Zagadnienie klasyfikacji (dyskryminacji)

Przykład. Bank chce klasyfikować klientów starających się o pożyczkę do jednej z dwóch grup: niskiego ryzyka (spłacających pożyczki terminowo) lub wysokiego ryzyka. Obserwując pewne cechy charakteryzujące klienta należy skonstruować regułę postępowania klasyfikującą ewentualnych pożyczkobiorców do jednej z dwóch wymienionych grup.

Populacje: $\pi_1, \ldots, \pi_k$

Obiekt: $\mathbf{X} = (X_1, \ldots, X_p)$

Zadanie
Przypisać obiekt do jednej z populacji $\pi_1, \ldots, \pi_k$

Rozwiązanie: podział zbioru $\mathbf{R}^p$ na takie obszary $R_1, \ldots, R_p$, że

$$\bigcup_{i=1}^{k} R_i = \mathbf{R}^p, \qquad R_i \cap R_j = \emptyset, i \neq j$$

Reguła klasyfikacyjna (dyskryminacyjna)

Jeżeli $\mathbf{X} \in R_i$, to obiekt zaliczamy do π_i

Problem: znaleźć zbiory R_i

Kryterium

$P\{\mathbf{X} \in R_i |$ obiekt pochodzi z populacji $\pi_i\} = \max!$

Rozwiązanie zagadnienia

Założenia
1. Dla populacji π_i: $\mathbf{X} \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$
2. $\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$
3. $P\{$obiekt pochodzi z $\pi_i\} = 1/k$

Klasyfikacja dla dwóch populacji $k = 2$

Idea: obserwacja $\mathbf{X}$ pochodzi z tej populacji, dla której odległość obserwacji od wektora średnich jest mniejsza.

Formalnie: Niech $W(\mathbf{X}) =$

$$(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)'\mathbf{C}^{-1}\mathbf{X} - \frac{1}{2}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)'\mathbf{C}^{-1}(\bar{\mathbf{X}}_1 + \bar{\mathbf{X}}_2)$$

Reguła klasyfikacyjna
$\mathbf{X}$ zaklasyfikować do populacji π_1, jeżeli $W(\mathbf{X}) > 0$
$\mathbf{X}$ zaklasyfikować do populacji π_2, jeżeli $W(\mathbf{X}) < 0$

Funkcja $W(\mathbf{X})$: *funkcja dyskryminacyjna*

Klasyfikacja dla wielu populacji $k > 2$

Idea: obserwacja $\mathbf{X}$ pochodzi z tej populacji, dla której odległość obserwacji od wektora średnich jest najmniejsza.

Formalnie: Niech

$$W_{ij}(\mathbf{X}) =$$
$$(\bar{\mathbf{X}}_i - \bar{\mathbf{X}}_j)'\mathbf{C}^{-1}\mathbf{X} - \frac{1}{2}(\bar{\mathbf{X}}_i - \bar{\mathbf{X}}_j)'\mathbf{C}^{-1}(\bar{\mathbf{X}}_i + \bar{\mathbf{X}}_j)$$

Reguła klasyfikacyjna

obserwację $\mathbf{X}$ zaklasyfikować do populacji π_i,
jeżeli $W_{ij}(\mathbf{X}) > 0$ dla wszystkich $i \neq j$

Funkcje W_{ij} — *funkcje dyskryminacyjne*

Przykład. W celu oceny stopnia ryzyka udzielanych kredytów bankowych, wybrano losowo 26 klientów i 12 z nich oceniono jako klientów o niskim stopniu ryzyka (klienci spłacali pożyczki w terminie), zaś 14 klientów z wylosowanej grupy oceniono jako klientów o wysokim stopniu ryzyka (klienci ci nie spłacali pożyczek w terminie). Przyjmując, że spłata pożyczek w terminie jest funkcją następujących cech:

X_1 — płeć,
X_2 — okres współpracy z bankiem,
X_3 — liczba posiadanych dzieci,
X_4 — wielkość dochodu
X_5 — zaakceptowane oprocentowanie pożyczek,

skonstruować funkcję pozwalającą na ocenę czy ubiegający się o pożyczkę i posiadający określone cechy należy do grupy niskiego ryzyka, czy też należy do grupy wysokiego ryzyka.

