

Eksploracyjna Analiza Danych.

Metoda geometryczna

Andrzej Dąbrowski

1 stycznia 2019

Contents

1	Wstęp	5
2	Macierz danych	7
3	Model liniowy	9
3.1	Reifikacja modelu	11
3.2	Przykład	12
4	Grupowanie - analiza skupień	17
4.1	Podział macierzy danych na klasy	17
4.2	Podział Woronoja	22
5	Dyskryminacja (klasyfikacja z nauczycielem)	27
5.1	Zmienne dyskryminacyjne	27
5.2	Podział dychotomiczny	28
6	Literatura	31

Chapter 1

Wstęp

Dane są jak ludzie. Wystarczy je trochę pomęczyć a powiedzą całą prawdę. [Ronald Coase]

Chapter 2

Macierz danych

Definicja 2.1. Macierz danych jest tablicą o wymiarach $n \times p$

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = [X^1 \quad X^2 \quad \cdots \quad X^p] = \begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{bmatrix}$$

Macierz danych przedstawia zbiór n przypadków, każdy z nich opisany jest przez p zmiennych (cech). Kolumny X^1, X^2, \dots, X^p są n -wymiarowymi wektorami cech. Wiersze X_1, X_2, \dots, X_n są p -wymiarowymi wektorami przypadków. Zazwyczaj macierz danych (dla $p = 2$) jest przedstawiana jako 2-wymiarowy wykres n punktów (*wykres rozrzutu*)

Definicja 2.2. Macierzą klonów wektora $A \in \mathbf{R}^p$ jest tablica o wymiarach $n \times p$

$$A^n = \begin{bmatrix} A^T \\ A^T \\ \vdots \\ A^T \end{bmatrix}$$

Na przykład, jeżeli \mathbf{a} jest liczbą to \mathbf{a}^n jest n -wymiarowym wektorem

$$\begin{bmatrix} a \\ a \\ \vdots \\ a \end{bmatrix}$$

Wektory w macierzy X są zapisane w standardowym, kartezjańskim układzie współrzędnych o początku w punkcie $\mathbf{0}$. Dla dowolnych wektorów A, B , symbol AB oznacza wektor o początku w punkcie A i końcu w punkcie B .

Definicja 2.3. Środkiem ciężkości macierzy danych X jest punkt G , spełniający równanie

$$\sum_{i=1}^n GX_i = \mathbf{0}$$

Twierdzenie 2.1. Każda macierz ma jedyny środek ciężkości. Oznaczmy go $G = g(X)$. Co więcej,

$$g(X) = \frac{1}{n} X^T \mathbf{1}^n$$

Dla macierzy danych X naturalny jest układ współrzędnych o centrum w środku ciężkości $g(X)$. Taka operacja nazywa się **centrowaniem** X i oznacza symbolem X^0 . Mamy więc:

$$X^0 = X - g(X)^n$$

Definicja 2.4. Macierz wariancji/kowariancji.

Niech X i Y będą macierzami o wymiarach odpowiednio $n \times p$ i $n \times q$.

Macierz kowariancji między X i Y jest macierz o wymiarach $p \times q$

$$V(X, Y) = \frac{1}{n} (X^0)^T Y^0$$

Macierz wariancji dla X nazywamy macierz kwadratową o wymiarach $p \times p$: $V(X) \stackrel{def}{=} V(X, X)$

Definicja 2.5. Standaryzacja macierzy danych

Niech

$$X^0 = [Z^1 \quad Z^2 \quad \dots \quad Z^p]$$

Standaryzacją macierzy X jest tablica:

$$S(X) = \begin{bmatrix} \frac{Z^1}{\|Z^1\|} & \frac{Z^2}{\|Z^2\|} & \dots & \frac{Z^p}{\|Z^p\|} \end{bmatrix}$$

Symbol $\|A\|$ oznacza długość wektora A

Definicja 2.6. Macierz korelacji macierzy X i Y o wymiarach odpowiednio $n \times p$ i $n \times q$ jest macierz

$$R(X, Y) \stackrel{def}{=} V(S(X), S(Y))$$

$$R(X) \stackrel{def}{=} R(X, X).$$

Definicja 2.7. Odległość Frobeniusa macierzy X i Y o tych samych wymiarach $n \times p$ jest liczba

$$d^2(X, Y) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^p (x_{ij} - y_{ij})^2 = \frac{1}{n} \sum_{i=1}^n \|X_i - Y_i\|^2 = \frac{1}{n} \sum_{j=1}^p \|X^j - Y^j\|^2$$

Definicja 2.8. Bezwładność macierzy X o wymiarach $n \times p$ jest liczba

$$J(X) = d^2(X, g(X)^n)$$

Propozycja 2.1.

$$J(X) = \text{Tr}(V(X))$$

Chapter 3

Model liniowy

Definicja 3.1. Hiperpłaszczyzna H wymiaru $d \geq 0$ w przestrzeni \mathbf{R}^p o kierunku podprzestrzeni U i przechodząca przez punkt $A \in \mathbf{R}^p$

$$H = H(U, A) \stackrel{def}{=} U + A, \dim(U) = d$$

Przykład 3.1. Gdy $U = \{0\}$ to hiperpłaszczyzna przechodząca przez A jest punktem A . Przyjmujemy, że wymiar takiej hiperpłaszczyzny jest równy 0.

Gdy U jest jednowymiarową podprzestrzenią \mathbf{R}^p o wektorze bazowym B to hiperpłaszczyzna wymiaru 1 przechodząca przez A jest prostą o kierunku wektora B przechodzącą przez punkt A

Dla macierzy X i hiperpłaszczyzny H Oznaczenie $X \subset H$ jest równoważne spełnieniu warunku

$$X_i \in H \quad i = 1, 2, \dots, n$$

Definicja 3.2. Model liniowy

Niech X macierz danych $n \times p$. Modelem liniowym macierzy X o wymiarze d jest macierz $Y \subset H$ wymiaru $n \times p$, dla pewnej hiperpłaszczyzny H wymiaru $0 \leq d \leq p$.

Definicja 3.3. Najlepszy model liniowy wymiaru d

Niech Y będzie dowolnym modelem liniowym macierzy X o wymiarze d .

Model Y^* jest najlepszym modelem liniowym macierzy X o wymiarze d gdy spełnia warunek

$$d^2(X, Y^*) \leq d^2(X, Y) \quad \forall Y$$

Twierdzenie 3.1. Twierdzenie Pitagorasa dla punktu

$$\forall A \in \mathbf{R}^p \quad d^2(X, A^n) = d^2(X, g(X)^n) + d^2(A^n, g(X)^n) = J(X) + \|g(X) - A\|^2$$

Wniosek 3.1.

$$\forall A \in \mathbf{R}^p \quad d^2(X, g(X)^n) \leq d^2(X, A^n) J(X) \leq d^2(X, A^n)$$

