Statystyka wielowymiarowa

Statystyka wielowymiarowa

- Model
- Redukcja wymiaru
- Analiza zależności cech

Model jednowymiarowy

- Niech X będzie badaną cechą populacji

$$X_i = \mu + \varepsilon_i, \qquad i = 1,2,\dots,n$$

  gdzie
  μ – wartość oczekiwana badanej cechy
  ε_i – reszty (błędy), niezależne zmienne losowe o jednakowym rozkładzie z zerową wartością oczekiwaną i wariancją równą σ^2

- **Uwaga:**
  Często przyjmujemy dodatkowo, że X_i mają rozkłady normalne o gęstości

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad x \in R$$

Model wielowymiarowy

- Niech $X_1, X_2, \dots, X_p$ będą badanymi cechami populacji

$$X_i = \mu + \varepsilon_i, \qquad i = 1,2 \dots, n$$

gdzie
$X_i = (X_{i1}, X_{i2}, \dots, X_{ip})'$ - wektor badanych cech populacji
μ – wektor wartości oczekiwanych

$$\mu = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$$

ε_i - reszty (błędy), niezależne wektory losowe o jednakowym rozkładzie z zerowym wektorem wartości oczekiwanych i dodatnio określoną macierzą kowariancji Σ

Model wielowymiarowy

$$\Sigma = \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & \cdots & Cov(X_1, X_p) \\ Cov(X_1, X_2) & Var(X_2) & \cdots & Cov(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_1, X_p) & Cov(X_2, X_P) & \cdots & Var(X_p) \end{bmatrix}$$

Uwaga:

Często przyjmujemy dodatkowo, że wektory X_i mają p-wymiarowe rozkłady normalne o gęstości

$$f(x) = \frac{1}{\sqrt{|\Sigma|(2\pi)^p}} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}, \qquad x \in R^p$$

Model wielowymiarowy

Estymatorami nieobciążonymi (a przy dodatkowym założeniu normalności również **estymatorami nieobciążonymi o minimalnej wariancji**) parametrów μ oraz Σ są statystyki:

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\Sigma} = S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})'$$

Analiza składowych głównych

Analiza składowych głównych (PCA) jest techniką redukcji wymiaru. Jej celem jest znalezienie niewielkiej liczby (zazwyczaj dwóch lub trzech) składowych głównych, które wyjaśniają w maksymalnym stopniu całkowita wariancję z próby p zmiennych pierwotnych $X_1, X_2, \ldots, X_p$ tj. wielkość

$$\sum_{i=1}^{p} Var(X_i) = tr(\Sigma)$$

Analiza składowych głównych

Składowe główne są unormowanymi kombinacjami liniowymi zmiennych pierwotnych:

$$Z_1 = a_1'X$$

$$Z_2 = a_2'X$$

$$\dots$$

$$Z_P = a_p'X$$

Przekształcone zmienne (składowe główne) są ortogonalne i nieskorelowane

Uwaga:

Ponieważ macierz Σ nie jest znana, posługujemy się jej oszacowaniem z próby, tj. macierzą S

Algorytm składowych głównych

1. Wyznaczamy współczynniki $a_1 = (a_{11}, \dots, a_{1p})'$ pierwszej składowej głównej, tak aby
   a) Zmaksymalizować wariancję zmiennej Z_1
   $$a_1' S a_1$$
   b) Długość wektora a_1 była równa jeden
   $$a_1' a_1 = 1$$
2. Wyznaczamy współczynniki $a_2 = (a_{21}, \dots, a_{2p})'$ drugiej składowej głównej, tak aby
   a) Zmaksymalizować wariancję zmiennej Z_2
   $$a_2' S a_2$$
   b) Długość wektora a_2 była równa jeden
   $$a_2' a_2 = 1$$
   c) Składowa Z_2 była nieskorelowana z Z_1
   $$a_2' a_1 = 0$$
3. Powtarzamy krok 2 (dla następnych składowych głównych) aż do otrzymania współczynników wszystkich p składowych głównych

Własności składowych głównych

- Wektor a_i jest wektorem charakterystycznym odpowiadającym i-tej co do wielkości wartości własnej λ_i macierzy S
- Zachodzi równość
$$\sum_{i=1}^{p} Var(X_i) = \sum_{i=1}^{p} Var(Z_i) = \sum_{i=1}^{p} \lambda_i = tr(S)$$
- W analizie składowych głównych oczekujemy, że dla pewnego małego k, suma $\lambda_1 + \lambda_2 + \cdots + \lambda_k$ będzie bliska $tr(S) = \lambda_1 + \lambda_2 + \cdots + \lambda_p$. Jeśli tak jest, to k pierwszych składowych głównych wyjaśnia dobrze zmienność wektora X.

Własności składowych głównych

- Pozostałe $p-k$ składowe główne wnoszą niewiele, ponieważ mają małe wariancje z próby
- Wskaźnik
$$\frac{\lambda_1+\cdots+\lambda_k}{\lambda_1+\cdots+\lambda_p}100\%$$
  jest procentową miarą wyjaśnienia zmienności wektora X przez pierwszych k składowych głównych
- Wartość modułu współczynnika a_{ji} w j-tej składowej głównej, pokazuje wkład w jej budowę i-tej zmiennej pierwotnej (z uwzględnieniem udziału pozostałych zmiennych pierwotnych)

R

Funkcje związane z analizą składowych głównych:

princomp – analiza składowych głównych, procedura główna