

Witold Miszczak

Akademia Ekonomiczna we Wrocławiu

REGRESJA ORTOGONALNA I KLASYFIKACJA

1. Wstęp

Zadanie jest następujące: dla zaobserwowanej zmiennej niezależnej X i odpowiadającej jej wartości zmiennej zależnej Y znaleźć linię regresji ortogonalnej mającej tę własność, że suma odległości wszystkich obserwacji od regresji jest minimalna. Innymi słowy, jeżeli regresja ma postać $AY + BX + C = 0$, to parametry A , B i C trzeba wyznaczyć z następującego kryterium

$$\sum_{i=1}^n \frac{|Ay_i + Bx_i + C|}{\sqrt{A^2 + B^2}} = \min_{A,B,C}.$$

Zadanie to jest nieco skomplikowane i spróbujemy znaleźć parametry regresji na podstawie innego kryterium opartego na metodzie najmniejszych kwadratów (MNK):

$$\sum_{i=1}^n \frac{(Ay_i + Bx_i + C)^2}{A^2 + B^2} = \min_{A,B,C}.$$

Uprościmy nieco to wyrażenie:

$$\sum_{i=1}^n \frac{\left(y_i + \frac{B}{A}x_i + \frac{C}{A}\right)^2}{1 + \left(\frac{B}{A}\right)^2} = \sum_{i=1}^n \frac{(y_i + ax_i + b)^2}{1 + a^2} = \min_{a,b}.$$

Zadanie znalezienia trzech parametrów zredukowaliśmy do znalezienia dwu. Obliczmy pochodną po b :

$$\frac{2}{1+a^2} \sum_{i=1}^n (y_i + ax_i + b) = \frac{2}{1+a^2} \left(\sum_{i=1}^n y_i + a \sum_{i=1}^n x_i + nb \right) = 0$$

Stąd po prostych obliczeniach mamy

$$b = -\bar{y} - a\bar{x}.$$

Obliczmy teraz pochodną funkcji kryterium po a :

$$\begin{aligned} \frac{1}{1+a^2} \sum_{i=1}^n (y_i + ax_i - \bar{y} - a\bar{x})^2 &= \min_a, \\ \frac{1}{1+a^2} \sum_{i=1}^n (y_i - \bar{y} + a(x_i - \bar{x}))^2 &= \frac{n}{1+a^2} (\sigma_y^2 + a^2 \sigma_x^2 + 2a \operatorname{cov}(x, y)) = \min_a, \\ -\frac{2na}{(1+a^2)^2} (\sigma_y^2 + a^2 \sigma_x^2 + 2a \operatorname{cov}(x, y)) &+ \frac{2n}{1+a^2} (2a \sigma_x^2 + 2 \operatorname{cov}(x, y)) = 0, \\ a \sigma_y^2 + a^3 \sigma_x^2 + 2a^2 \operatorname{cov}(x, y) - a \sigma_x^2 - a^3 \sigma_x^2 - \operatorname{cov}(x, y) - a^2 \operatorname{cov}(x, y) &= 0, \\ a^2 \operatorname{cov}(x, y) + a(\sigma_y^2 - \sigma_x^2) - \operatorname{cov}(x, y) &= 0. \end{aligned}$$

Rozwiązaniem powyższego równania kwadratowego jest:

$$\begin{aligned} a_1 &= \frac{\sigma_x^2 - \sigma_y^2 - \sqrt{(\sigma_y^2 - \sigma_x^2)^2 + 4 \operatorname{cov}^2(x, y)}}{2 \operatorname{cov}(x, y)}, \\ a_2 &= \frac{\sigma_x^2 - \sigma_y^2 + \sqrt{(\sigma_y^2 - \sigma_x^2)^2 + 4 \operatorname{cov}^2(x, y)}}{2 \operatorname{cov}(x, y)}. \end{aligned}$$

Otrzymaliśmy, co nietrudno sprawdzić, współczynniki kątowe dwu prostych wzajemnie ortogonalnych. Jedna z nich realizuje minimum funkcji kryterium, a druga jego maksimum.

2. Wielozmienna regresja ortogonalna

Postawmy teraz ogólniejsze zadanie: niech macierz $\mathbf{X}$ ma postać

$$\mathbf{X} = \begin{pmatrix} y_1 & x_{11} & \dots & x_{m1} & 1 \\ \dots & \dots & \dots & \dots & \dots \\ y_n & x_{1n} & \dots & x_{mn} & 1 \end{pmatrix},$$

a wektor parametrów $\mathbf{a}$

$$\mathbf{a} = \begin{pmatrix} 1 \\ a_1 \\ \dots \\ a_m \\ c \end{pmatrix}.$$

Wprowadźmy dodatkowo następujące macierze i wektory:

$$\mathbf{a}_c = \begin{pmatrix} 1 \\ a_1 \\ \dots \\ a_m \end{pmatrix}, \quad \mathbf{X}_c = \begin{pmatrix} y_1 & x_{11} & \dots & x_{m1} \\ \dots & \dots & \dots & \dots \\ y_n & x_{1n} & \dots & x_{mn} \end{pmatrix} \text{ i } \mathbf{1} = \begin{pmatrix} 1 \\ \dots \\ 1 \end{pmatrix}_{n \times 1}$$