Wniosek 3.2. Najlepszym modelem 0-wymiarowym macierzy X jest $g(X)^n$

Najlepszym modelem p -wymiarowym macierzy X jest X

Definicja 3.4. Kwadratowy błąd względny modelu

Niech Y będzie modelem liniowym macierzy X . Kwadratowy błąd względny modelu to liczba

$$RSE(Y) = \frac{d^2(X, Y)}{J(X)}$$

Twierdzenie 3.2. Niech Y_1 i Y_2 będą najlepszymi modelami liniowymi macierzy X o wymiarach $0 \leq d_1 \leq d_2 \leq p$. Wtedy

$$0 \leq RSE(Y_2) \leq RSE(Y_1) \leq 1$$

Definicja 3.5. Rzut prostopadły na hiperpłaszczyznę

Niech $H = H(U, A)$ będzie hiperpłaszczyzną o kierunku podprzestrzeni U o wymiarze $d > 0$ i przechodzącą przez punkt $A \in \mathbf{R}^p$. Rzutem punktu $Q \in \mathbf{R}^p$ na H jest punkt

$$P_H(Q) = P_U(Q - A) + A = P_U(Q) + (A - P_U(A))$$

gdzie $P_U()$ jest rzutem prostopadłym na podprzestrzeń U

Wniosek 3.3. P_H jest operatorem liniowym $\iff H$ jest podprzestrzenią liniową

Wniosek 3.4. Dla $n \times p$ macierzy W

$$P_H(W) = W \iff W \subset H$$

Twierdzenie 3.3. Twierdzenie Pitagorasa dla hiperpłaszczyzn

Niech H będzie hiperpłaszczyzną wymiaru $0 < d \leq p$, X, Y macierzami danych wymiaru $n \times p$, $Y \subset H$

$$d^2(X, Y) = d^2(X, P_H(X)) + d^2(P_H(X), Y)$$

Wniosek 3.5. Niech Y^* będzie najlepszym modelem liniowym wymiaru d dla macierzy X , $Y^* \subset H^*$. Wtedy

$$Y^* = P_{H^*}(X)$$

Z ostatniego wniosku wynika, że budowanie najlepszego modelu Y^* jest równoważne szukaniu najlepszej hiperpłaszczyzny rzutu H^*

Dlatego wymiennie można używać jako synonimu modelu macierzy danych Y^* i hiperpłaszczyzny rzutu H^*

Wniosek 3.6.

$$RSE(X, P_{H^*}(X)) = 1 - \frac{J(P_{H^*}(X))}{J(X)}$$

Definicja 3.6. Współczynnik determinacji najlepszego modelu liniowego wymiaru d

$$r^2(P_{H^*}(X)) \stackrel{\text{def}}{=} \frac{J(P_{H^*}(X))}{J(X)}$$

Twierdzenie 3.4.

$$H^* = U^* + g(X)$$

dla pewnej podprzestrzeni $U^* \subset \mathbf{R}^p$

Budowanie najlepszego modelu wymiaru d można ograniczyć do szukania najlepszej podprzestrzeni wymiaru d .

Twierdzenie 3.5. Najlepszy model d wymiarowy $H^* = U^* + g(X)$ dla macierzy danych X spełnia jeden z dwóch warunków;

$$J(P_{U^*+g(X)}(X)) \geq J(P_{U+g(X)}(X)) \text{Tr}(V(P_{U^*}(X^0))) \geq J(P_U(X^0))$$

dla dowolnej podprzestrzeni U wymiaru d

Dla zadanego układu współrzędnych operator rzutu na podprzestrzeń d -wymiarową można utożsamić z macierzą C wymiaru $d \times p$, której wiersze są wektorami bazy (ortonormalnej) przestrzeni rzutu U :

$$C = \begin{bmatrix} C_1^T \\ C_2^T \\ \vdots \\ C_d^T \end{bmatrix}$$

Twierdzenie 3.6. *Macierz C^* odpowiadająca rzutowi na optymalną podprzestrzeń dla macierzy X o macierzy wariancji $V = V(X)$ spełnia warunek:*

$$C^* = \operatorname{argmax}(Tr(CVC^T))$$

z warunkiem ubocznym

$$\|C_i\| = 1, \quad i = 1, 2, \dots, d$$

Korzystając z metody mnożników Lagrange'a zadanie maksymalizacji z poprzedniego twierdzenia może być przedstawione jako zagadnienie maksymalizacji funkcji, gdzie zmienną jest macierz C wymiaru $d \times p$

$$Tr(CVC^T) - Tr(C^T \Lambda C)$$

Macierz $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_d)$ jest diagonalna. Wartości λ_i muszą być tak dobrane, aby spełniony był warunek $\|C_i\| = 1, \quad i = 1, 2, \dots, d$

Skorzystamy tu z warunku dostatecznego na istnienie ekstremum funkcji rzeczywistej z argumentem macierzowym (uogólnienie pojęcia gradientu).

Definicja 3.7. Pochodna macierzowa funkcji rzeczywistej macierzy D rozmiaru $d \times p$ jest macierzą rozmiaru $d \times p$:

$$\left[\frac{\partial f}{\partial D}(D) \right]_{ij} = \frac{\partial f(D)}{\partial D_{ij}}$$

Twierdzenie 3.7. *Jeżeli funkcja $f(D)$ ma ekstremum dla macierzy D_0 to*

$$\left. \frac{\partial f}{\partial D}(D) \right|_{D=D_0} = 0$$

Lemat 3.1. *Dla symetrycznej macierzy W*

$$\frac{\partial Tr(A^T W A)}{\partial A} = 2W A \frac{\partial Tr(BW B^T)}{\partial B} = 2BW$$

Twierdzenie 3.8. *Macierz C odpowiadająca rzutowi na optymalną podprzestrzeń dla macierzy X o macierzy wariancji $V = V(X)$ (twierdzenie 3.6) spełnia równanie*

$$CV = \Lambda C$$

Wniosek 3.7. *Wiersze macierzy C (**składowe główne**) są wektorami własnymi macierzy V , odpowiadającymi jej d największym wartościom własnym $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$*

Wniosek 3.8. *Współczynnik determinacji najlepszego modelu d -wymiarowego wyraża się wzorem;*

$$r^2 = \frac{\sum_{i=1}^d \lambda_i}{\sum_{i=1}^p \lambda_i}$$

3.1 Reifikacja modelu

Składowe główne są zbiorem nowych cech w d -wymiarowej przestrzeni najbliższej w sensie odległości Frobeniusa danym z macierzy X . Interesujące jest uzyskanie praktycznej interpretacji tych nowych cech. Taki proces nazywa się **reifikacją** modelu.

Zazwyczaj wszystkie analizy prowadzi się dla danych scentrowanych X^0 , to znaczy w układzie współrzędnych, którego początek umieszcza się w środku ciężkości X .

Model danych X w przestrzeni składowych głównych (ang. *PCA scores*) jest macierzą $Y = X^0 C^T$.

O związku między kolumnami X^0 i Y świadczy macierz korelacji $R(X, Y)$

Table 3.1: Wyniki 5 najlepszych zawodniczek.

	hurdles	highjump	shot	run200m	longjump	javelin	run800m
Joyner-Kersey (USA)	12.69	1.86	15.80	22.56	7.27	45.66	128.51
John (GDR)	12.85	1.80	16.23	23.65	6.71	42.56	126.12
Behmer (GDR)	13.20	1.83	14.20	23.10	6.68	44.54	124.20
Sablovskaite (URS)	13.61	1.80	15.23	23.92	6.25	42.78	132.24
Choubenkova (URS)	13.51	1.74	14.76	23.93	6.32	47.46	127.90

Propozycja 3.1. Niech $V = V(X) = [v_{ij}]$ będzie macierzą wariancji X , $C = [c_{ij}]$ macierzą, której wiersze są składowymi głównymi

$$r_{ij} \stackrel{\text{def}}{=} [R(X, Y)]_{ij} = \sqrt{\frac{\lambda_j}{v_{ii}}} c_{ij}$$

Propozycja 3.2.

$$\sum_{j=1}^p r_{ij}^2 = 1$$

Liczba r_{ij} jest korelacją między X^i a Y^j , czyli cosinusem kąta między nimi. Im ten kąt mniejszy (a więc korelacja większa) tym bardziej nowa zmienna Y^j jest związana z X^i . Propozycja 3.2, mówi że wektor $[r_{i1}, r_{i2}, \dots, r_{ip}]$, reprezentujący zmienną X^i leży na sferze jednostkowej w układzie współrzędnych korelacyjnych między X^i a Y^1, Y^2, \dots, Y^p .

Wykres w układzie dwóch pierwszych składowych korelacyjnych nazywa się *kołem korelacyjnym*. Pokazuje on jak “stare” zmienne są związane z dwiema najważniejszymi składowymi głównymi.

