

# Analiza Wielowymiarowa

## Analiza dyskryminacyjna

Maciej Nasiński, Paweł Strawiński

Zajęcia 7  
25 listopada 2021

- 1 Wprowadzenie
  - O analizie
  - Metoda
  - Założenia i metody
- 2 Liniowa Analiza Dyskryminacyjna
- 3 Kwadratowa Analiza Dyskryminacyjna
- 4 Inne metody

# Metody grupowania

- Jednym z działów Analizy Wielowymiarowej są metody grupowania
- Służą one do badania podobieństwa obiektów pod względem wartości charakterystyk grupujących oraz cech tych obiektów
- Metody grupowania dzielą się na metody dyskryminacyjne i metody klasyfikacyjne
- Dyskryminacją nazywany jest przydział obiektów do znanych wcześniej klas
- Klasyfikacją nazywany jest podział obiektów na nieznane wcześniej klasy

# Analiza dyskryminacyjna

- Analiza dyskryminacyjna jest stosowana do rozstrzygania, które zmienne wyróżniają (dyskryminują) dwie lub większą liczbę naturalnie wyłaniających się grup
- Szuka reguły przyporządkowania wielowymiarowych obiektów do jednej z wielu klas przy możliwie minimalnych błędach klasyfikacji
- Główna idea to rozstrzygnięcie, czy grupy różnią się ze względu na średnią wartość pewnej cechy (zmiennej), a następnie wykorzystanie tej zmiennej do przewidywania przynależności do grupy

# Analiza dyskryminacyjna

- Liniowa Analiza Dyskryminacyjna została niezależnie i równocześnie zaproponowana przez dwóch badaczy, R. A. Fishera (1936) i P. C. Mahalanobisa (1936) jako sposób na rozwiązanie dwóch problemów
- Fisher zaproponował podejście opisowe, Opisowa Analiza Dyskryminacyjna, które pozwala na znalezienie takiej kombinacji liniowej cech, która w optymalny sposób rozdzieli grupy obiektów
- Z kolei Mahalanobis zaproponował Predykcyjną Analizę Dyskryminacyjną. Pozwala ona na przypisanie obiektu do grupy obiektów podobnych

# Cel analizy

- Celem opisowej analizy dyskryminacyjnej jest opisanie różnic między grupami obiektów
- Celem predykcyjnej analizy dyskryminacyjnej jest klasyfikowanie obserwacji, o których nie wiadomo, do której grupy należą
- Może być traktowana jako rozszerzenie analizy wariancji
- Jeśli występują istotne statystycznie różnice w średnich wartościach cech obiektów pomiędzy grupami, to mogą być one wykorzystane do przewidywania przynależności do grupy

# Założenia analizy dyskryminacyjnej

- Rozkład cech obiektów jest rozkładem wielowymiarowym normalnym
- Wariancje cechy w grupach są zbliżone (homogeniczne)
- Analiza dyskryminacyjna jest odporna wobec skośności rozkładów, o ile pozostaje co najmniej 20 stopni swobody w każdej grupie
- Obserwacje odstające zaburzają wyniki analizy
- Zalecenie praktyczne: najmniej liczna grupa powinna mieć kilkakrotnie (4-5) razy większą liczbę obserwacji niż jest zmiennych dyskryminujących

## Opis metody

- Zmienna objaśniana, wskazująca na przynależność do klas (grup), przyjmuje dwie albo większą (skończoną) liczbę wartości
- Poszukiwana jest taka kombinacja liniowa cech obiektów, która w optymalny sposób przypisze je do klas
- Procedura minimalizuje wariancję cech obiektu wewnątrz klasy i maksymalizuje wariancję pomiędzy klasami
- W tym celu poszukiwany jest kierunek wektora  $a$



## Wzory analityczne dla dwóch grup

- Na podstawie danych wyznaczane są:
  - średnie grupowe

$$\bar{x}_k = \sum_{i=1}^{n_k} x_{ki}$$

- macierze wariancji wewnątrzgrupowej

$$\mathbf{W} = \frac{1}{n-2} \sum_{k=1}^K (n_k - 1) S_k = \frac{1}{n-2} \sum_{k=1}^K \sum_{i=1}^{n_k} (x_{ki} - \bar{x}_k)^2$$

## Wzory analityczne dla dwóch grup

- Fisher zadanie analizy dyskryminacyjnej zdefiniował następująco:

*znaleźć taki kierunek  $a$ , który maksymalizuje odległość między zrzuconymi średnimi obu prób przy uwzględnieniu wariancji rzutu (odległości średnich)*