Przekształćmy teraz następującą formę kwadratową:

$$\begin{aligned} \mathbf{a}^T \mathbf{X}^T \mathbf{X} \mathbf{a} &= \sum_{i=1}^n \left(y_i + \sum_{j=1}^m a_j x_{ji} + c \right)^2 = \\ &= \sum_{i=1}^n \left(y_i + \sum_{j=1}^m a_j x_{ji} \right)^2 + 2c \sum_{i=1}^n \left(y_i + \sum_{j=1}^m a_j x_{ji} + c \right) + nc^2 = \\ &= \mathbf{a}_c^T \mathbf{X}_c^T \mathbf{X}_c \mathbf{a}_c + 2c \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c + nc^2. \end{aligned}$$

Ogólnie kryterium zapiszemy w postaci

$$K = \frac{\mathbf{a}_c^T \mathbf{X}_c^T \mathbf{X}_c \mathbf{a}_c + 2c \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c + nc^2}{\mathbf{a}_c^T \mathbf{a}_c} = \min_{\mathbf{a}_c, c}. \quad (1)$$

Znajdźmy minimum po c :

$$\frac{\partial K}{\partial c} = \frac{2}{\mathbf{a}_c^T \mathbf{a}_c} (nc + \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c) = 0.$$

Stąd

$$c = -\frac{\mathbf{1}^T \mathbf{X}_c \mathbf{a}_c}{n} = -\bar{y} - \sum_{i=1}^m a_i \bar{x}_i.$$

Wstawmy do K obliczone c :

$$K = \frac{\mathbf{a}_c^T \mathbf{X}_c^T \mathbf{X}_c \mathbf{a}_c - \frac{2}{n} \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c + \frac{1}{n} \mathbf{a}_c^T \mathbf{X}_c^T \mathbf{1} \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c}{\mathbf{a}_c^T \mathbf{a}_c} = \min_{\mathbf{a}_c}.$$

Możemy posłużyć się dwiema metodami znalezienia tego minimum: obliczyć pochodną K po $\mathbf{a}_c$ lub przekształcić bezpośrednio powyższe wyrażenie. Wybierzemy najpierw drugą metodę:

$$\begin{aligned} nK \mathbf{a}_c^T \mathbf{a}_c &= n \mathbf{a}_c^T \mathbf{X}_c^T \mathbf{X}_c \mathbf{a}_c - 2 \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c + \mathbf{a}_c^T \mathbf{X}_c^T \mathbf{1} \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c = \\ &= \mathbf{a}_c^T \left(n \mathbf{X}_c^T \mathbf{X}_c + \mathbf{X}_c^T \mathbf{1} \mathbf{1}^T \mathbf{X}_c \right) \mathbf{a}_c + \\ &- 2 \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c = \mathbf{a}_c^T \left(n^2 \left(\frac{1}{n} \mathbf{X}_c^T \mathbf{X}_c - \frac{1}{n^2} \mathbf{X}_c^T \mathbf{1} \mathbf{1}^T \mathbf{X}_c + \frac{2}{n^2} \mathbf{X}_c^T \mathbf{1} \mathbf{1}^T \mathbf{X}_c \right) \mathbf{a}_c + \right. \\ &- 2 \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c = n^2 \mathbf{a}_c^T \mathbf{S} \mathbf{a}_c + 2 \mathbf{a}_c^T \mathbf{X}_c^T \mathbf{1} \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c - 2 \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c = \\ &= n^2 \mathbf{a}_c^T \mathbf{S} \mathbf{a}_c + 2 \left(\mathbf{a}_c^T \mathbf{X}_c^T \mathbf{1} - \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c \right) \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c = n^2 \mathbf{a}_c^T \mathbf{S} \mathbf{a}_c. \end{aligned}$$