3.2 Przykład

Dane są związane z wynikami zawodów olimpijskich w Seulu (1988) w siedmioboju kobiet.

System punktowy przelicza wyniki 7 dyscyplin na punkty. Porównamy ten system punktów z modelem składowych głównych.

Słowniczek

hurdles = 100 m płotki

shot = pchnięcie kulą

javelin = oszczep

```
load("siedmioboj.Rda")
hm <- as.matrix(siedmioboj[,1:7])
knitr::kable(
  hm[1:5,], booktabs = TRUE,
  caption = 'Wyniki 5 najlepszych zawodniczek.'
)
```

```
kor <- round(cor(hm), 2)
knitr::kable(
  kor, booktabs = TRUE,
  caption = 'Korelacje między zmiennymi'
)
```

```
siedmioboj_pca <- prcomp(hm, scale = TRUE)
knitr::kable(
  siedmioboj_pca$rotation, booktabs = TRUE, digits=4,
```

Table 3.2: Korelacje między zmiennymi

	hurdles	highjump	shot	run200m	longjump	javelin	run800m
hurdles	1.00	-0.81	-0.65	0.77	-0.91	-0.01	0.78
highjump	-0.81	1.00	0.44	-0.49	0.78	0.00	-0.59
shot	-0.65	0.44	1.00	-0.68	0.74	0.27	-0.42
run200m	0.77	-0.49	-0.68	1.00	-0.82	-0.33	0.62
longjump	-0.91	0.78	0.74	-0.82	1.00	0.07	-0.70
javelin	-0.01	0.00	0.27	-0.33	0.07	1.00	0.02
run800m	0.78	-0.59	-0.42	0.62	-0.70	0.02	1.00

Table 3.3: Składowe główne

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
hurdles	0.4529	-0.1579	-0.0451	0.0265	-0.0949	-0.7833	-0.3802
highjump	-0.3772	0.2481	0.3678	-0.6800	-0.0188	-0.0994	-0.4339
shot	-0.3631	-0.2894	-0.6762	-0.1243	-0.5117	0.0509	-0.2176
run200m	0.4079	0.2604	0.0836	-0.3611	-0.6498	0.0250	0.4534
longjump	-0.4562	0.0559	-0.1393	-0.1113	0.1843	-0.5902	0.6121
javelin	-0.0754	-0.8417	0.4716	-0.1208	-0.1351	0.0272	0.1729
run800m	0.3750	-0.2245	-0.3959	-0.6034	0.5043	0.1556	0.0983

```
caption = 'Składowe główne'
)
```

scale=TRUE oznacza, że używamy danych standaryzowanych

```
lambdy <- siedmioboj_pca$sdev^2
lmb_tab <- data.frame(lambda=round(lambdy,4),r2=round(cumsum(lambdy)/7,2))
lmb_tab <- t(lmb_tab)
colnames(lmb_tab) <- 1:7
knitr::kable(
  lmb_tab, booktabs = TRUE,
  caption = 'Lambdy i współczynniki determinacji jako funkcja wymiaru modelu d'
)
```

```
punkty_pca <- -predict(siedmioboj_pca)[,1]
pca_sco <- data.frame(score=siedmioboj$score,pca=punkty_pca)
rownames(pca_sco) <- rownames(siedmioboj)
knitr::kable(
  t(pca_sco[1:5,]), booktabs = TRUE,digits = 2,
  caption = 'Porównanie punktów siedmioboju i punktów z pca dla 5 zawodniczek'
```

Table 3.4: Lambdy i współczynniki determinacji jako funkcja wymiaru modelu d

	1	2	3	4	5	6	7
lambda	4.4603	1.1943	0.521	0.4572	0.2453	0.073	0.049
r2	0.6400	0.8100	0.880	0.9500	0.9800	0.990	1.000

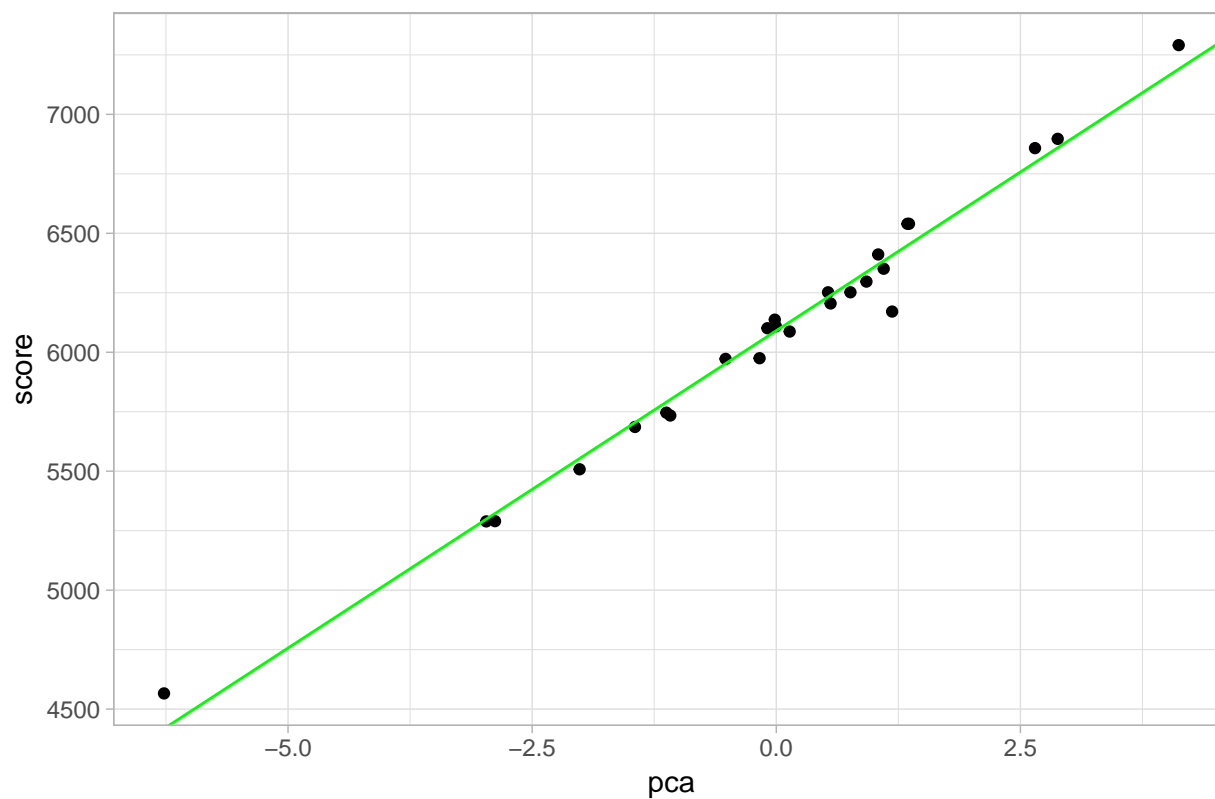
Table 3.5: Porównanie punktów siedmioboju i punktów z pca dla 5 zawodniczek

	Joyner-Kersey (USA)	John (GDR)	Behmer (GDR)	Sablovskaite (URS)	Choubenkova (URS)
score	7291.00	6897.00	6858.00	6540.00	6540.00
pca	4.12	2.88	2.65	1.34	1.36

)