π_1 — grupa niskiego ryzyka
π_2 — grupa wysokiego ryzyka

Nowy klient: $\mathbf{X} = (X_1, \ldots, X_5)'$

Funkcja dyskryminacyjna:

$$W(\mathbf{X}) = -0.98855 + 0.91522X_1 + 0.34271X_2 + 0.80272X_3 - 0.20583X_4 - 0.20061X_5.$$

Jeżeli $W(\mathbf{X}) < 0$, to klasyfikujemy klienta $\mathbf{X}$ do π_1.
Jeżeli $W(\mathbf{X}) > 0$, to klasyfikujemy klienta $\mathbf{X}$ do π_2.

Wniosek kredytowy złożył bezdzietny ($X_3 = 0$) mężczyzna ($X_1 = 0$) współpracujący z bankiem jeden rok ($X_2 = 1$) deklarujący uzyskiwany dochód na poziomie 500 złotych ($X_4 = 5$) oraz akceptujący 4% jako tygodniowe oprocentowanie pożyczki ($X_5 = 4$).

Wartość funkcji dyskryminacyjnej

$$\begin{aligned} W &= -0.98855 + 0.91522 \cdot 0 + 0.34271 \cdot 1 \\ &\quad + 0.80272 \cdot 0 - 0.20583 \cdot 5 - 0.20061 \cdot 4 \\ &= -2.4041. \end{aligned}$$

Ponieważ jest to wartość ujemna, więc klienta klasyfikujemy do grupy małego ryzyka.

Przykład. Przykład pochodzi od Fishera i przeszedł do klasyki przykładów analizy dyskryminacji. Badano trzy populacje kwiatów: *Iris virginica*, *Iris versicolor* oraz *Iris setosa*. Dla każdego kwiatu mierzono długość i szerokość działki kielicha (SL i SW) oraz długość i szerokość płatka (PL i PW).

Zadanie: na podstawie czterech pomiarów zaklasyfikować nowy kwiat do jednej z trzech populacji

Dla każdej z populacji dokonano po 50 obserwacji i uzyskano następujące średnie próbkowe

Iris	SL	SW	PL	PW
Virginica	6.588	2.974	5.552	2.062
Versicolor	5.936	2.770	4.260	1.326
Setosa	5.006	3.428	1.462	0.246

Macierz średnich kwadratów i iloczynów ma postać:

$$\mathbf{C} = \frac{1}{150-3}\begin{bmatrix} 102.17 & -6.59 & 189.51 & 77.12 \\ & 28.31 & -49.12 & -18.12 \\ & & 464.33 & 193.05 \\ & & & 86.57 \end{bmatrix}$$

Dwie funkcje dyskryminacyjne:

$$W_{12} = -3.246SL - 3.391SW + 7.553PL + 14.636PW - 31.523$$

$$W_{13} = -11.076SL - 19.916SW + 29.187PL + 38.461PW - 18.093$$

Reguła klasyfikacyjna ma postać:

Zaklasyfikować kwiat *Iris* o obserwacji $\mathbf{X}$ jako

virginica, jeżeli $W_{12}(\mathbf{X}) > 0$ i $W_{13}(\mathbf{X}) > 0$

versicolor, jeżeli $W_{12}(\mathbf{X}) < 0$ i $W_{13}(\mathbf{X}) > W_{12}(\mathbf{X})$

setosa, jeżeli $W_{12}(\mathbf{X}) < 0$ i $W_{13}(\mathbf{X}) < 0$

Analiza skupień

$\mathbf{X}_1, \ldots, \mathbf{X}_n$ — p–wymiarowe obserwacje jednostek

Założenie

Przyjmujemy, że obserwacje $\mathbf{X}_1, \ldots, \mathbf{X}_n$ pochodzą z nieznanej liczby k populacji.