Drugi składnik przed ostatnim znakiem równości w powyższym wyrażeniu jest równy zeru. W przekształceniach skorzystaliśmy z następujących, łatwych do udowodnienia wzorów:

$$\begin{aligned} \frac{1}{n} \mathbf{X}_c^T \mathbf{1} &= \begin{pmatrix} \bar{y} \\ \bar{x}_1 \\ \dots \\ \bar{x}_m \end{pmatrix}, \\ \mathbf{S} &= \frac{1}{n} \mathbf{X}_c^T \mathbf{X}_c - \frac{1}{n^2} \mathbf{X}_c^T \mathbf{1} \mathbf{1}^T \mathbf{X}_c, \end{aligned}$$

gdzie $\mathbf{S}$ jest macierzą kowariancji zmiennych $(Y, X_1, \dots, X_m)$. Przekształcając dalej, mamy

$$\begin{aligned} n^2 \mathbf{a}_c^T \mathbf{S} \mathbf{a}_c - nK \mathbf{a}_c^T \mathbf{a}_c &= 0, \\ \mathbf{a}_c^T (n\mathbf{S} - K\mathbf{I}) \mathbf{a}_c &= 0. \end{aligned}$$

Ostatnia równość jest spełniona, gdy

$$\det(n\mathbf{S} - K\mathbf{I}) = 0.$$

Oznacza to, że K jest najmniejszą wartością własną macierzy $\mathbf{S}$, a $\mathbf{a}_c$ jest odpowiadającym jej wektorem własnym z pierwszą współrzędną równą 1.

Algorytm estymowania parametrów regresji ortogonalnej jest następujący:

1. Wyznaczamy macierz kowariancji $\mathbf{S}$.

2. Znajdujemy jej najmniejszą wartość własną i odpowiadający jej wektor własny. Wśród wszystkich możliwych wektorów własnych wybieramy ten, którego pierwsza składowa jest równa 1. Wektor ten jest wektorem parametrów $\mathbf{a}_c$.

3. Stałą c wyznaczamy z wzoru

$$c = -\bar{y} - \sum_{i=1}^m a_i \bar{x}_i.$$

Równanie regresji ortogonalnej ma postać

$$Y = -\sum_{i=1}^m a_i X_i - c.$$

Spróbujmy teraz obliczyć $\mathbf{a}_c$ pierwszym sposobem, tzn. rozwiązując równanie

$$\frac{\partial K}{\partial \mathbf{a}_c} = \mathbf{0}.$$

$$\begin{aligned} \frac{\partial K}{\partial \mathbf{a}_c} &= -\frac{1}{(\mathbf{a}_c^T \mathbf{a}_c)^2} \frac{\partial (\mathbf{a}_c^T \mathbf{a}_c)}{\partial \mathbf{a}_c} (\mathbf{a}_c^T \mathbf{X}_c^T \mathbf{X}_c \mathbf{a}_c + 2c \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c + nc^2) + \\ &+ \frac{1}{\mathbf{a}_c^T \mathbf{a}_c} \left(\frac{\partial \mathbf{a}_c^T \mathbf{X}_c^T \mathbf{X}_c \mathbf{a}_c}{\partial \mathbf{a}_c} + 2c \frac{\partial \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c}{\partial \mathbf{a}_c} \right) = \\ &= -\frac{2}{(\mathbf{a}_c^T \mathbf{a}_c)^2} (\mathbf{a}_c \mathbf{a}_c^T \mathbf{X}_c^T \mathbf{X}_c \mathbf{a}_c + 2c \mathbf{a}_c \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c + nc^2 \mathbf{a}_c) + \\ &+ \frac{2}{\mathbf{a}_c^T \mathbf{a}_c} (\mathbf{X}_c^T \mathbf{X}_c \mathbf{a}_c + c \mathbf{X}_c^T \mathbf{1}) = \mathbf{0}. \end{aligned}$$

Przekształcając dalej, mamy

$$\begin{aligned} \mathbf{X}_c^T \mathbf{X}_c \mathbf{a}_c \mathbf{a}_c^T \mathbf{a}_c + c \mathbf{X}_c^T \mathbf{1} \mathbf{a}_c^T \mathbf{a}_c - \mathbf{a}_c \mathbf{a}_c^T \mathbf{X}_c^T \mathbf{X}_c \mathbf{a}_c - 2c \mathbf{a}_c \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c - nc^2 \mathbf{a}_c &= \mathbf{0}. \\ \mathbf{X}_c^T \mathbf{X}_c \mathbf{a}_c \mathbf{a}_c^T \mathbf{a}_c - \frac{1}{n} \mathbf{X}_c^T \mathbf{1} \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c \mathbf{a}_c^T \mathbf{a}_c - \mathbf{a}_c \mathbf{a}_c^T \mathbf{X}_c^T \mathbf{X}_c \mathbf{a}_c + \\ + \frac{2}{n} \mathbf{a}_c \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c - \frac{1}{n} \mathbf{a}_c \mathbf{a}_c^T \mathbf{X}_c^T \mathbf{1} \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c &= \\ = n \left(\frac{1}{n} \mathbf{X}_c^T \mathbf{X}_c - \frac{1}{n^2} \mathbf{X}_c^T \mathbf{1} \mathbf{1}^T \mathbf{X}_c \right) \mathbf{a}_c \mathbf{a}_c^T \mathbf{a}_c + \\ - n \mathbf{a}_c \mathbf{a}_c^T \left(\frac{1}{n} \mathbf{X}_c^T \mathbf{X}_c + \frac{1}{n^2} \mathbf{X}_c^T \mathbf{1} \mathbf{1}^T \mathbf{X}_c \right) \mathbf{a}_c + \frac{2}{n} \mathbf{a}_c \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c \mathbf{1}^T \mathbf{X}_c \mathbf{a}_c &= \end{aligned}$$