Znak “-” przy `predict(siedmioboj_pca)[,1]` został wybrany tak, aby zwrot wektora `pca` i `score` był taki sam

$$\text{score} = 266.77 \cdot \text{pca} + 6090.6 \quad r^2 = 0.982$$

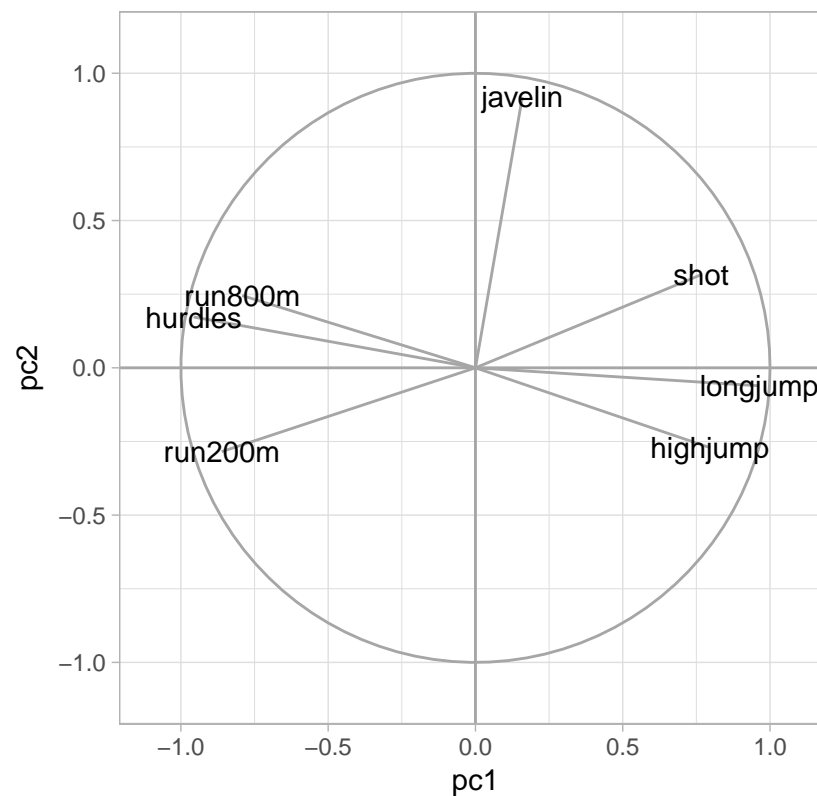


```
require(FactoMineR)
pca.FM <- PCA(hm, graph = FALSE)
knitr::kable(
  pca.FM$var$coord, booktabs = TRUE, digits = 2,
  caption = 'Korelacje zmiennych oryginalnych ze składowymi głównymi'
)
```

Table 3.6: Korelacje zmiennych oryginalnych ze składowymi głównymi

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
hurdles	-0.96	0.17	0.03	-0.02	0.05
highjump	0.80	-0.27	-0.27	0.46	0.01
shot	0.77	0.32	0.49	0.08	0.25
run200m	-0.86	-0.28	-0.06	0.24	0.32
longjump	0.96	-0.06	0.10	0.08	-0.09
javelin	0.16	0.92	-0.34	0.08	0.07
run800m	-0.79	0.25	0.29	0.41	-0.25

Kolo korelacyjne



Składowa główna pc1 jest dodatnio i silnie skorelowana ze sportami siłowymi i z prędkościami (z czasem biegu ujemnie!). Jedynym niepasującym sportem jest rzut oszczepem, związanym z drugą składową główną. Czyli punktacja, silnie związana ze składową 1 *nie uwzględnia w istocie rzutu oszczepem*.

Chapter 4

Grupowanie - analiza skupień

4.1 Podział macierzy danych na klasy

Definicja 4.1. Podział macierzy danych X na k klas wyznaczony jest przez podział indeksów wierszy na rozłączne zbiory I_1, I_2, \dots, I_k . Wyznaczone są w ten sposób podmacierze

$$X_{[j]} = \begin{bmatrix} X_{i_1} \\ X_{i_2} \\ \vdots \\ X_{i_{n_j}} \end{bmatrix}$$

gdzie

$$I_j = \{i_1, i_2, \dots, i_{n_j}\}$$

Macierz X można zapisać ¹:

$$X = \begin{bmatrix} X_{[1]} \\ X_{[2]} \\ \vdots \\ X_{[k]} \end{bmatrix}$$

Definicja 4.2. Centroidem G_j j -tej klasy podziału \mathcal{P} nazywamy środek ciężkości $X_{[j]}$:

$$G_j = g(X_{[j]})$$

Macierz

$$G_{\mathcal{P}} = \begin{bmatrix} G_1^{n_1} \\ G_1^{n_2} \\ \vdots \\ G_1^{n_k} \end{bmatrix}$$

nazywamy **macierzą centroidów podziału \mathcal{P}**

Lemat 4.1. Niech G_1, G_2, \dots, G_k będą centroidami podziału, $G = g(X)$.

1. G jest wypukłą kombinacją G_j :

$$G = \sum_{i=1}^k p_i G_i, \quad p_i = \frac{n_i}{n}, \quad \sum_{i=1}^k p_i = 1$$

2. G jest środkiem ciężkości $G_{\mathcal{P}}$.

¹po odpowiednim przestawieniu wierszy

Dobry podział charakteryzuje się dwiema cechami:

1. Dane w podmacierzach $X_{[1]}, X_{[2]}, \dots, X_{[k]}$ są maksymalnie zwarte ²
2. Centroidy podziału są maksymalnie odległe ³

Realizacją postulatów 1. jest aby bezwładności $J(X_{[j]})$ były małe, zaś postulat 2. - aby bezwładność $J(G_{\mathcal{P}})$ była jak największa.

Zazwyczaj trudno jest zrealizować jednocześnie tak dwa sprzeczne cele. Okaże się, że (twierdzenie 4.1) wystarczy realizować jeden z tych postulatów

Definicja 4.3. Bezwładność wewnątrzklasowa podziału \mathcal{P} jest liczbą

$$J_W(\mathcal{P}) = \sum_{j=1}^k p_j J(X_{[j]})$$

Bezwładność międzyklasowa podziału \mathcal{P} jest liczbą $J_M(\mathcal{P}) = J(G_{\mathcal{P}})$

Twierdzenie 4.1. Twierdzenie Pitagorasa dla podziału Dla każdego podziału \mathcal{P}

$$J(X) = J_W(\mathcal{P}) + J_M(\mathcal{P})$$

Wniosek 4.1. $J_W(\mathcal{P})$ maleje wtedy i tylko wtedy gdy $J_M(\mathcal{P})$ rośnie

Definicja 4.4. Porządek między podziałami

$$\mathcal{P} \succ \mathcal{Q} \iff J_M(\mathcal{P}) > J_M(\mathcal{Q}) \quad \mathcal{P} \succeq \mathcal{Q} \iff J_M(\mathcal{P}) \geq J_M(\mathcal{Q})$$

Propozycja 4.1.

$$J_M(\mathcal{P}) = \sum_{j=1}^k p_j \|G_j - G\|^2 \quad J_W(\mathcal{P}) = \frac{1}{n} \sum_{j=1}^k \sum_{i \in I_j} \|X_i - G_j\|^2$$

4.1.1 Przykład

4.1.1.1 Użyteczne funkcje w R

```
SrodekCiezkosci <- function(dane){
  g <- apply(dane, MARGIN = 2, mean)
  return(list(g = g, n = nrow(dane)))
}
```

```
srodki_ciezkosci <- function(dane, podzial){
  require(dplyr)
  k <- length(levels(podzial))
  nc <- ncol(dane)
  gg <- numeric(k*(nc+1))
  gg <- array(gg, dim=c(k,nc+1))
  gg <- as.data.frame(gg)
  colnames(gg) <- c(colnames(dane), "n")

  for (i in 1:k) {
    dane %>%
      as.data.frame() %>%
      filter(podzial == i) %>%
```

²Dane wewnątrz klas są do siebie podobne

³Klasy są do siebie niepodobne

```

  SrodekCiezkosci() -> gg_rob
  gg[i,1:nc] <- gg_rob$g
  gg[i,(nc+1)] <- gg_rob$n
}
return(gg)
}

JM <- function(dane, podzial){
  gg <- srodki_ciezkosci(dane, podzial)
  g <- SrodekCiezkosci(dane)$g
  jm <- 0
  for (j in 1: (ncol(gg)-1))
    jm <- jm + gg$n %*%(gg[,j]-g[j])^2
  jm <- as.numeric(jm/nrow(dane))
  return(jm)
}