- Poszukiwane jest zatem rozwiązanie problemu

$$\arg \max_a \frac{(a' \bar{x}_1 - a' \bar{x}_2)^2}{a' \mathbf{W} a}$$

- Rozwiązaniem jest

$$a = \mathbf{W}^{-1}(\bar{x}_1 - \bar{x}_2)$$

# Reguła dyskryminacyjna

- W celu wyznaczenia reguły dyskryminacyjnej
  - Obserwacje są rzutowane na hiperpłaszczyznę o kierunku  $a$
  - Średnie grupowe  $\bar{x}_1$  oraz  $\bar{x}_2$  są rzutowane na hiperpłaszczyznę o kierunku  $a$
  - Obserwacje są przypisywane do tej grupy, której rzut środka jest bliższy
- Reguła dyskryminacyjna Fishera. Obiekt przypisywany jest do grupy 1, gdy

$$(\bar{x}_1 - \bar{x}_2)\mathbf{W}^{-1}(x - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)) > 0$$

- W przeciwnym przypadku obiekt przypisywany jest do grupy 2
- W przypadku równości, przydział dokonywany jest metodą ekspercką

## Wzory dla więcej niż dwóch grup

- Poszukiwane jest rozwiązanie problemu dla  $g$  grup

$$\arg \max_a \frac{a' \mathbf{B} a}{a' \mathbf{W} a}$$

- gdzie
  - wariancja międzygrupowa

$$\mathbf{B} = \frac{1}{g-1} \sum_{k=1}^g n_k (\bar{x}_k - \bar{x})^2$$

- wariancja wewnątrzgrupowa

$$\mathbf{W} = \frac{1}{n-g} \sum_{k=1}^g (n_k - 1) S_k = \frac{1}{n-2} \sum_{k=1}^K \sum_{i=1}^{n_k} (x_{ki} - \bar{x}_k)^2$$

# Terminologia

- Kierunek  $a$  nazywany jest kierunkiem kanonicznym, jest to kierunek najlepiej rozdzielający klasy
- Wektor o kierunku  $a$  nazywany jest pierwszym wektorem kanonicznym
- Zmienna  $a'x$  jest pierwszą zmienną kanoniczną odpowiadającą wektorowi  $x$ .

# Kanoniczna LAD

- Przy znanych postaciach macierzy  $\mathbf{W}$  oraz  $\mathbf{B}$  konstruowana jest macierz  $\mathbf{W}^{-1}\mathbf{B}$
- Kolejne wektory własne macierzy  $\mathbf{W}^{-1}\mathbf{B}$  odpowiadające uporządkowanym malejąco wartościom własnym nazywane są wektorami kanonicznymi
- Jest to połączenie analizy kanonicznej i analizy dyskryminacyjnej
- Jest ono użyteczne w przypadku, gdy badacz dysponuje dużą liczbą potencjalnych zmiennych dyskryminujących

# Interpretacja

- Współczynniki funkcji dyskryminacyjnej nie posiadają interpretacji ilościowej
- Ilościowo można interpretować standaryzowane współczynniki funkcji dyskryminacyjnej. Wyrażone są w jednostkach odchylenia standardowego. Wskazują, jak silny jest wpływ danej zmiennej dyskryminacyjnej na różnicowanie grup.

# Kwadratowa Analiza Dyskryminacyjna

- Kwadratowa Analiza Dyskryminacyjna została zaproponowana w artykule Smith (1947).
- Jest uogólnieniem Liniowej Analizy Dyskryminacyjnej
- Jest bardziej ogólna od LAD, gdyż nie wymaga spełnienia założenia o równości macierzy wariancji-kowariancji w grupach
- Ale wymaga oszacowania dwóch macierzy wariancji-kowariancji
- Oraz, gdy najmniej liczebna grupa liczy mniej obserwacji niż jest zmiennych dyskryminujących, nie można przeprowadzić analizy



## Metoda najbliższych sąsiadów

- Metoda najbliższych sąsiadów jest techniką wykorzystywaną głównie w predykcyjnej analizie danych
- Jest metodą nieparametryczną, dzięki czemu zbędne są założenia dotyczące rozkładu danych
- Obserwacja przypisywana jest do grupy na podstawie analizy jej najbliższych sąsiadów
- Jej zaletą jest możliwość wyróżniania grup o nieregularnych kształtach
- Z drugiej strony, wyniki zależą od wyboru liczby sąsiadów i metryki stosowanej do określenia odległości

# Najczęściej wykorzystywane metryki

- $L_1$ , czyli metryka miejska
- $L_2$ , czyli metryka Euklidesowa
- $L_\infty$ , czyli największa odległość
- współczynnik korelacji
- odległość Mahalanobisa

# Logistyczna analiza dyskryminacyjna

- Jest częściowo parametryczną metodą, znajdującą się pomiędzy liniową analizą dyskryminacyjną a metodą najbliższych sąsiadów
- Wykorzystuje model dla dyskretnej zmiennej zależnej szacujący prawdopodobieństwo sukcesu (model logistyczny) w celu przeprowadzenia analizy dyskryminacyjnej
- Rozwiązanie polega na maksymalizacji funkcji wiarygodności