$$\begin{aligned}
&= n\mathbf{S}\mathbf{a}_c\mathbf{a}_c^T\mathbf{a}_c - n\mathbf{a}_c\mathbf{a}_c^T\left(\frac{1}{n}\mathbf{X}_c^T\mathbf{X}_c - \frac{1}{n^2}\mathbf{X}_c^T\mathbf{1}\mathbf{1}^T\mathbf{X}_c\right)\mathbf{a}_c + \\
&\quad - \frac{2}{n}\mathbf{a}_c\mathbf{a}_c^T\mathbf{X}_c^T\mathbf{1}\mathbf{1}^T\mathbf{X}_c\mathbf{a}_c + \frac{2}{n}\mathbf{a}_c\mathbf{1}^T\mathbf{X}_c\mathbf{a}_c\mathbf{1}^T\mathbf{X}_c\mathbf{a}_c = \\
&= n\mathbf{S}\mathbf{a}_c\mathbf{a}_c^T\mathbf{a}_c - n\mathbf{a}_c\mathbf{a}_c^T\mathbf{S}\mathbf{a}_c - \frac{2}{n}\mathbf{a}_c\left(\mathbf{a}_c^T\mathbf{X}_c^T\mathbf{1} - \mathbf{1}^T\mathbf{X}_c\mathbf{a}_c\right)\mathbf{1}^T\mathbf{X}_c\mathbf{a}_c = \\
&= n\mathbf{S}\mathbf{a}_c\mathbf{a}_c^T\mathbf{a}_c - n\mathbf{a}_c\mathbf{a}_c^T\mathbf{S}\mathbf{a}_c = \mathbf{0}.
\end{aligned}$$

Stąd

$$\begin{aligned}
&(\mathbf{S}\mathbf{a}_c\mathbf{a}_c^T - \mathbf{a}_c\mathbf{a}_c^T\mathbf{S})\mathbf{a}_c = \mathbf{0}, \\
&\left(\left(\mathbf{a}_c\mathbf{a}_c^T\mathbf{S}\right)^T - \mathbf{a}_c\mathbf{a}_c^T\mathbf{S}\right)\mathbf{a}_c = \mathbf{0}.
\end{aligned} \tag{2}$$

Rozwiązując ostatni układ równań ze względu na $\mathbf{a}_c$, otrzymujemy poszukiwane parametry regresji ortogonalnej. Do wektora parametrów $\mathbf{a}_c$ należy dołączyć wyraz wolny c . Jak widać, ten ostatni sposób jest znacznie bardziej skomplikowany niż poprzedni. Nie sądzę, żeby istniał pakiet, który potrafi szybko rozwiązać układ równań wielomianowych trzeciego stopnia. Inna trudność polega na tym, że symetryczna macierz kwadratowa $\mathbf{a}_c\mathbf{a}_c^T$ nie jest odwracalna, gdyż jej wyznacznik jest równy zeru ze względu na współliniowość wszystkich wierszy tej macierzy z pierwszym wierszem. Również nieodwracalna jest niesymetryczna macierz kwadratowa $\mathbf{a}_c\mathbf{a}_c^T\mathbf{S}$. Pewnym ułatwieniem jest to, że macierz drugiego układu w nawiasie (2) ma na przekątnej same zera, a elementy pod przekątną mają znaki przeciwne do elementów nad przekątną. Jedno z równań jest kombinacją liniową pozostałych. Nietrudno się przekonać, że rozwiązaniem jest wektor własny macierzy $\mathbf{S}$. Z własności wektorów własnych mamy

$$\lambda\mathbf{a}_c = \mathbf{S}\mathbf{a}_c.$$

Stąd

$$\mathbf{S}\mathbf{a}_c\mathbf{a}_c^T\mathbf{a}_c - \mathbf{a}_c\mathbf{a}_c^T\mathbf{S}\mathbf{a}_c = \lambda\mathbf{a}_c\mathbf{a}_c^T\mathbf{a}_c - \lambda\mathbf{a}_c\mathbf{a}_c^T\mathbf{a}_c = \mathbf{0}.$$

Jeżeli wybierzemy wektor własny odpowiadający minimalnej wartości własnej, to będzie on minimalizował kryterium (1), gdyż odległości liczone są wzdłuż najkrótszej osi elipsoidy rozrzutu punktów. Gdybyśmy wybrali maksymalną wartość własną, to kryterium (1) przyjęłoby wartość maksymalną. Tym razem odległości byłyby liczone wzdłuż najdłuższej osi elipsoidy rozrzutu punktów. Prostych ortogonalnych jest tyle, ile niezerowych wartości własnych ma macierz $\mathbf{S}$ plus jedna odpowiadająca zerowej wartości własnej (gdyby taka była), co oznacza, że wszyst-

kie zaobserwowane punkty są współliniowe. Ta prosta, która odpowiada najmniejszej wartości własnej, jest regresją ortogonalną.