```

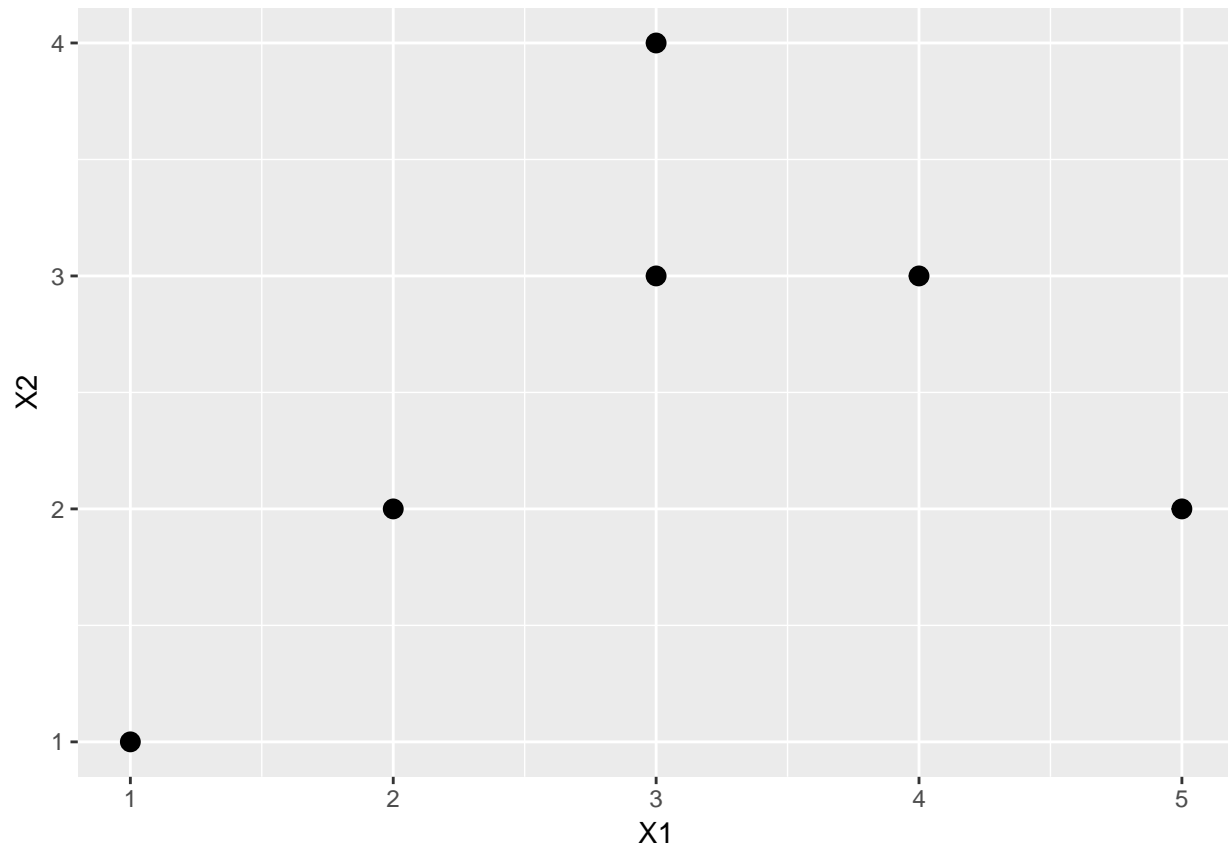
4.1.1.2 Dane

```

##   X1 X2
## 1  1  1
## 2  2  2
## 3  3  3
## 4  3  4
## 5  4  3
## 6  5  2

ggplot(X,aes(X1,X2)) + geom_point(size=3)

```



4.1.1.3 Porównanie podziałów

```
##   X1 X2 P1 P2 P3
## 1  1  1  2  1  1
## 2  2  2  2  2  1
## 3  3  3  2  1  2
## 4  3  4  2  2  2
## 5  4  3  1  2  2
## 6  5  2  1  1  2
```

```
p1_plot <- ggplot(XP,aes(X1,X2)) + geom_point(size=3,aes(color=P1))
p2_plot <- ggplot(XP,aes(X1,X2)) + geom_point(size=3,aes(color=P2))
p3_plot <- ggplot(XP,aes(X1,X2)) + geom_point(size=3,aes(color=P3))
grid.arrange(p1_plot,p2_plot,p3_plot,nrow=2)
```

Środki ciężkości

```
data.frame(
  Podzial=c("P1 K1","P1 K2","P2 K1","P2 K2","P3 K1","P3 K2"),
  rbind(
    srodki_ciezkosci(X,P1),
    srodki_ciezkosci(X,P2),
    srodki_ciezkosci(X,P3)
  )
)
```

```
##   Podzial   X1  X2 n
```

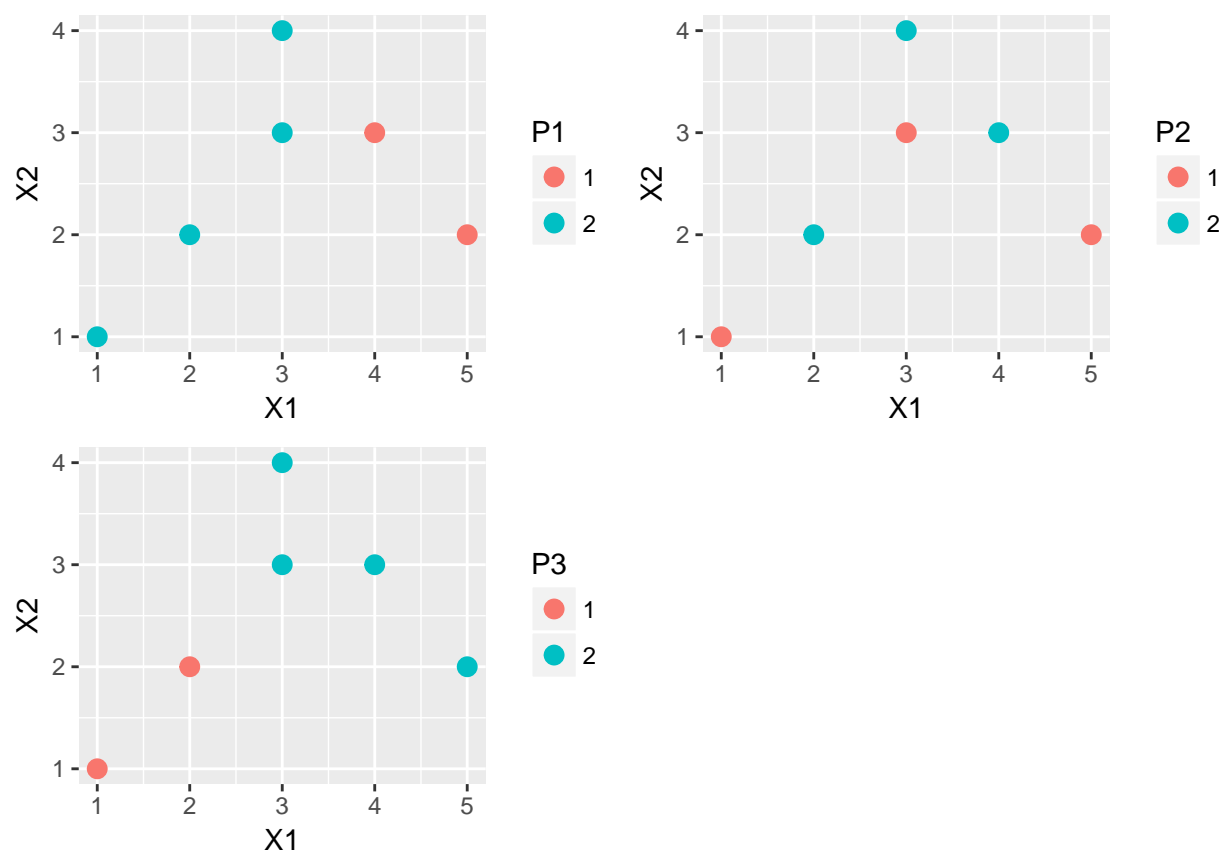


Figure 4.1: Trzy podziały tych samych danych

```
## 1   P1 K1 4.50 2.5 2
## 2   P1 K2 2.25 2.5 4
## 3   P2 K1 3.00 2.0 3
## 4   P2 K2 3.00 3.0 3
## 5   P3 K1 1.50 1.5 2
## 6   P3 K2 3.75 3.0 4
```

