DATA MINING – praca domowa nr 1

MARTA
SOMMER
BSMAD
19.03.2014 r.

1) $$(a'\bar{x}_2 - a'\bar{x}_1)^2 \overset{?}{=} a'(\bar{x}_2 - \bar{x}_1)(\bar{x}_2 - \bar{x}_1)'a$$

Dow:

$$a = \begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix} \qquad \bar{x}_1 = \begin{bmatrix} x_1^{(1)} \\ \vdots \\ x_p^{(1)} \end{bmatrix} \qquad \bar{x}_2 = \begin{bmatrix} x_1^{(2)} \\ \vdots \\ x_p^{(2)} \end{bmatrix}$$

$$a'\bar{x}_2 = [a_1 \dots a_p] \cdot \begin{bmatrix} x_1^{(2)} \\ \vdots \\ x_p^{(2)} \end{bmatrix} = \sum_{i=1}^{p} a_i x_i^{(2)}$$

$$a'\bar{x}_1 = \sum_{i=1}^{p} a_i x_i^{(1)}$$

$$(a'\bar{x}_2 - a'\bar{x}_1)^2 = \left(\sum_{i=1}^{p} a_i x_i^{(2)} - \sum_{i=1}^{p} a_i x_i^{(1)} \right)^2 =$$

$$= \left[\sum_{i=1}^{p} \left(a_i x_i^{(2)} - a_i x_i^{(1)} \right) \right]^2 = \left[\sum_{i=1}^{p} a_i \left(x_i^{(2)} - x_i^{(1)} \right) \right]^2$$

$$a'(\bar{x}_2 - \bar{x}_1) = [a_1 \dots a_p] \begin{bmatrix} x_1^{(2)} - x_1^{(1)} \\ \vdots \\ x_p^{(2)} - x_p^{(2)} \end{bmatrix} = \sum_{i=1}^{p} a_i \left(x_i^{(2)} - x_i^{(1)} \right)$$

$$(\bar{x}_2 - \bar{x}_1)'a = [x_1^{(2)} - x_1^{(1)}, \dots, x_p^{(2)} - x_p^{(2)}] \cdot \begin{bmatrix} a_1 \\ \vdots \\ a_p \end{bmatrix} =$$

$$= \sum_{i=1}^{p} a_i \left(x_i^{(2)} - x_i^{(1)} \right)$$

$$a'(\bar{x}_2 - \bar{x}_1)(\bar{x}_2 - \bar{x}_1)'a =$$

$$= \left[\sum_{i=1}^{p} a_i \left(x_i^{(2)} - x_i^{(1)} \right) \right] \left[\sum_{i=1}^{p} a_i \left(x_i^{(2)} - x_i^{(1)} \right) \right] =$$

$$= \left[\sum_{i=1}^{p} a_i \left(x_i^{(2)} - x_i^{(1)} \right) \right]^2$$

$$L = P$$

2) $$\overline{x} \stackrel{?}{=} \frac{n_1}{n_1+n_2}\overline{x_1} + \frac{n_2}{n_1+n_2}\overline{x_2}$$

Dow:

Dla przypadku jednowymiarowego:

$$\underbrace{x_1, \ldots\ldots, x_{n_1}}_{n_1}, \underbrace{x_{n_1+1}, \ldots\ldots, x_{n_1+n_2}}_{n_2}$$

$$\overline{x_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i$$

$$\overline{x_2} = \frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2} x_i$$

$$\overline{x} = \frac{1}{n_1+n_2}\sum_{i=1}^{n_1+n_2} x_i$$

$$\overline{x} = \frac{1}{n_1+n_2}\sum_{i=1}^{n_1+n_2} x_i = \frac{1}{n_1+n_2}\left[\sum_{i=1}^{n_1} x_i + \sum_{i=n_1+1}^{n_1+n_2} x_i\right] =$$

$$= \frac{1}{n_1+n_2}\left[n_1\left(\frac{1}{n_1}\sum_{i=1}^{n_1} x_i\right) + n_2\left(\frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2} x_i\right)\right] =$$

$$= \frac{1}{n_1+n_2}\left[n_1\overline{x_1} + n_2\overline{x_2}\right] = \frac{n_1}{n_1+n_2}\overline{x_1} + \frac{n_2}{n_1+n_2}\overline{x_2}$$

Rozumowanie przenosi się na elementy wektorów.