Nawiasem mówiąc, układy równań (2) podają pewien inny sposób wyznaczania wszystkich wektorów własnych macierzy $\mathbf{S}$, bez uprzedniego wyliczania wartości własnych. Pierwszy układ wyznacza wektory własne dla dowolnych macierzy, a drugi dla macierzy symetrycznych. Wartości własne dla znanych wektorów własnych łatwo jest policzyć z wzoru

$$\lambda = \frac{\mathbf{a}_c^T \mathbf{S} \mathbf{a}_c}{\mathbf{a}_c^T \mathbf{a}_c}.$$

Problem leży w tym, jak te układy rozwiązać. Być może istnieje jakiś prosty algorytm iteracyjny, który pozwala przynajmniej w przybliżeniu podać rozwiązania. Jest to materiał do dalszego przemyślenia.

3. Wielozmienna ważona regresja ortogonalna

Żałómy teraz, że nie wszystkie obserwacje będą traktowane równoprawnie i niech przekątniowa macierz wag ma postać

$$\mathbf{W} = \begin{pmatrix} w_1 & 0 & 0 \\ 0 & w_2 & 0 \\ \dots & \dots & \dots \\ 0 & 0 & w_n \end{pmatrix}.$$

Wtedy kryterium K zapiszemy w następujący sposób

$$K = \frac{\mathbf{a}_c^T \mathbf{X}_c^T \mathbf{W} \mathbf{X}_c \mathbf{a}_c + 2c \mathbf{1}^T \mathbf{W} \mathbf{X}_c \mathbf{a}_c + c^2 \mathbf{1}^T \mathbf{W} \mathbf{1}}{\mathbf{a}_c^T \mathbf{a}_c} = \min_{\mathbf{a}_c, c}.$$

Powtarzając powyższe obliczenia, otrzymujemy

$$c = -\frac{\mathbf{1}^T \mathbf{W} \mathbf{X}_c \mathbf{a}_c}{\mathbf{1}^T \mathbf{W} \mathbf{1}} = -\bar{y}^* - \sum_{i=1}^m a_i \bar{x}_i^*,$$

gdzie

$$\bar{x}_i^* = \frac{\sum_{j=1}^n w_j x_{ij}}{\sum_{j=1}^n w_j}, \quad \bar{y}^* = \frac{\sum_{j=1}^n w_j y_j}{\sum_{j=1}^n w_j}.$$

Wektor parametrów $\mathbf{a}_c$ jest wektorem własnym macierzy $\mathbf{S}^*$, odpowiadającym jej najmniejszej wartości własnej, gdyż w wyniku przekształceń według drugiego ze sposobów znajdowania rozwiązania dla kryterium K dostajemy

$$\det(\mathbf{1}^T \mathbf{W} \mathbf{I} \mathbf{S}^* - K \mathbf{I}) = 0,$$

gdzie

$$\mathbf{S}^* = \begin{pmatrix} s_{00} & s_{01} & s_{0m} \\ s_{01} & s_{11} & s_{1m} \\ \dots & & \dots \\ s_{0m} & s_{1m} & s_{mm} \end{pmatrix}.$$

Macierz $\mathbf{S}^*$ jest macierzą ważonych kowariancji zmiennych $Y, X_1, \dots, X_m$, o elementach

$$s_{ij} = \frac{\sum_{k=1}^n w_k x_{ik} x_{jk}}{\sum_{k=1}^n w_k} - \bar{x}_i^* \bar{x}_j^*,$$

gdzie obserwacje y_i oznaczamy jako x_{0i} .