Zadanie

Oszacować liczbę k populacji oraz rozpoznać, które obserwacje pochodzą z kolejnych populacji.

Grupy obserwacji uznane za pochodzące z tych samych populacji nazywane są *skupieniami* lub *segmentami* (ang. *cluster*).

Techniki analizy skupień zwane są *procedurami segmentacji* lub *aglomeracji*.

Idea

Dwie obserwacje uznajemy za pochodzące z tej samej populacji, jeżeli są dostatecznie „blisko" siebie.

Techniki segmentacji
— techniki hierarchiczne
— techniki optymalnego podziału
— techniki natężenia
— techniki grupowania

Metody hierarchiczne

Macierz odległości $[d_{ij}]$ między obiektami i skupieniami.

Odległość d_{ij} między obiektami

$$\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})' \qquad \mathbf{X}_j = (X_{j1}, \ldots, X_{jp})'$$

$$d_{ij} = \sum_{l=1}^{p} (X_{il} - X_{jl})^2$$

Zasada działania metod hierarchicznych

1. zakładamy, że każdy z obiektów tworzy jednoelementowe skupienie
2. w macierzy odległości między skupieniami szukamy takiej pary skupień q i r $(q < r)$ dla której odległość jest najmniejsza:

$$d_{qr} = \min_{i<j} d_{ij}$$

3. łączymy obiekty q i r w jedno skupienie, nadajemy mu numer q i wyznaczamy nową macierz odległości
4. powyższe kroki powtarzamy aż do uzyskania jednego skupienia

Metoda najbliższego sąsiedztwa

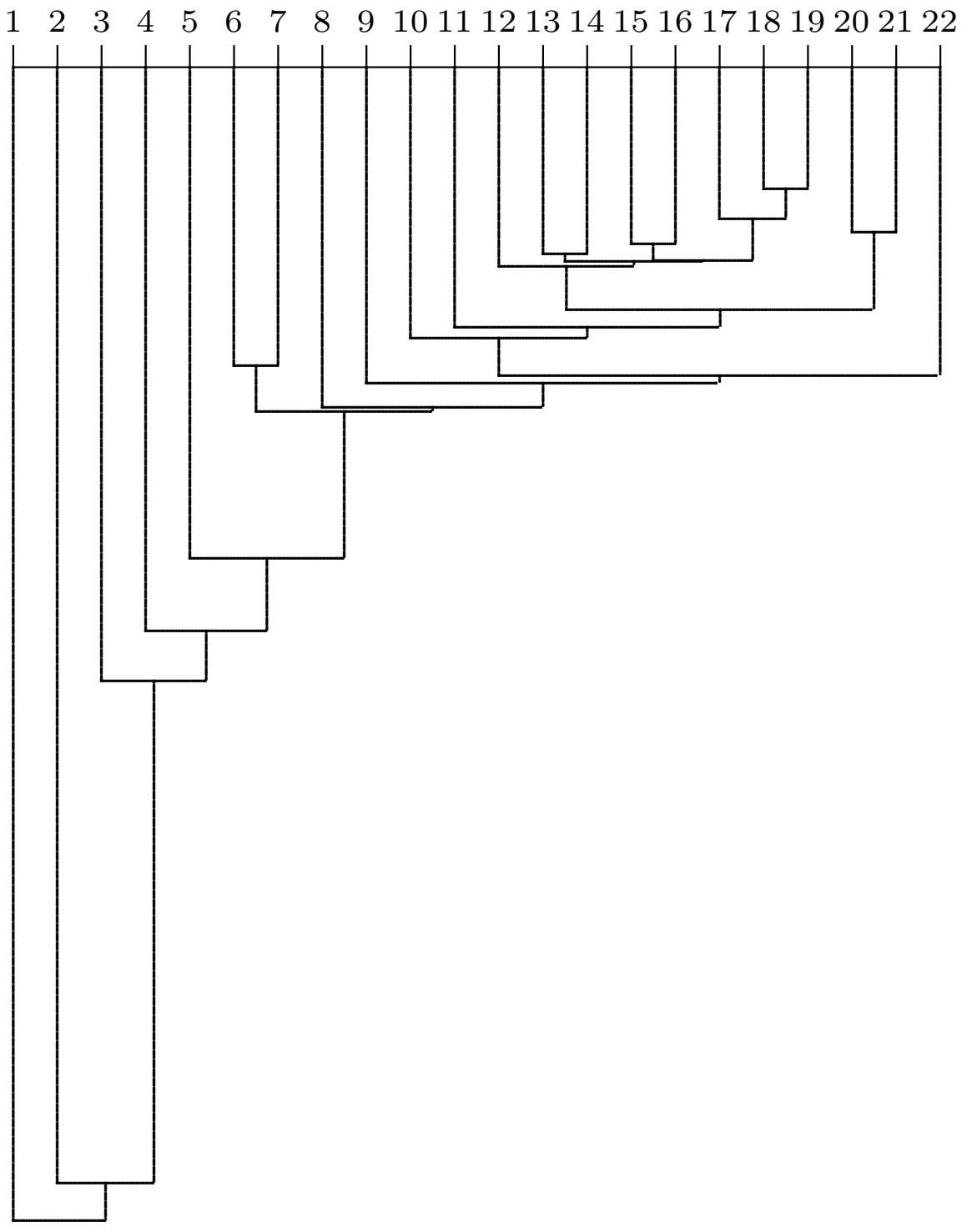
$$d_{q't} = \min_{t\neq q,r} \{d_{qt}, d_{rt}\}$$