3) $$a'(\overline{x_2}-\overline{x_1})(\overline{x_2}-\overline{x_1})'a \stackrel{?}{=} \frac{n_1+n_2}{n_1 n_2}a'\left[\sum_{k=1}^{2} n_k(\overline{x_k}-\overline{x})(\overline{x_k}-\overline{x})'\right]a$$

Wystarczy oczywiście wykazać, że:

$$(\overline{x_2}-\overline{x_1})(\overline{x_2}-\overline{x_1})' = \frac{n_1+n_2}{n_1 n_2}\sum_{k=1}^{2} n_k(\overline{x_k}-\overline{x})(\overline{x_k}-\overline{x})'$$

Czyli:

$$(\overline{x_2}-\overline{x_1})(\overline{x_2}-\overline{x_1})' \stackrel{?}{=} \frac{n_1+n_2}{n_1 n_2}n_1(\overline{x_1}-\overline{x})(\overline{x_1}-\overline{x})' + \frac{n_1+n_2}{n_1 n_2}n_2(\overline{x_2}-\overline{x})(\overline{x_2}-\overline{x})'$$

Dow:

$$\frac{n_1+n_2}{\cancel{n_1} n_2}\cancel{n_1}(\overline{x_1}-\overline{x})(\overline{x_1}-\overline{x})' =$$

$$= \frac{n_1+n_2}{n_2}\left(\bar{x}_1 - \frac{n_1}{n_1+n_2}\bar{x}_1 - \frac{n_2}{n_1+n_2}\bar{x}_2\right)\left(\frac{\cancel{n_1}+n_2-\cancel{n_1}}{n_1+n_2}\bar{x}_1 - \frac{n_2}{n_1+n_2}\bar{x}_2\right)' =$$

$$= \frac{n_1+n_2}{n_2}\left(\frac{n_2}{n_1+n_2}\bar{x}_1 - \frac{n_2}{n_1+n_2}\bar{x}_2\right)\left(\frac{n_2}{n_1+n_2}\bar{x}_1 - \frac{n_2}{n_1+n_2}\bar{x}_2\right)' =$$

$$= \frac{\cancel{n_1+n_2}}{\cancel{n_2}}\left(\frac{n_2}{n_1+n_2}\right)^{\cancel{2}}(\bar{x}_1-\bar{x}_2)(\bar{x}_1-\bar{x}_2)' =$$

$$= \frac{n_2}{n_1+n_2}(\bar{x}_2-\bar{x}_1)(\bar{x}_2-\bar{x}_1)'$$

$$\frac{n_1+n_2}{n_1\cancel{n_2}}\cancel{n_2}(\bar{x}_2-\bar{x})(\bar{x}_2-\bar{x})' =$$

$$= \frac{n_1+n_2}{n_1}\left(\bar{x}_2 - \frac{n_1}{n_1+n_2}\bar{x}_1 - \frac{n_2}{n_1+n_2}\bar{x}_2\right)\left(\frac{n_1+\cancel{n_2}-\cancel{n_2}}{n_1+n_2}\bar{x}_2 - \frac{n_1}{n_1+n_2}\bar{x}_1\right)'$$

$$= \frac{\cancel{n_1+n_2}}{\cancel{n_1}}\left(\frac{n_1}{n_1+n_2}\right)^{\cancel{2}}(\bar{x}_2-\bar{x}_1)(\bar{x}_2-\bar{x}_1)' =$$

$$= \frac{n_1}{n_1+n_2}(\bar{x}_2-\bar{x}_1)(\bar{x}_2-\bar{x}_1)'$$

$$\frac{n_2}{n_1+n_2}(\bar{x}_2-\bar{x}_1)(\bar{x}_2-\bar{x}_1)' + \frac{n_1}{n_1+n_2}(\bar{x}_2-\bar{x}_1)(\bar{x}_2-\bar{x}_1)' =$$

$$= \frac{\cancel{n_1+n_2}}{\cancel{n_1+n_2}}(\bar{x}_2-\bar{x}_1)(\bar{x}_2-\bar{x}_1)' = (\bar{x}_2-\bar{x}_1)(\bar{x}_2-\bar{x}_1)'$$

$\blacksquare$

4) dla $g=2$:

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

Więc:

$$\frac{(a'\bar{x}_2 - a'\bar{x}_1)^2}{a'\mathcal{W}a} = \frac{\frac{n_1+n_2}{n_1n_2}a'Ba}{a'\mathcal{W}a}$$

Chcąc maksymalizować to wyrażenie możemy oczywiście opuścić stałą. Szukamy więc:

$$\underset{a\in\mathbb{R}^p}{\operatorname{argmax}}\ \frac{a'Ba}{a'\mathcal{W}a}$$

$\blacksquare$