4. Klasyfikacja danych niejednorodnych ze względu na regresję ortogonalną

Zakładamy, że zbiór danych $\mathbf{X}$ nie jest jednorodny i trzeba w nim wyodrębnić klasy o właściwych dla nich parametrach liniowych regresji ortogonalnych. Założmy na początku, że liczba klas jest dana i równa k . Oznaczmy klasy przez C_i , $i = \overline{1, k}$. Rozpatrzmy dwa przypadki klasyfikacji: minimalizującą sumę kwadratów odległości obserwacji zmiennej zależnej Y od hiperpłaszczyzny regresji

$$K_1 = \sum_{i=1}^k \sum_{\mathbf{x}_j \in C_i} d_{ji} = \min$$

i klasyfikację rozmytą

$$K_2 = \sum_{j=1}^n \sum_{i=1}^k w_{ij}^2 d_{ji} = \min,$$

gdzie

$$d_{ji} = \frac{(\mathbf{a}_{ci}^T \mathbf{x}_j + c_i)^2}{\mathbf{a}_{ci}^T \mathbf{a}_{ci}}$$

jest kwadratem odległości obserwacji $\mathbf{x}_j$ rozumianej jako

$$\mathbf{x}_j^T = (y_j \quad x_{1j} \quad \dots \quad x_{mj})$$

od regresji ortogonalnej w klasie C_i , zaś

$$w_{ij} = \varphi_i(\mathbf{x}_j),$$

jest wartością funkcji przynależności ϕ obserwacji $\mathbf{x}_j$ do klasy C_i . Postulujemy, by

$$\sum_{i=1}^k w_{ij} = 1. \quad (3)$$

Jak widać, oba te przypadki kryteriów można połączyć w jedno

$$K = \sum_{j=1}^n \sum_{i=1}^k w_{ij}^2 d_{ji} = \min,$$

gdzie w przypadku K_1 wagi w_{ij} przyjmują wartość 1, gdy $\mathbf{x}_j$ należy do klasy C_i , i zero w przeciwnym razie.

Pomimo że oba kryteria można zapisać w ten sam sposób, to w pierwszym przypadku będziemy szukali minimum bezwarunkowego, w drugim zaś warunkowego ze względu na (3). Minimum warunkowe pozwoli nam na wyznaczenie optymalnych wag. Zapiszmy je więc

$$K_2 = \sum_{j=1}^n \sum_{i=1}^k w_{ij}^2 d_{ji} - \sum_{j=1}^n \lambda_j \left(\sum_{i=1}^k w_{ij} - 1 \right) = \min_{w_{ij}},$$

gdzie λ_j są mnożnikami Lagrange'a. Obliczenie pochodnej K_2 po w_{ij} prowadzi do równania

$$2w_{ij}d_{ji} - \lambda_j = 0,$$

z którego otrzymujemy

$$\sum_{i=1}^k w_{ij} = \sum_{i=1}^k \frac{\lambda_j}{2d_{ji}} = 1.$$

Po obliczeniu λ_j i prostych przekształceniach otrzymujemy

$$w_{ij} = \frac{1}{\sum_{i=1}^k \frac{1}{d_{ji}}}.$$

Ponieważ kryterium K można zapisać jako kryterium dla wielozmiennej ważonej regresji ortogonalnej, w dalszej analizie skorzystamy z wyników poprzedniego punktu. Macierz diagonalna $\mathbf{W}$ zawiera na przekątnej wagi będące kwadratami stopni przynależności obserwacji do klasy. Takich macierzy konstruujemy tyle, ile jest klas. Dla kryterium K_i na przekątnej macierzy $\mathbf{W}_i$, $i = \overline{1, k}$ znajdują się zera i jedynki.

Kryterium dla zagadnienia klasyfikacji przy ustalonych wagach ma postać

$$K = \sum_{i=1}^k \frac{\mathbf{a}_{ci}^T \mathbf{X}_c^T \mathbf{W}_i \mathbf{X}_c \mathbf{a}_{ci} + 2c_i \mathbf{1}^T \mathbf{W}_i \mathbf{X}_c \mathbf{a}_{ci} + c_i^2 \mathbf{1}^T \mathbf{W}_i \mathbf{1}}{\mathbf{a}_{ci}^T \mathbf{a}_{ci}} = \min_{\mathbf{a}_{ci}, c_i, i=\overline{1, k}}, \quad (4)$$

gdzie

$$\mathbf{a}_{ci} = \begin{pmatrix} 1 \\ a_{li} \\ \dots \\ a_{mi} \end{pmatrix}$$

jest wektorem parametrów, a c_i wyrazem wolnym regresji ortogonalnej w klasie C_i .

Pochodna K po c_i ma postać

$$\frac{\partial K}{\partial c_i} = \frac{2}{\mathbf{a}_{ci}^T \mathbf{a}_{ci}} (n_i c_i + \mathbf{1}^T \mathbf{W}_i \mathbf{X}_c \mathbf{a}_{ci}) = 0.$$

W wyniku otrzymujemy

$$c_i = -\frac{\mathbf{1}^T \mathbf{W}_i \mathbf{X}_c \mathbf{a}_{ci}}{\mathbf{1}^T \mathbf{W}_i \mathbf{1}} = -\bar{y}_i^* - \sum_{j=1}^m a_{ji} \bar{x}_{ji}^*,$$

gdzie $\bar{y}_i^*$, $\bar{x}_{ji}^*$, $i = \overline{1, k}$, $j = \overline{1, m}$ są średnimi ważonymi w klasie C_i .