Metoda najdalszego sąsiedztwa

$$d_{q't} = \max_{t\neq q,r} \{d_{qt}, d_{rt}\}$$

Przykład. Badano 22 samochody różnych marek pod względem czterech cech: *ceny* (X_1), *przyspieszenia* (X_2), *hamowania* (X_3), *trzymania się drogi* (X_4) oraz *zużycia paliwa* (X_5).

	X_1	X_2	X_3	X_4	X_5
Acura	−0.5211	0.4773	−0.0066	0.3816	2.0788
Audi	0.8657	0.2080	0.3187	−0.0914	−0.6771
BMW	0.4959	−0.8015	0.1922	−0.0914	−0.1538
Buick	−0.6135	1.6887	0.9331	−0.2096	−0.1538
Corvette	1.2354	−1.8111	−0.4945	0.9729	−0.6771
Chrysler	−0.6135	0.0734	0.4271	−0.2096	−0.1538
Dodge	−0.7060	−0.1958	0.4813	0.1451	−0.1538
Eagle	−0.6135	1.2176	−4.1989	−0.2096	−0.6771
Ford	−0.7060	−1.5419	0.9873	0.1451	−1.7236
Honda	−0.4286	0.4099	−0.0066	0.0269	0.3695
Isuzu	−0.7984	0.4099	−0.0608	−4.2301	1.0671
Mazda	0.1261	0.6792	−0.1331	0.4999	−1.7236
Mercedes	1.0505	0.0061	0.1199	−0.0914	−0.1538
Mitsub.	−0.6135	−1.0035	0.0838	0.3816	0.7183
Nissan	−0.4286	0.0734	−0.0066	0.2634	0.9974
Olds	−0.6135	−0.7342	0.4090	0.3816	2.1136
Pontiac	−0.6135	0.6792	0.5355	0.1451	0.1950
Porsche	3.4542	−2.2149	−0.2957	0.6181	−1.0259
Saab	0.5883	0.6792	0.2464	0.2634	0.0206
Toyota	−0.0588	1.2176	0.2283	0.7364	−0.8515
VW	−0.7060	−0.1285	0.1019	0.3816	0.1950
Volvo	0.2185	0.6119	0.1380	−0.2096	0.3695


