet vzdálenosti mezi shlukem 2-3 a shlukem 4-5 použijeme hodnotu 7,81, což je vzdálenost odpovídající bodům 3 a 4, což jsou nejbližší sousedé jmenovaných shluků.

1	0	2,42	10
2 – 3	2,24	0	7,81
4 – 5	10	7,81	0

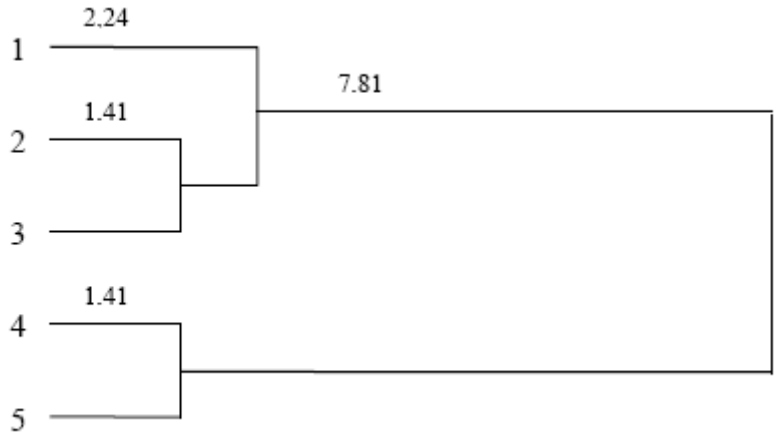
3. krok
Nejmenší vzdálenosti mají bod 1 a shluk bodů 2-3 a je to hodnota 2,24. Vytvořily se pouze dva shluky (1-2-3 a 4-5).

4. krok
Provedeme výpočet matice vzdáleností. Pro výpočet vzdáleností mezi shlukem 1-2-3 a shlukem 4-5 použijeme hodnotu 7,81, což je vzdálenost odpovídající bodům 3 a 4.

1 – 2 – 3	0	7,81
4 – 5	7,81	0

5. krok
V tomto kroku jsme všechny objekty seskupily do jednoho shluku. Seskupení shluků znázorníme dendrogramem.

Obrázek 1: Dendrogram (metoda nejbližšího souseda)



2.9 Aplikace metod s latentními proměnnými

V prvé fázi provedeme přípravu dat standardizací, kterou odstraníme vliv rozměru sledovaných znaků jednotlivých objektů. V další fázi se provádí exploratorní analýza, která v tomto případě spočívá ve vyhledávání podobných objektů (objektů se stejnými znaky). K tomu lze využít grafické diagnostiky jako např. tzv. Indexové grafy (Chernoffovy tváře aj.). Potom se provede výpočet korelační, resp. kovarianční matice. Pak následuje výpočet latentních proměnných a komunalit. Dále se určí minimální počet latentních proměnných potřebných k určení modelu. K tomu slouží např. grafická diagnostika Scree plot, resp. tzv. sutinový graf. Na závěr se pokusíme jednotlivým latentním proměnným přiřadit fyzikální význam. K tomu využíváme vhodné grafické diagnostiky jako biolog nebo graf komponentního skóre. Tento postup umožňuje určit významné znaky a odlišit je od nevýznamných.

Příklad praktického využití – termická hydrolýza odželezněného titanového roztoku. Hydrolýza je složitý proces jak po stránce chemismu tak i z hlediska počtu parametrů, kterými je ovlivněna. Z hlediska kvality polotovaru je pro mne důležitá sedimentační rychlost charakterizovaná např. velikostí částic. Za použití metod s latentními proměnnými mohu určit z velkého počtu parametrů hydrolýzy takové parametry, které jsou skutečně významné. Ty pak mohu použít jako základ pro modelování procesu.