Po uwzględnieniu w K wartości c_i należy zminimalizować wyrażenie

$$\begin{aligned} K &= \sum_{i=1}^k \frac{\mathbf{a}_{ci}^T \mathbf{X}_c^T \mathbf{W}_i \mathbf{X}_c \mathbf{a}_{ci} + \frac{2}{\mathbf{1}^T \mathbf{W}_i \mathbf{1}} \mathbf{1}^T \mathbf{W}_i \mathbf{X}_c \mathbf{a}_{ci} \mathbf{1}^T \mathbf{W}_i \mathbf{X}_c \mathbf{a}_{ci} + \frac{1}{\mathbf{1}^T \mathbf{W}_i \mathbf{1}} \mathbf{a}_{ci}^T \mathbf{X}_c^T \mathbf{1} \mathbf{W}_i \mathbf{1}^T \mathbf{X}_c \mathbf{a}_{ci}}{\mathbf{a}_{ci}^T \mathbf{a}_{ci}} = \\ &= \sum_{i=1}^k \mathbf{1}^T \mathbf{W}_i \mathbf{1} \frac{\mathbf{a}_{ci}^T \mathbf{S}_i^* \mathbf{a}_{ci}}{\mathbf{a}_{ci}^T \mathbf{a}_{ci}} = \sum_{i=1}^k K_i. \end{aligned}$$

S_i^* jest ważoną macierzą kowariancji w klasie C_i . Ponieważ wszystkie składniki K są nieujemne i każdy z nich jest niezależny, więc by zminimalizować K , wystarczy zminimalizować każdy ze składników K_i , gdzie

$$K_i = \mathbf{1}^T \mathbf{W}_i \mathbf{1} \frac{\mathbf{a}_{ci}^T \mathbf{S}_i^* \mathbf{a}_{ci}}{\mathbf{a}_{ci}^T \mathbf{a}_{ci}}.$$

Przekształcając ostatnie wyrażenie jak uprzednio, otrzymujemy

$$\det(\mathbf{1}^T \mathbf{W}_i \mathbf{1} \mathbf{S}_i^* - K_i \mathbf{I}) = 0,$$

a to oznacza, że K_i jest minimalną wartością własną macierzy $\mathbf{S}_i^*$, a odpowiadający jej wektor własny jest szukanym wektorem parametrów $\mathbf{a}_{ci}$ w klasie C_i .

Algorytm postępowania jest następujący:

1. Ustalamy liczbę klas k . Gdy jest ona nieznana, zaczynamy od minimalnej przewidywalnej liczby klas. Zwiększając liczbę klas o jeden, dojdziemy w pewnym momencie do klas zbyt mało licznych, by dało się wyznaczyć w sposób jednoznaczny regresję ortogonalną. Liczba obserwacji w klasie musi być równa co najmniej m . Gdy zaczną się pojawiać klasy mniej liczne, oznacza to np., że liczba klas jest za duża i w próbie są być może obserwacje nietypowe. Może też być tak, że populacja jest silnie zróżnicowana i istnieją w niej klasy obiektów o niewielkiej liczności. Wtedy próba nie będzie ich zawierała i nie uda się w rzeczywistości zidentyfikować wszystkich klas w populacji bądź próba będzie nadreprezentatywna i takie obserwacje będą traktowane jako nietypowe; gdy się ich wcześniej nie usunie, zostaną na siłę dołączone do innej klasy, mocno zniekształcając całą klasyfikację obiektów. Przez próbę nadreprezentatywną rozumie się próbę reprezentatywną, w której mogą wystąpić elementy z populacji, które ze względu na małe prawdopodobieństwo ich wylosowania nie powinny się w próbie o zadanej liczebności znaleźć, co daje co prawda informację o ich istnieniu w populacji, ale nie pozwala na wyodrębnienie ich jako klasy obiektów ze względu na zbyt małą reprezentację.

2. Dokonujemy wstępnego generowania wektorów wagowych $\mathbf{w}_i$, $i = \overline{1, k}$, których j -ta składowa ($j = \overline{1, n}$) określa przynależność j -tej obserwacji do klasy C_i

$$\mathbf{w}_i^T = (w_{i1} \quad w_{i2} \quad \dots \quad w_{in}).$$

Można to zrobić, stosując

- znane wagi ustalone na podstawie analizy problemu lub zasugerowane przez ekspertów,
- generując wagi losowo,
- ustalając wagi proporcjonalnie, np. przypisujemy jedynki pierwszej składowej wektora $\mathbf{w}_i$, drugiej składowej $\mathbf{w}_{i+1}$, trzeciej składowej $\mathbf{w}_{i+2}$, $k-i+1$ -szej składowej

wej $\mathbf{w}_k$, $k-i+2$ -giej składowej wektora $\mathbf{w}_1$ i tak dalej, aż zapełnimy wszystkie elementy przekątniowe całego zestawu k wektorów wagowych. Wyboru i można dokonać losowo.

