Chap. 12

本レジュメは(12.9)式の導出を目指すものである。教科書とは異なる表現※を用いるため、まず各種前提から整理させてもらう。

※ i番目のデータをベクトル $\boldsymbol{x}_i$ で表現します。教科書では x_{ik} のようにベクトル成分を露わに書いていますが、扱いづらくなるため、あえてベクトル表示を使います。

2つのクラスターA, Bを考える。各クラスターに所属するi番目のデータは

Aに所属する場合; $\boldsymbol{x}_i$ $(i \in A)$, データ数は n_A コ

Bに所属する場合; $\boldsymbol{x}_i$ $(i \in B)$, データ数は n_B コ

のように表現できる。今2つのクラスターA, Bを併合して、新しいクラスターを作る、この新しいクラスター $A \cup B$ に所属するデータは

$\boldsymbol{x}_i$ $(i \in A \cup B)$

と書くことができる。

各クラスター内における平方和の def. は

$$A;\ \sum_{i \in A} \left| \boldsymbol{x}_i - \sum_{j \in A} \frac{\boldsymbol{x}_j}{n_A} \right|^2$$

$$B;\ \sum_{i \in B} \left| \boldsymbol{x}_i - \sum_{j \in B} \frac{\boldsymbol{x}_j}{n_B} \right|^2$$

$$A \cup B;\ \sum_{i \in A \cup B} \left| \boldsymbol{x}_i - \sum_{j \in A \cup B} \frac{\boldsymbol{x}_j}{n_A + n_B} \right|^2$$

となる。いずれの平方和も各クラスター内のデータの平均からの偏差平方和として約束されている。

ここで

$$\Delta = \sum_{i \in A \cup B} \left| \boldsymbol{x}_i - \sum_{j \in A \cup B} \frac{\boldsymbol{x}_j}{n_A + n_B} \right|^2 - \sum_{i \in A} \left| \boldsymbol{x}_i - \sum_{j \in A} \frac{\boldsymbol{x}_j}{n_A} \right|^2 - \sum_{i \in B} \left| \boldsymbol{x}_i - \sum_{j \in B} \frac{\boldsymbol{x}_j}{n_B} \right|^2$$

のように Δ を定義する。Δ は merging cost と呼ばれ、この Δ を min 化するようなクラスタリング方法を「ウォード法 (Ward Method)」と呼ぶ。

実は Δ は concise に表現することが可能で

$$\Delta = \frac{n_A n_B}{n_A + n_B} (m_A - m_B)^2 \quad , \text{with} \quad m_A = \sum_{i \in A} \frac{x_i}{n_A} \quad \text{and} \quad m_B = \sum_{i \in B} \frac{x_i}{n_B}$$

となる。以下では Δ の def. から上式を導くことにする。

$$\Delta = \sum_{i \in A \cup B} \left\{ |x_i|^2 - 2x_i \cdot \frac{\sum_{j \in A \cup B} x_j}{n_A + n_B} + \left| \frac{\sum_{j \in A \cup B} x_j}{n_A + n_B} \right|^2 \right\}$$

$$- \sum_{i \in A} \left\{ |x_i|^2 - 2x_i \cdot \frac{\sum_{j \in A} x_j}{n_A} + \left| \frac{\sum_{j \in A} x_j}{n_A} \right|^2 \right\}$$

$$- \sum_{i \in B} \left\{ |x_i|^2 - 2x_i \cdot \frac{\sum_{j \in B} x_j}{n_B} + \left| \frac{\sum_{j \in B} x_j}{n_B} \right|^2 \right\}$$

ここで、以降の計算のポイントを押える。

① $|x_i|^2$ の項はキャンセル

② $\sum_{j \in A \cup B}$ がかかる項は $\sum_{j \in A}$ と $\sum_{j \in B}$ に分離

③ m_A, m_B が現われるように変形

②, ③ を実行すると

$$\frac{\sum_{j \in A \cup B} x_j}{n_A + n_B} = \frac{\sum_{j \in A} x_j + \sum_{j \in B} x_j}{n_A + n_B} = \frac{1}{n_A + n_B} \left(n_A \sum_{j \in A} \frac{x_j}{n_A} + n_B \sum_{j \in B} \frac{x_j}{n_B} \right)$$

$$= \frac{1}{n_A + n_B} (n_A m_A + n_B m_B) \quad \cdots (☆)$$

① と (☆) によって

$$\Delta = -2(n_A + n_B) \sum_{i \in A \cup B} \frac{x_i}{n_A + n_B} \cdot \sum_{j \in A \cup B} \frac{x_j}{n_A + n_B} + (n_A + n_B) \left| \frac{\sum_{j \in A \cup B} x_j}{n_A + n_B} \right|^2$$

$$+ 2n_A \sum_{i \in A} \frac{x_i}{n_A} \cdot \sum_{j \in A} \frac{x_j}{n_A} - n_A \left| \sum_{j \in A} \frac{x_j}{n_A} \right|^2$$

$$+ 2n_B \sum_{i \in B} \frac{x_i}{n_B} \cdot \sum_{j \in B} \frac{x_j}{n_B} - n_B \left| \sum_{j \in B} \frac{x_j}{n_B} \right|^2 \quad (\because ①)$$

$$= -2(n_A + n_B)\frac{1}{n_A + n_B}(n_A m_A + n_B m_B)\cdot\frac{1}{n_A + n_B}(n_A m_A + n_B m_B)$$

$$+ (n_A + n_B)\left|\frac{1}{n_A + n_B}(n_A m_A + n_B m_B)\right|^2$$

$$+ 2n_A m_A \cdot m_A - n_A |m_A|^2 + 2n_B m_B \cdot m_B - n_B |m_B|^2 \quad (\because (\star))$$

$$= -\frac{2}{n_A + n_B}(n_A m_A + n_B m_B)^2 + \frac{1}{n_A + n_B}(n_A m_A + n_B m_B)^2$$

$$+ n_A m_A^2 + n_B m_B^2$$

$$= \frac{-(n_A m_A + n_B m_B)^2 + (n_A + n_B) n_A m_A^2 + (n_A + n_B) n_B m_B^2}{n_A + n_B}$$

$$= \cdots = \frac{n_A n_B}{n_A + n_B}(m_A - m_B)^2$$

$\blacksquare$