Bezwładność międzyklasowa i wewnątrzklasowa

```
jm1 <- JM(X,P1)
jm2 <- JM(X,P2)
jm3 <- JM(X,P3)
(J <- sum(diag(cov(X))))
```

```
## [1] 3.1
```

```
bzw1 <- rbind(
  c(jm1,J-jm1),
  c(jm2,J-jm2),
  c(jm3,J-jm3)
)
colnames(bzw1) <- c("JM","JW")
rownames(bzw1) <- c("P1","P2","P3")
bzw1
```

```
##      JM    JW
## P1 1.125 1.975
## P2 0.250 2.850
## P3 1.625 1.475
```

$P3 \succ P1 \succ P2$

4.2 Podział Woronoja

Definicja 4.5. Podziałem Woronoja przestrzeni \mathbf{R}^p o centrach C_1, C_2, \dots, C_k jest rodzina podzbiorów \mathbf{R}^p :

$$W_j^* = \{P \in \mathbf{R}^p : \|P - C_j\| = \min\{\|P - C_r\|, r = 1, 2, \dots, k\}\}$$

Źródło: <https://chichacha.netlify.com/2018/11/10/voronoi-diagram-with-ggvoronoi-package-with-train-station-data/>

Zbiory W_j^* nie są rozłączne (mają wspólne granice). Można je skonstruować tak, aby uzyskały rozłączność:

$$W_j = W_j^* - \bigcup_{r=1}^{j-1} W_r^*$$

Definicja 4.6. Podziałem Woronoja macierzy danych X o centrach C_1, C_2, \dots, C_k jest rodzina macierzy $X_{[j]}$ taka, że

$$I_j = \{i : X_i \in W_j\}$$

Twierdzenie 4.2. Niech \mathcal{P} będzie podziałem macierzy X na k klas $X_{[1]}, X_{[2]}, \dots, X_{[k]}$, $g(X_{[j]}) = G_j$ ($j = 1, 2, \dots, k$) - centroidami podziału \mathcal{P} . Niech \mathcal{Q} będzie podziałem Woronoja o centrach G_1, G_2, \dots, G_k . Wtedy $\mathcal{Q} \succeq \mathcal{P}$.

4.2.1 Przykład (cd)

Korzystając z twierdzenia 4.2 znajdziemy dla każdego z podziałów $P1, P2, P3$ maksymalnie najlepszy podział

東京都 (Tokyo)

943 stations 80 lines operated by 17 companies in Tokyo

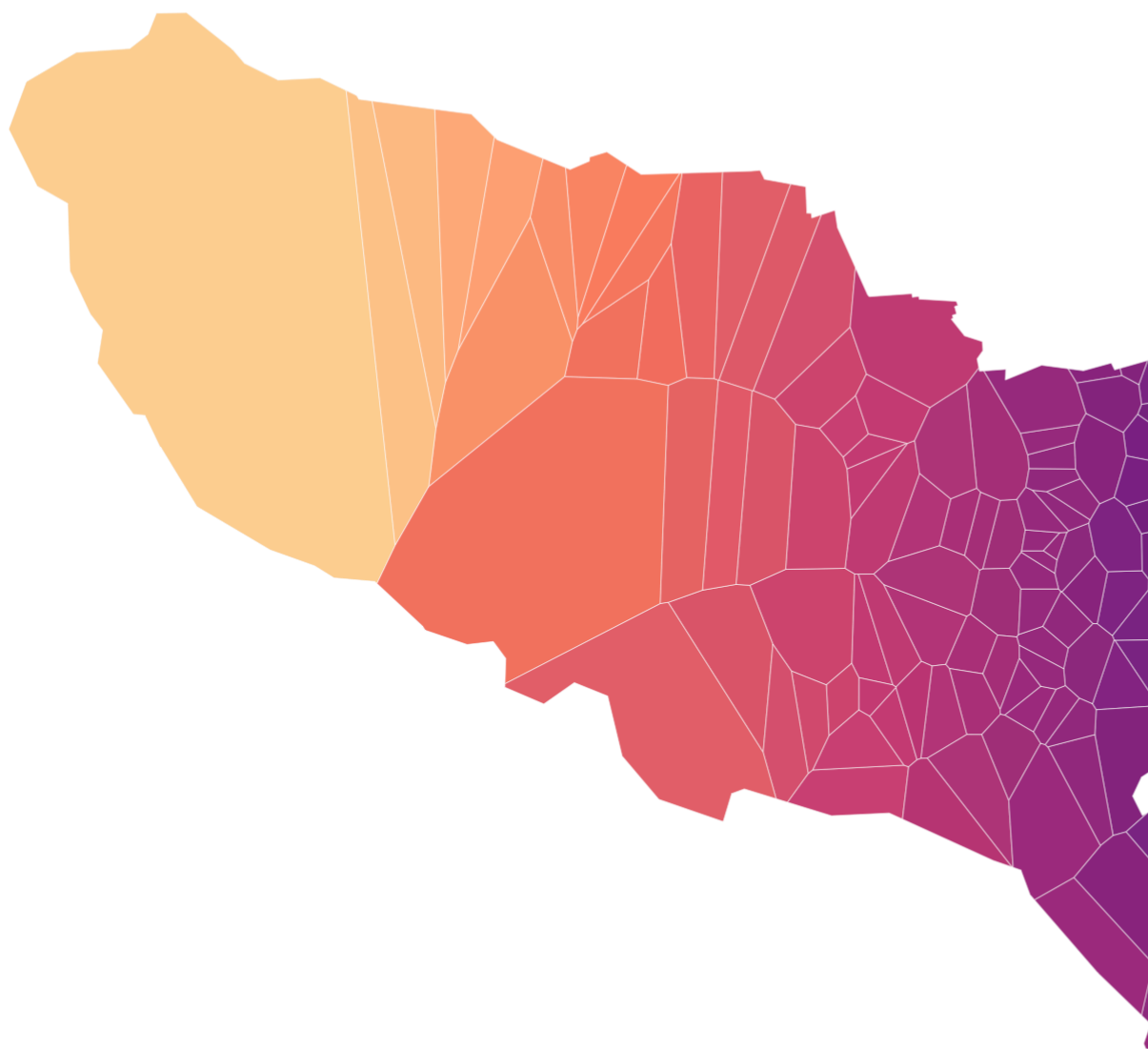


Figure 4.2: Mapa Tokio z podziałem Woronoja. Centrami są stacje kolejowe. Kolor zależy od odległości od stacji Shinjuku - centralnej stacji w Tokio (biały krzyżyk)

