



Elementy statystyki

STA - Wykład 7

dr hab. Waldemar Wołyński
Wydział Matematyki i Informatyki
Uniwersytet im. Adama Mickiewicza



Składowe główne



Zakładamy, że każda jednostka (obiekt) opisany jest za pomocą p skorelowanych zmiennych (cech) X_1, X_2, \dots, X_p . Ponadto zakładamy, że wektor (dokładnie: wektor losowy) $\mathbf{X} = (X_1, X_2, \dots, X_p)'$ ma zerową wartość oczekiwaną

$$E(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}$$

oraz dodatnio określoną macierz kowariancji

$$\text{Var}(\mathbf{X}) = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_p) \\ \text{Cov}(X_1, X_2) & \text{Var}(X_2) & \dots & \text{Cov}(X_2, X_p) \\ \vdots & \vdots & & \vdots \\ \text{Cov}(X_1, X_p) & \text{Cov}(X_2, X_p) & \dots & \text{Var}(X_p) \end{bmatrix} = \Sigma.$$

Zmienne X_1, X_2, \dots, X_p nazywamy **zmiennymi pierwotnymi**. Całkowitą zmienność wektora \mathbf{X} opisuje wielkość

$$\sum_{i=1}^p \text{Var}(X_i).$$

Składowe główne

Poszukujemy nowych zmiennych Z_1, Z_2, \dots, Z_p , **składowych głównych**, opisujących daną jednostkę (obiekt) o następujących własnościach:

1. Każda nowa zmienna (składowa główna) jest liniową kombinacją zmiennych pierwotnych, tzn.

$$Z_j = \sum_{i=1}^p a_{ij} X_i, \quad j = 1, 2, \dots, p.$$

2. Składowe główne są nieskorelowane.
3. Pierwsza składowa główna ma największą wariancję spośród wszystkich liniowych kombinacji zmiennych pierwotnych, druga składowa główna ma największą wariancję spośród wszystkich liniowych kombinacji zmiennych pierwotnych nieskorelowanych z pierwszą składową główną, itd.
4. $\sum_{i=1}^p \text{Var}(X_i) = \sum_{i=1}^p \text{Var}(Z_i)$.





Fakt

Wektor $\mathbf{a}_j = (a_{1j}, a_{2j}, \dots, a_{pj})'$ jest wektorem charakterystycznym, odpowiadającym j -tej co do wielkości, wartości własnej λ_j macierzy Σ .

Ponadto, $\lambda_j = \text{Var}(Z_j)$, $j = 1, 2, \dots, p$.

Uwaga: Ponieważ macierz Σ nie jest znana, posługujemy się jej oszacowaniem z próby.

Fakt

Niech $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ będzie próbą prostą z populacji o p -wymiarowym rozkładzie z zerowym wektorem wartości oczekiwanych i dodatnio określonej macierzy kowariancji Σ . Wtedy statystyka

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'$$

jest nieobciążonym estymatorem macierzy kowariancji Σ .



W analizie składowych głównych oczekujemy, że dla pewnego małego k , suma $\lambda_1 + \lambda_2 + \dots + \lambda_k$ będzie bliska $\lambda_1 + \lambda_2 + \dots + \lambda_p$. Jeśli tak jest, to k pierwszych składowych głównych wyjaśnia dobrze zmienność wektora $\mathbf{X} = (X_1, X_2, \dots, X_p)'$ i pozostałe $p - k$ składowe główne wnoszą niewiele, ponieważ mają one małe wariancje.

Wskaźnik

$$\frac{\lambda_1 + \dots + \lambda_k}{\lambda_1 + \dots + \lambda_p} 100\%$$

jest procentową miarą wyjaśniania zmienności wektora \mathbf{X} przez pierwszych k składowych głównych.

Dobór liczby składowych głównych

Popularne metody ustalenia liczby użytecznych składowych głównych.

1. Jeśli dla pewnego k wskaźnik

$$\frac{\lambda_1 + \dots + \lambda_k}{\lambda_1 + \dots + \lambda_p} 100\% \geq \beta,$$

np. $\beta = 80\%$, to pozostałe $p - k$ składowe główne pomijamy.

