10 Metody wielowymiarowe

Analiza skupień

Analiza skupień jest narzędziem analizy danych służącym do grupowania jednostek eksperymentalnych (lub/i zmiennych), opisanych za pomocą wektora obserwacji, w K niepustych, rozłącznych i możliwie "jednorodnych" grup - skupień. Jednostki (zmienne) należące do danego skupienia powinny być "podobne" od siebie (używa się w tym celu różnych miar podobieństwa, a w zasadzie niepodobieństwa), a jednostki (zmienne) należące do różnych skupień powinny być z kolei możliwie mocno "niepodobne" do siebie. Głównym celem tej analizy jest wykrycie z zbiorze danych, tzw. "naturalnych" skupień, czyli skupień, które dają się w sensowny sposób interpretować.

Algorytmy analizy skupień:

1. hierarchiczne (np. metoda aglomeracyjna),
2. niehierarchiczne (np. metoda k-średnich).

Metoda aglomeracyjna

Podstawowe miary niepodobieństwa jednostek (zmiennych):

1. odległość euklidesowa - stosowana w przypadku grupowania jednostek opisanych zmiennymi ilościowymi,
2. niezgodność procentowa - stosowana w przypadku grupowania jednostek opisanych zmiennymi jakościowymi,
3. $1 - r$ Pearsona - stosowana w przypadku grupowania zmiennych.

Podstawowe sposoby wiązania skupień:

1. metoda pojedynczego wiązania (najbliższego sąsiedztwa) - miara niepodobieństwa pomiędzy dwoma skupieniami jest określona jako najmniejsza miara niepodobieństwa między dwoma jednostkami (zmiennymi) należącymi do różnych skupień,
2. metoda pełnego wiązania (najdalszego sąsiedztwa) - miara niepodobieństwa pomiędzy dwoma skupieniami jest określona jako największa miara niepodobieństwa między dwoma jednostkami (zmiennymi) należącymi do różnych skupień,
3. metoda średniego wiązania - miara niepodobieństwa pomiędzy dwoma skupieniami jest określona jako średnia miara niepodobieństwa między wszystkimi parami jednostek (zmiennych) należących do różnych skupień.

Algorytm aglomeracyjny

1. W pierwszym kroku każda z jednostek (zmiennych) tworzy oddzielne skupienie.
2. Łączymy (wiążemy ze sobą) dwa najbardziej podobne do siebie skupienia - w sensie wybranej miary niepodobieństwa skupień - zmniejszając w ten sposób liczbę skupień o jeden.
3. Powtarzamy krok drugi do momentu uzyskania zadeklarowanej, końcowej liczby skupień K lub do połączenia wszystkich jednostek (zmiennych) w jedno skupienie.

Uwaga: Graficzną ilustracją przebiegu aglomeracji jest wykres zwany **dendrogramem**.

Metoda k-średnich

Główną ideą metody K-średnich jest taka alokacja jednostek, która minimalizuje zmienność wewnątrz powstałych skupień, a co za tym idzie maksymalizuje zmienność pomiędzy skupieniami.

Niech C_K oznacza funkcję, która każdej jednostce (dokładnie jego numerowi), przyporządkowuje numer skupienia do którego jest ona przyporządkowana (przy podziale na K skupień).

Dla ustalonej funkcji C_K, oznaczmy macierze zmienności (analogicznie jak w ANOVA):

- $SSE(C_K)$ - macierz zmienności wewnątrz skupień,
- $SSA(C_K)$ - macierz zmienności pomiędzy skupieniami,
- SST - macierz zmienności całkowitej, przy czym

$$SST = SSE(C_K) + SSA(C_K).$$

Powszechnie stosowane algorytmy metody K-średnich minimalizują ślad macierzy $SSE(C_K)$.

Oznaczmy przez C_K^* funkcję realizującą optymalny podział jednostek na K skupień.

Wtedy

$$C_K^* = \min_{C_K} \operatorname{tr}[SSE(C_K)].$$

Algorytm K-średnich

1. Rozmieszczamy n jednostek w K skupieniach. Niech funkcja $C_K^{(1)}$ opisuje to rozmieszczenie.
2. Dla każdego z K skupień obliczamy wektory średnich $\bar{\boldsymbol{x}}_k$, $(k = 1, 2, \ldots, K)$.
3. Rozmieszczamy ponownie jednostki w K skupieniach, w taki sposób że

$$C_K^{(l)}(i) = \arg \min_{1 \leq k \leq K} \mathrm{tr}[SSE(C_K)].$$

4. Powtarzamy kroki drugi i trzeci aż do momentu, gdy przyporządkowanie jednostek do skupień pozostanie niezmienione, tzn. aż do momentu, gdy $C_K^{(l)} = C_K^{(l-1)}$.

Istnieje wiele modyfikacji powyższego algorytmu. Przykładowo, losowe rozmieszczenie elementów w skupieniach - krok pierwszy algorytmu, zastąpione zostaje narzuconym podziałem, mającym na celu szybsze ustabilizowanie się algorytmu. Inna modyfikacja polega na przeliczeniu średnich skupień natychmiast po przesunięciu pomiędzy skupieniami choćby jednego elementu.