4.2.1.1 Użyteczne funkcje w R

```

deuc <- function(x,y){ # odleglosc euklidesowa
  sum((x-y)^2)
}

odSc <- function(x,sc){ #odleglosc x od centroidów sc
  apply(sc,1,function(y) deuc(x,y))
}

prox <- function(x,sc){ # podaje numer centroidu najblizszego punktowi x
  which.min(odSc(x,sc))
}

proxV <- function(dane,sc){# podaje numery centroidow najblizszych punktom danych dane
  as.factor(apply(dane,1,function(x) prox(x,sc)))
}

proxVpod <- function(dane,podzial){# podaje numery klas najblizszych punktom danych dane
  nc <- ncol(dane)
  sc <- srodki_ciezkosci(dane,podzial)[,1:nc]
  proxV(dane,sc)
}

proxWyn <- function(dane,podzial){# sprawdza czy nowy podzial lepszy od starego
  w <- proxVpod(dane,podzial)
  rowne <- all.equal.factor(podzial,w)
  return(list(stare=podzial,nowe=w,rowne=rowne))
}

```

4.2.1.2 Poprawianie podziałów

```
(P11 <- proxWyn(X,P1))
```

```

## $stare
## [1] 2 2 2 2 1 1
## Levels: 1 2
##
## $nowe
## [1] 2 2 2 2 1 1
## Levels: 1 2
##
## $rowne
## [1] TRUE

```

Podział $P1$ jest podziałem Woronoja

```
(P21 <- proxWyn(X,P2))
```

```

## $stare
## [1] 1 2 1 2 2 1
## Levels: 1 2
##
## $nowe

```

```
## [1] 1 1 2 2 2 1
## Levels: 1 2
##
## $rowne
## [1] "2 string mismatches"
(P22 <- proxWyn(X,P21$nowe))
```

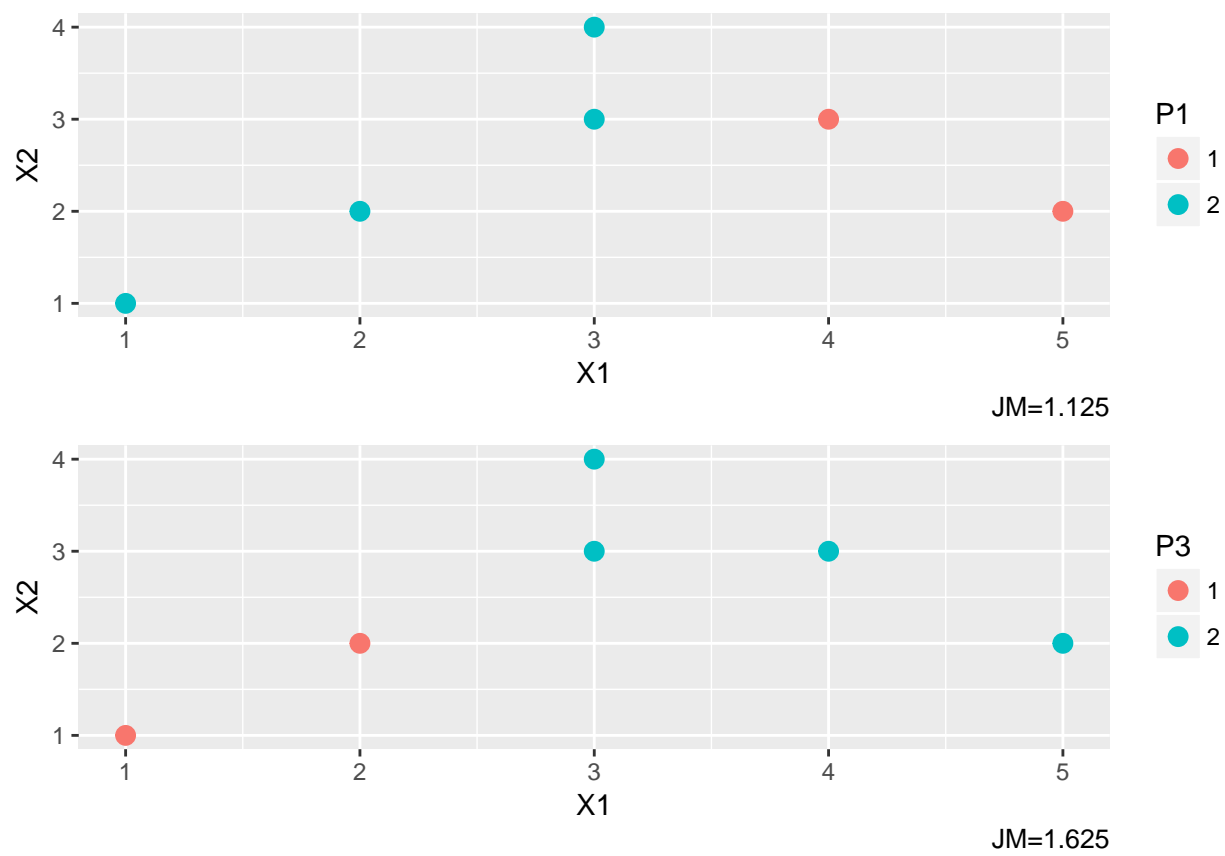
```
## $stare
## [1] 1 1 2 2 2 1
## Levels: 1 2
##
## $nowe
## [1] 1 1 2 2 2 2
## Levels: 1 2
##
## $rowne
## [1] "1 string mismatch"
(P23 <- proxWyn(X,P22$nowe))
```

```
## $stare
## [1] 1 1 2 2 2 2
## Levels: 1 2
##
## $nowe
## [1] 1 1 2 2 2 2
## Levels: 1 2
##
## $rowne
## [1] TRUE
```

Podział $P22\$nowe = 112222$ jest lepszy podziału $P2 = 121221$ i jest równy podziałowi $P3$. Z tego wynika, że $P3$ jest podziałem Woronoja. Nie wiadomo, czy są inne podziały Woronoja dla tych danych

```
p1_plot <- ggplot(XP,aes(X1,X2)) + geom_point(size=3,aes(color=P1)) + labs(caption="JM=1.125")
p3_plot <- ggplot(XP,aes(X1,X2)) + geom_point(size=3,aes(color=P3)) + labs(caption="JM=1.625")
grid.arrange(p1_plot,p3_plot,nrow=2)
```

Z tych dwóch podziałów Woronoja podział $P3$ jest lepszy.

Figure 4.3: Podziały Woronoja dla danych X

Chapter 5

Dyskryminacja (klasyfikacja z nauczycielem)

Zakładamy, że zadany jest podział \mathcal{P} macierzy danych X rozmiaru $n \times p$. Podział ten odzwierciedla stan naszej wiedzy o zebranych (historycznych) danych i ich podziale na klasy. Przykładem takiej sytuacji jest medycyna, gdzie X jest macierzą p objawów zgromadzonych wśród n pacjentów, zaś podział odzwierciedla klasyfikację na k jednostek chorobowych. Zadaniem dyskryminacji jest opracowanie prostych reguł zakwalifikowania do jednej z klas obiektu $x \in \mathbf{R}^p$ nie odwołujących się do macierzy X ¹. W tym rozdziale zajmiemy się **dyskryminacją liniową**, w której reguły dyskryminacji oparte są na operatorach liniowych w \mathbf{R}^p .

Definicja 5.1. Niech podmacierzami podziału \mathcal{P} będą $X_{[1]}, X_{[2]}, \dots, X_{[k]}$, G_1, G_2, \dots, G_k - centroidami podziału, $G_{\mathcal{P}}$ - macierzą centroidów. **Macierzą wariancji wewnątrzklasowej (odp. międzyklasowej)** nazywamy macierz

$$V_W \stackrel{def}{=} V_W(X, \mathcal{P}) = \sum_{j=1}^k p_j V(X_{[j]}) \quad V_M \stackrel{def}{=} V_M(X, \mathcal{P}) = V(G_{\mathcal{P}})$$

Twierdzenie 5.1. Twierdzenie Pitagorasa o macierzach wariancji podziału

Dla każdego podziału \mathcal{P} zachodzi

$$V(X) = V_W(X, \mathcal{P}) + V_M(X, \mathcal{P})$$

Propozycja 5.1.

$$V_M = \sum_{j=1}^k p_j (G_j - G)(G_j - G)^T \text{rz}(V_M) \leq k - 1$$

5.1 Zmienne dyskryminacyjne

Podobnie jak w przypadku wyboru najlepszego modelu liniowego, szukamy macierzy przekształcenia liniowego wymiaru $d \times p$ ($d \leq p$):

$$U = \begin{bmatrix} U_1^T \\ U_2^T \\ \vdots \\ U_d^T \end{bmatrix}$$

której wiersze stanowią zmienne dyskryminacyjne.