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22

Metoda k–średnich

$\mathbf{X}_1, \ldots, \mathbf{X}_n$ — p–wymiarowe obserwacje jednostek

Założenie

Przyjmujemy, że obserwacje $\mathbf{X}_1, \ldots, \mathbf{X}_n$ pochodzą z k populacji.

$\mathcal{J} = \{I_1, \ldots, I_k\}$: podział zbioru $\{1, \ldots, n\}$ na rozłączne podzbiory

$$\bar{\mathbf{X}}_j = \frac{1}{n_j} \sum_{i \in I_j} \mathbf{X}_i$$

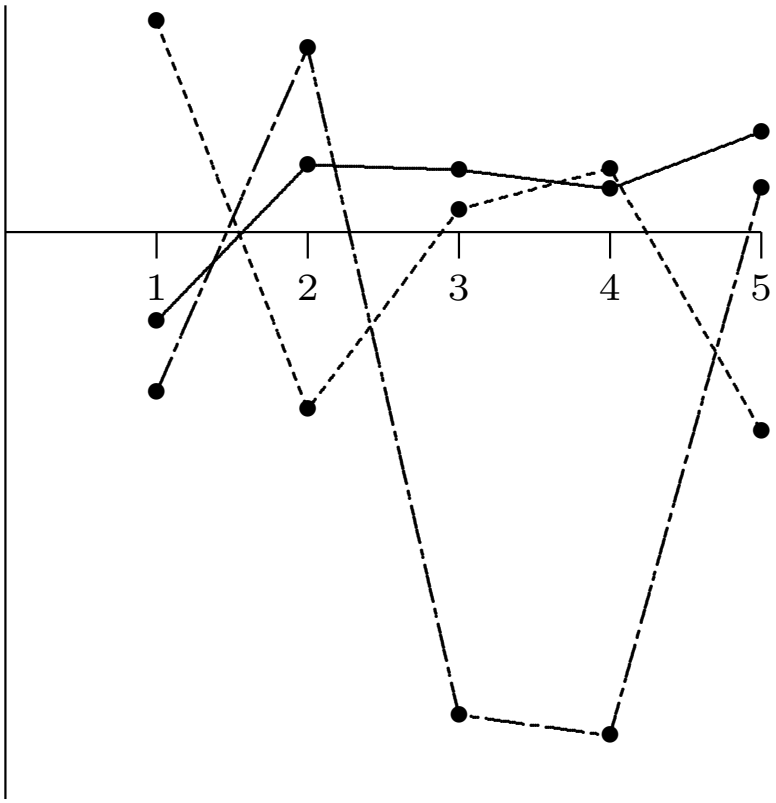
$$D(\mathcal{J}) = \sum_{i=1}^{k} \sum_{i \in I_j} (\mathbf{X}_i - \bar{\mathbf{X}}_j)^2$$

Znaleźć takie $\mathcal{J}^*$, że

$$D(\mathcal{J}^*) = \min D(\mathcal{J})$$

Przykład. (cd.)

Cecha	Średnie		
	1	2	3
X_1	−0.39307	0.93169	−0.70597
X_2	0.29605	−0.78231	0.81378
X_3	0.27422	0.09927	−2.12984
X_4	0.19061	0.28027	−2.21984
X_5	0.44191	−0.87640	0.19503



Grupa 1:
Acura Buick Chrysler Dodge Honda Mitsub. Nissan Olds Pontiac Saab Toyota VW Volvo

Grupa 2:
Audi BMW Corvette Ford Mazda Mercedes Porsche

Grupa 3:
Eagle Isuzu