Wszystkie wersje algorytmu K-średnich są zbieżne. Nie gwarantują one jednak zbieżności do optymalnego rozwiązania C_K^*. Niestety, w zależności od początkowego podziału, algorytm zbiega do zazwyczaj różnych lokalnie optymalnych rozwiązań. W związku z tym, aby uzyskać najlepszy podział, zaleca się często wielokrotne stosowanie tego algorytmu z różnymi, wstępnymi rozmieszczeniami jednostek.

Funkcje związane z analizą skupień:

dist - obliczanie macierzy odległości,

hclust - metoda aglomeracyjna, procedura główna,

plclust - dendrogram,

cutree - podział na skupienia,

kmeans - metoda k-średnich, procedura główna.

Klasyfikacja

W zagadnieniach klasyfikacyjnych poszukujemy reguły **(klasyfikatora)** (zazwyczaj oznaczanego literą d) pozwalającego na przyporządkowanie obiektu (opisanego przez p-wymiarowy wektor obserwowanych cech $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)'$) do jednej z K $(K \geq 2)$ klas.

$$d : \qquad \rightarrow \{1, 2, \ldots, K\}.$$

Uwaga: Liczby $1, 2, \ldots, K$ oznaczają **etykiety** klas.

Oznaczenia:

$\pi_1 = P(Y = 1), \ldots, \pi_K = P(Y = K)$ - prawdopodobieństwa **a priori**,

$p_1(\boldsymbol{x}) = P(Y = 1 | \boldsymbol{X} = \boldsymbol{x}), \ldots, p_K(\boldsymbol{x}) = P(Y = K | \boldsymbol{X} = \boldsymbol{x})$ - prawdopodobieństwa **a posteriori**.

Wzór Bayesa

$$p_k(\boldsymbol{x}) = \frac{\pi_k f_k(\boldsymbol{x})}{\sum_{i=1}^{K} \pi_i f_i(\boldsymbol{x})},\ k = 1, \ldots, K,$$

gdzie $f_k(\boldsymbol{x})$ oznacza gęstość rozkładu wektora $\boldsymbol{X}$ w k-tej klasie.

Bayesowska reguła klasyfikacyjna

$$d_B(\boldsymbol{x}) = \arg\max_k p_k(\boldsymbol{x}) = \arg\max_k \pi_k f_k(\boldsymbol{x}),$$

gdzie $\arg\max_k$ oznacza tę wartość k, która maksymalizuje dane wyrażenie.

Jeżeli $f_k(\boldsymbol{x})$ jest gęstością p-wymiarowego rozkładu normalnego $N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, to klasyfikator bayesowski ma postać

$$d_B(\boldsymbol{x}) = \arg\max_k \delta_k(\boldsymbol{x}),$$

gdzie

$$\delta_k(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln \pi_k.$$

Parametry π_k, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}$, $k = 1, 2, \ldots, K$, występujące w funkcji klasyfikującej $\delta_k(\boldsymbol{x})$ nie są znane i należy zastąpić je ich estymatorami.

$$\hat{\delta}_k(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{S}^{-1}\bar{\boldsymbol{Y}}_k - \frac{1}{2}\bar{\boldsymbol{Y}}_k'\boldsymbol{S}^{-1}\bar{\boldsymbol{Y}}_k + \ln \frac{n_k}{n},$$

gdzie n_k oznacza liczbę obserwacji z k-tej grupy.

Prognoza

Dokonujemy prognozy dla nowego obiektu opisanego wektorem p cech $\boldsymbol{x}_p$. Prognozowaną wartość etykiety Y równą y_p wyznaczamy ze wzoru

$$y_p = \arg\max_k \hat{\delta}_k(\boldsymbol{x}_p).$$

Szacowanie jakości klasyfikatora

1. Metoda ponownego podstawienia. Korzystając z całej próby (uczącej) uzyskujemy klasyfikator $\hat{d}$. Następnie, wykorzystując ten klasyfikator, prognmozujemy etykiety wszystkich elementów próby (obiektów). Miarą jakości jest proporcja poprawnie zaklasyfikowanych obiektów.

2. Metoda sprawdzania krzyżowego. W losowy sposób dzielimy całą próbę (uczącą) na v podzbiorów (zazwyczaj $v = 3, 5$ lub 10). $v - 1$ z nich tworzy nową próbę (uczącą) na bazie której uzyskujemy klasyfikator $\hat{d}$. Dla pozostałych elementów próby dokonujemy prognozy (z wykorzystaniem klasyfikatora $\hat{d}$. Procedure tę powtarzamy kolejno v-razy. Miarą jakości jest proporcja poprawnie zaklasyfikowanych obiektów.

Funkcje związane z klasyfikacją:

lda (MASS) - metoda LDA, procedura główna,

predict - klasyfikacja nowych obiektów.