2. Pomijamy te składowe główne, których wartości własne są mniejsze od średniej

$$\bar{\lambda} = \frac{1}{p} \sum_{j=1}^p \lambda_j.$$

Uwaga: W ustaleniu liczby użytecznych składowych głównych, pomocny jest również **wykres osypiska**.





Wartość modułu współczynnika a_{ij} , w j -tej składowej głównej, pokazuje wkład w jej budowę i -tej zmiennej pierwotnej (z uwzględnieniem udziału pozostałych zmiennych pierwotnych).

R: princomp – analiza składowych głównych.



Analiza skupień

Zakładamy, że każda jednostka (obiekt) opisany jest za pomocą p skorelowanych zmiennych (cech) X_1, X_2, \dots, X_p . Ponadto zakładamy, że dysponujemy wartościami tych cech, dla n obiektów:

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})', \quad i = 1, 2, \dots, n.$$

Obiekty te grupujemy w K niepustych, rozłącznych i możliwie "jednorodnych" skupień. Obiekty należące do danego skupienia powinny być "podobne" od siebie (używa się w tym celu różnych miar podobieństwa, a w zasadzie niepodobieństwa obiektów), a obiekty należące do różnych skupień powinny być z kolei możliwie mocno "niepodobne" do siebie.

Metody analizy skupień:

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





1. Odległość euklidesowa:

$$\rho_1(\mathbf{x}_r, \mathbf{x}_s) = ((\mathbf{x}_r - \mathbf{x}_s)'(\mathbf{x}_r - \mathbf{x}_s))^{1/2} = \left(\sum_{i=1}^p (x_{ri} - x_{si})^2 \right)^{1/2}$$

2. Odległość miejska:

$$\rho_3(\mathbf{x}_r, \mathbf{x}_s) = \sum_{i=1}^p |x_{ri} - x_{si}|,$$

3. Odległość Czebyszewa:

$$\rho_4(\mathbf{x}_r, \mathbf{x}_s) = \max_{1 \leq i \leq p} |x_{ri} - x_{si}|.$$

Metoda aglomeracyjna - sposoby wiązania skupień

1. **Metoda pojedynczego wiązania (najbliższego sąsiedztwa).** Miara niepodobieństwa pomiędzy dwoma skupieniami jest określona jako najmniejsza miara niepodobieństwa między dwoma obiektami należącymi do różnych skupień, tzn. $\rho(R, S) = \min_{i \in R, j \in S} \rho(\mathbf{x}_i, \mathbf{x}_j)$.
2. **Metoda pełnego wiązania (najdalszego sąsiedztwa).** Miara niepodobieństwa pomiędzy dwoma skupieniami jest określona jako największa miara niepodobieństwa między dwoma obiektami należącymi do różnych skupień, tzn. $\rho(R, S) = \max_{i \in R, j \in S} \rho(\mathbf{x}_i, \mathbf{x}_j)$.
3. **Metoda średniego wiązania.** Miara niepodobieństwa pomiędzy dwoma skupieniami jest określona jako średnia miara niepodobieństwa między wszystkimi parami obiektów należących do różnych skupień, tzn.

$$\rho(R, S) = \frac{1}{n_R n_S} \sum_{i \in R} \sum_{j \in S} \rho(\mathbf{x}_i, \mathbf{x}_j),$$

gdzie n_R i n_S są liczbami obiektów wchodzących w skład skupień R i S odpowiednio.





Algorytm aglomeracyjny

1. W pierwszym kroku każdy z obiektów tworzy oddzielne skupienie. Zatem skupień tych jest n .
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 połączenia wszystkich obiektów w jedno skupienie.



Graficzną ilustracją przebiegu aglomeracji jest wykres zwany **dendrogramem**. Jest to (binarne) drzewo którego węzły reprezentują skupienia, a liście pojedyncze obiekty. Liście umieszczone są na poziomie zerowym, pozostałe węzły drzewa umieszczone są na wysokości odpowiadającej mierze niepodobieństwa pomiędzy skupieniami reprezentowanymi przez węzły potomki.

Funkcje w **R** związane z aglomeracyjną analizą skupień:

dist – obliczanie macierzy odległości,

hclust – procedura główna,

cutree – podział na skupienia.



