V-3-3. マハラノビスの距離

多変量解析では、しばしば、データの類似性が議論されます。一般的には、類似性は多次元空間における距離によって評価されます。したがって、２つのデータの類似性が高い場合、それらのデータは多次元空間で近い場所にプロットされます。しかし、測定項目間に相関があったり分散が大きく違っている場合には、それらを考慮しなければ、距離をそのまま類似性だと考えることはできません。

二つのデータ $\boldsymbol{a}$ と $\boldsymbol{b}$を次のようにベクトルで表します。

$$\boldsymbol{a}=\begin{pmatrix}a_1\\a_2\\a_3\end{pmatrix}、\ \mathbf{b}=\begin{pmatrix}b_1\\b_2\\b_3\end{pmatrix}$$

このように表現すると、このデータ間の距離はベクトルの矢印の先端間の距離 $|\mathbf{a}-\mathbf{b}|$となります。

$$\boldsymbol{a}-\boldsymbol{b}=\begin{pmatrix}a_1-b_1\\a_2-b_2\\a_3-b_3\end{pmatrix}$$

$$|\mathbf{a}-\mathbf{b}|=\sqrt{(\boldsymbol{a}-\boldsymbol{b})^T(\boldsymbol{a}-\boldsymbol{b})}$$

$$=\sqrt{(a_1-b_1\quad a_2-b_2\quad a_3-b_3)\begin{pmatrix}a_1-b_1\\a_2-b_2\\a_3-b_3\end{pmatrix}}$$

$$\sqrt{(a_1-b_1)^2+(a_2-b_2)^2+(a_3-b_3)^2}$$

これらのベクトルが行列$\boldsymbol{A}$によって変換され$\boldsymbol{Aa}$、$\boldsymbol{Ab}$ になったとすると、変換されデータ間の距離は次のようになります。

$$|\boldsymbol{Aa}-\boldsymbol{Ab}|=|\boldsymbol{A}(\boldsymbol{a}-\boldsymbol{b})|$$

$$=\sqrt{\big(\boldsymbol{A}(\boldsymbol{a}-\boldsymbol{b})\big)^T\boldsymbol{A}(\boldsymbol{a}-\boldsymbol{b})}$$

$$\sqrt{(a_1-b_1\quad a_2-b_2\quad a_3-b_3)\boldsymbol{A}^T\boldsymbol{A}\begin{pmatrix}a_1-b_1\\a_2-b_2\\a_3-b_3\end{pmatrix}}$$

$$|\mathbf{A}(\mathbf{a}-\mathbf{b})|^2=(a_1-b_1\quad a_2-b_2\quad a_3-b_3)\boldsymbol{A}^T\boldsymbol{A}\begin{pmatrix}a_1-b_1\\a_2-b_2\\a_3-b_3\end{pmatrix}$$

$$|(\mathbf{a}-\mathbf{b})\mathbf{A}|^2=(\mathbf{a}-\mathbf{b})^T\boldsymbol{A}^T\boldsymbol{A}(\mathbf{a}-\mathbf{b})$$

ここで $\mathbf{a}-\mathbf{b}=\boldsymbol{x}$と表すことにします。

$$\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}$$

ここで、$\boldsymbol{A}$が分散共分散行列の平方根であったとします。

$$\boldsymbol{A}=\boldsymbol{\Sigma}^{-\frac{1}{2}}$$

$\boldsymbol{\Sigma}$: 分散共分散行列

$$|\boldsymbol{Ax}| = \sqrt{\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{x}\right)^{T}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{x}}$$

$$= \sqrt{\boldsymbol{x}^{T}\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\right)^{T}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{x}} = \sqrt{\boldsymbol{x}^{T}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{x}} = \sqrt{\boldsymbol{x}^{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}$$

($\because \boldsymbol{\Sigma}^{-\frac{1}{2}}$ は対称行列ですから. $\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\right)^{T} = \boldsymbol{\Sigma}^{-\frac{1}{2}}$)

これをマハラノビスの距離と言います。

$$D_{\boldsymbol{a}-\boldsymbol{b}} = \sqrt{(\boldsymbol{a}-\boldsymbol{b})^{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{a}-\boldsymbol{b})}$$

($\boldsymbol{\Sigma}$: 分散共分散行列)

式 70

マハラノビスの距離は分散共分散による歪みを補正した標準化された距離です。項目間に相関があったり、項目間の分散違いが大きい場合に、そのまま距離を計算する（ユークリッド距離）ことが適切でないと判断して、マハラノビス距離を計算することは、クラスター分析等でしばしば行われます。

Prasanta Chandra　Mahalanobis はインドの統計学者で、インド統計研究所を創設しました。Fisher や Pearson とも交友があり、インドの社会主義的な計画経済の政策立案にも関与し、理論とともに実践面でも活躍しました。