3. Przekształcamy wektory wagowe $\mathbf{w}_i$ w macierze wagowe $\mathbf{W}_i$, wstawiając kwadraty składowych wektora wagowego na przekątną macierzy wagowej.

4. Obliczamy dla każdej klasy macierz kowariancji $\mathbf{S}_i^*$, $i = \overline{1, k}$:

$$\mathbf{S}_i^* = \frac{1}{\mathbf{1}^T \mathbf{W}_i \mathbf{1}} \mathbf{X}_c^T \mathbf{W}_i \mathbf{X}_c - \frac{1}{(\mathbf{1}^T \mathbf{W}_i \mathbf{1})^2} \mathbf{X}_c^T \mathbf{1} \mathbf{W}_i \mathbf{1}^T \mathbf{X}_c.$$

5. Obliczamy najmniejszą wartość własną i odpowiadający jej wektor własny $\mathbf{v}_i$ macierzy $\mathbf{S}_i^*$, $i = \overline{1, k}$. Wyznaczamy tak wektor własny, by jego pierwsza współrzędna była równa 1:

$$\mathbf{a}_{ci}^T = \left(1 \quad \frac{v_{2i}}{v_{1i}} \quad \frac{v_{3i}}{v_{1i}} \quad \dots \quad \frac{v_{mi}}{v_{1i}} \right),$$

oraz

$$c_i = -\frac{\mathbf{1}^T \mathbf{W}_i \mathbf{X}_c \mathbf{a}_{ci}}{\mathbf{1}^T \mathbf{W}_i \mathbf{1}}.$$

6. Mając obliczone parametry regresji ortogonalnej w klasach, wyznaczamy dla każdej obserwacji i wektor jej odległości od każdej regresji ($i = \overline{1, k}$, $j = \overline{1, n}$):

$$\mathbf{d}_j^T = (d_{j1} \quad d_{j2} \quad \dots \quad d_{jk}),$$

gdzie

$$d_{ji} = \frac{(\mathbf{a}_{ci}^T \mathbf{x}_j + c_i)^2}{\mathbf{a}_{ci}^T \mathbf{a}_{ci}},$$

i

$$\mathbf{x}_j^T = (y_j \quad x_{1j} \quad \dots \quad x_{mj}).$$

7. W zależności od sposobu klasyfikacji zmieniamy wagi w wektorach wag $\mathbf{w}_i$:

- W metodzie minimum sumy odległości od regresji j -tą składową wektora $\mathbf{w}_i$ zmieniamy na 1, gdy i -ta składowa wektora $\mathbf{d}_j$ osiągnęła wartość najmniejszą.
- W metodzie klasyfikacji rozmytej nowe składowe wektora $\mathbf{w}_i$ obliczamy z wzoru

$$w_{ij} = \frac{\frac{1}{d_{ji}}}{\sum_{i=1}^k \frac{1}{d_{ji}}}.$$

8. Obliczenia kończymy, gdy

$$\max_{j,i} |w_{ij}^{(n)} - w_{ij}^{(s)}| < \varepsilon$$

dla dowolnie dobranego małego $\varepsilon > 0$. Symbole (n) i (s) oznaczają odpowiednio nowe i stare z poprzedniej iteracji wartości wag.

9. Jeżeli punkt 8. nie jest spełniony, wracamy do punktu 3.

W wyniku działania algorytmu otrzymamy poklasyfikowane dane, ale nie ma żadnej gwarancji, że jest to klasyfikacja najlepsza z możliwych. Dla tak wyznaczonej klasyfikacji obliczamy wartość kryterium K z wzoru (4). Jest to podstawa do porównania z nową klasyfikacją, której należy dokonać dla innych wektorów wagowych w_i . Po wielokrotnych próbach klasyfikacji zauważymy, że trudno jest zmniejszyć wartość kryterium K . Jeżeli dla kilku z rzędu prób klasyfikacji nie uda się poprawić (4), to można z dużą wiarygodnością uważać, że znaleźliśmy klasyfikację optymalną bądź też bliską optymalnej.

THE ORTHOGONAL REGRESSION AND CLASSIFICATION

Summary

In this paper the well-known problem of estimating parameters of linear orthogonal regression have been presented. There are proposed any methods of estimating the eigenvectors for the symmetrical and unsymmetrical matrices without calculating its eigenvalues. Additionally there is given an algorithm for classification sample data obtained from no homogenous population in which subpopulations are stepping out with diverse equations of the orthogonal regression.