Klasyfikacja



Zakładamy, że badaną populację podzielić można na K , ($K \geq 2$) rozłącznych i niepustych klas. Każda jednostka (obiekt) opisany jest za pomocą p skorelowanych zmiennych (cech) X_1, X_2, \dots, X_p . Ponadto, niech Y oznacza dodatkową zmienną o wartości równej etykietcie (numerowi) klasy danego obiektu.

Niech $\mathbf{x}_0 = (x_{01}, x_{02}, \dots, x_{0p})'$ będą wartościami cech uzyskanymi dla pewnego obiektu należącego do tej populacji.

Poszukujemy reguły (**klasyfikatora**) pozwalającego na przyporządkowanie obiektu opisanego za pomocą wektora obserwacji \mathbf{x}_0 do konkretnej klasy.

Metoda bayesowska

Klasyfikator bayesowski dokonuje prognozy etykiety Y dla obserwacji $\mathbf{X} = \mathbf{x}_0$ w następujący sposób:

$$\hat{y}_0 = \arg \max_{1 \leq k \leq K} \hat{p}_k(\mathbf{x}_0),$$

gdzie $\hat{p}_k(\mathbf{x}_0)$ jest estymatorem prawdopodobieństwa a posteriori przynależności obserwacji \mathbf{x}_0 do k -tej klasy, tzn.

$$p_k(\mathbf{x}_0) = P(Y = k | \mathbf{X} = \mathbf{x}_0).$$

Fakt

$$p_k(\mathbf{x}_0) = \frac{\pi_k f_k(\mathbf{x}_0)}{\sum_{i=1}^K \pi_i f_i(\mathbf{x}_0)}, \quad k = 1, 2, \dots, K,$$

gdzie

π_k oznacza prawdopodobieństwo a priori przynależności klasyfikowanego obiektu do klasy K ,

f_k oznacza gęstość rozkładu wektora \mathbf{X} dla klasy K .





Estymacja prawdopodobieństw a posteriori

Zakładamy, że w każdej z K klas dysponujemy n_k elementowymi próbami prostymi (**próba ucząca**).

Ponadto zakładamy, że próby te pochodzą z p -wymiarowego rozkładu normalnego z nieznanymi wartościami oczekiwanymi $\mu_1, \mu_2, \dots, \mu_K$ oraz nieznaną wspólną macierzą kowariancji Σ .

Fakt

Statystyki

$$\hat{\mu}_k = \bar{\mathbf{x}}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{x}_{ki}, \quad k = 1, 2, \dots, K$$

są nieobciążonymi estymatorami wartości oczekiwanych $\mu_1, \mu_2, \dots, \mu_K$. Ponadto, statystyka

$$\hat{\Sigma} = \mathbf{S} = \frac{1}{n - K} \sum_{k=1}^K \sum_{i=1}^{n_k} (\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)(\mathbf{x}_{ki} - \bar{\mathbf{x}}_k)', \quad n = n_1 + n_2 + \dots + n_K$$

jest nieobciążonym estymatorem macierzy kowariancji Σ .



Przyjmując:

$$\hat{\pi}_k = \frac{n_k}{n}, \quad n = n_1 + n_2 + \cdots + n_K,$$

$$\hat{f}_k(\mathbf{x}_0) = (2\pi)^{-p/2} |\mathbf{S}|^{-p/2} \exp \left[-\frac{1}{2} (\mathbf{x}_0 - \bar{\mathbf{x}}_k)' \mathbf{S}^{-1} (\mathbf{x}_0 - \bar{\mathbf{x}}_k) \right],$$

otrzymujemy następującą prognozę etykiety:

$$\hat{y}_0 = \arg \max_{1 \leq k \leq K} \delta_k(\mathbf{x}_0),$$

gdzie

$$\delta_k(\mathbf{x}_0) = \mathbf{x}_0' \mathbf{S}^{-1} \bar{\mathbf{x}}_k - \frac{1}{2} \bar{\mathbf{x}}_k' \mathbf{S}^{-1} \bar{\mathbf{x}}_k + \ln \hat{\pi}_k.$$

Powyższa funkcja nosi nazwę **liniowej funkcji klasyfikującej związanej z k -tą klasą**.



Funkcje związane z klasyfikacją metodą LDA:

lda (MASS) – procedura główna,

predict – prognoza etykiety oraz estymacja prawdopodobieństw a posteriori.