¹Może to być macierz gigantycznych rozmiarów

Po przekształceniu U obrazem macierzy X będzie macierz $Y = XU^T$ o wymiarze $n \times d$. Z twierdzenia 5.1 wiemy, że

$$V(Y) = V_W(Y, \mathcal{P}) + V_M(Y, \mathcal{P})$$

Należy wybrać takie przekształcenie U , aby zdolność rozróżnienia klas była jak największa czyli bezwładność międzyklasowa

$$J_M(Y, \mathcal{P}) = \text{Tr}(V_M(Y, \mathcal{P}))$$

była jak największa.

Propozycja 5.2.

$$V_M(Y, \mathcal{P}) = UV_M(X, \mathcal{P})U^T$$

Ponieważ zarówno macierz X jak i podział \mathcal{P} są ustalone, przyjmujemy oznaczenia:

$$V \stackrel{\text{def}}{=} V(X), V_M \stackrel{\text{def}}{=} V_M(X, \mathcal{P})$$

Twierdzenie 5.2. *Optymalna macierz zmiennych dyskryminacyjnych spełnia warunek*

$$U^* = \text{argmax}(\text{Tr}(UV_MU^T))$$

z warunkiem ubocznym

$$\|Y_i\| = 1, i = 1, 2, \dots, d$$

Korzystając z metody mnożników Lagrange’a zadanie to można sprowadzić do maksymalizacji funkcji zmiennej macierzowej U :

$$\text{Tr}(UV_MU^T) - \text{Tr}(\Lambda UVU^T)$$

Macierz Λ jest przekątniową macierzą rozmiaru $d \times d$ współczynników Lagrange’a tak dobranych by $\|Y_i\| = 1, i = 1, 2, \dots, d$.

Twierdzenie 5.3. *Wiersze optymalnej macierzy zmiennych dyskryminacyjnych U^* są wektorami własnymi macierzy $V^{-1}V_M$ odpowiadającymi $\min(d, k-1)$ największym wartościom własnym². Wektory U_i mają długości spełniające warunki $U_i^T V U_i = 1$*

5.2 Podział dychotomiczny

Często w praktyce występuje podział na dwie klasy. Jest on o tyle interesujący, że jak wynika z twierdzenia 5.3, jest tylko jedna zmienna dyskryminacyjna, a więc kryterium dyskryminacyjne jest oparte na iloczynie skalarnym tej zmiennej z wektorem x mającym być obiektem klasyfikacji

Propozycja 5.3. *W przypadku podziału dychotomicznego, w którym frakcja przypadków próby uczacej, należących do klasy 1 jest równa p_1 a należących do klasy 2 $p_2 = 1 - p_1$, $G_j, j = 1, 2$ są centroidami klas*

$$V_M = p_1 p_2 (G_1 - G_2)(G_1 - G_2)^T u_0 = V^{-1}(G_1 - G_2)$$

u_0 jest równoległy do wektora dyskryminacyjnego

Propozycja 5.4. *Kryterium przynależności do klasy 1*

$$u_0^T x \geq c \stackrel{\text{def}}{=} \frac{u_0^T (G_1 + G_2)}{2}$$

Jeżeli próg c jest różny od 0, to wygodnie jest przyjąć ustaloną z góry (niezależnie od zadania dyskryminacji) wartość progu, równą c^* ³.

²a więc wektorów dyskryminacyjnych jest na ogół mało

³Ja (AD) przyjmuję wartość 100

Propozycja 5.5. *Kryterium przynależności do klasy 1 dla uniwersalnego progu.*

Niech $\alpha = \frac{c^*}{c}$, $u_* = \alpha u_0$

$$\begin{cases} u_*^T x \geq c^* & \text{gdy } \alpha > 0 \\ u_*^T x \leq c^* & \text{gdy } \alpha < 0 \end{cases}$$

5.2.1 Przykład

Badano zmiany zawartej w płazmie krwi stężenia glukozy [%] (zmienna 1) i wolnego kwasu tłuszczowego [mEq/l] u 12 schizofreników (grupa 1) i 13 zdrowych ochotników (grupa 2) po domięśniowym wstrzyknięciu insuliny.

Dane

Środki ciężkości

```
##                G1      G2
## glukoza [%]      -25.60 -31.10
## tluszcz [mEq/l]  -0.06  -0.15
```

Macierze kowariancji

```
##          [,1]  [,2]
## [1,] 278.0830 0.8291
## [2,]  0.8291 0.0092

##          [,1]  [,2]
## [1,] 269.9230 -0.2493
## [2,] -0.2493  0.0067
```

Obliczenia

```
p1 <- 12/25
p2 <- 13/25
g <- p1*g1+p2*g2
g
```

```
## [1] -28.4600 -0.1068
```

```
VW <- p1*V1+p2*V2
VW
```

```
##          [,1]  [,2]
## [1,] 273.839800 0.268332
## [2,]  0.268332 0.007900

VM <- p1*p2*(g1-g2)%*%t(g1-g2)
VM
```

```
##          [,1]  [,2]
## [1,] 7.550400 0.12355200
## [2,] 0.123552 0.00202176
```

```
V <- VW+VM
V
```

```
##          [,1]  [,2]
## [1,] 281.390200 0.39188400
## [2,]  0.391884 0.00992176
```

```
Vinv <- solve(V)
Vinv
```

```
##           [,1]      [,2]
## [1,]  0.003760646 -0.1485358
## [2,] -0.148535828 106.6553529
```

```
u0 <- Vinv %*% (g1-g2)
u0
```

```
##           [,1]
## [1,] 0.007315326
## [2,] 8.782034709
```

Punkt podziału

```
c <- 0.5 * sum(t(u0) * (g1+g2))
c
```

```
## [1] -1.129503
```

Reguła 100

```
ug <- 100*u0/c
ug
```

```
##           [,1]
## [1,]  -0.6476588
## [2,] -777.5130830
```

Dyskryminacja

```
x <- c(-30,-0.1)
x
```

```
## [1] -30.0 -0.1
```

```
sum(t(ug) * x)
```

```
## [1] 97.18107
```

Chory

```
x <- c(-20,-0.2)
x
```

```
## [1] -20.0 -0.2
```

```
sum(t(ug) * x)
```

```
## [1] 168.4558
```

Zdrowy

Chapter 6

Literatura

1. Le Roux, B., Ruanet, H. , Geometric data analysis: from correspondence analysis to structured data analysis, Kluwer
2. Du Toit, S.H.C., Steyn, A.G.W., Stumpf, R.H., Graphical Exploratory Data Analysis, Springer 1986
3. Gnanadesikan, R., Statistical Data Analysis, Lecture Notes for AMS, 1982, Proc. of Symposia in Applied Mathematics vol. 28, 1983
4. Hoaglin, D.C., Mosteller, F., Tukey J.W., Understanding Robust and Exploratory Data Analysis, Wiley 2000
5. Hoaglin, D.C., Mosteller, F., Tukey J.W., Exploring Data Tables Trends and Shapes, Wiley 2000
6. Krzanowski., W.J., Principles of Multivariate Analysis, A User's Perspective, Oxford University Press, 2000
7. Tukey, J.W., Exploratory Data Analysis, Reading 1977