

クラスタ分析

非階層的方法と分析の評価

村田 昇

講義の内容

- 第1日：クラスタ分析の考え方と階層的方法
- 第2日：非階層的方法と分析の評価

クラスタ分析の復習

クラスタ分析

- クラスタ分析 (cluster analysis) の目的
個体の間に隠れている**集まり=クラスタ**を個体間の“距離”にもとづいて発見する方法
- 個体間の類似度・距離 (非類似度) を定義
 - 同じクラスタに属する個体どうしは似通った性質
 - 異なるクラスタに属する個体どうしは異なる性質
- さらなるデータ解析やデータの可視化に利用
- 教師なし学習の代表的な手法の一つ

クラスタ分析の考え方

- 階層的方法
 - データ点およびクラスタの間に **距離** を定義
 - 距離に基づいてグループ化
 - * 近いものから順にクラスタを **凝集**
 - * 近いものが同じクラスタに残るように **分割**
- 非階層的方法
 - クラスタの数を事前に指定
 - クラスタの **集まりの良さ** を評価する損失関数を定義
 - 損失関数を最小化するようにクラスタを形成

階層的クラスタリング

- 凝集の手続き
 1. データ・クラスタ間の距離を定義
 - データ点間の距離
 - クラスタ間の距離
 2. データ点およびクラスタ間の距離を計算
 3. 最も近い2つを統合し新たなクラスタを形成
 4. クラスタ数が1つになるまで2-3の手続きを繰り返す

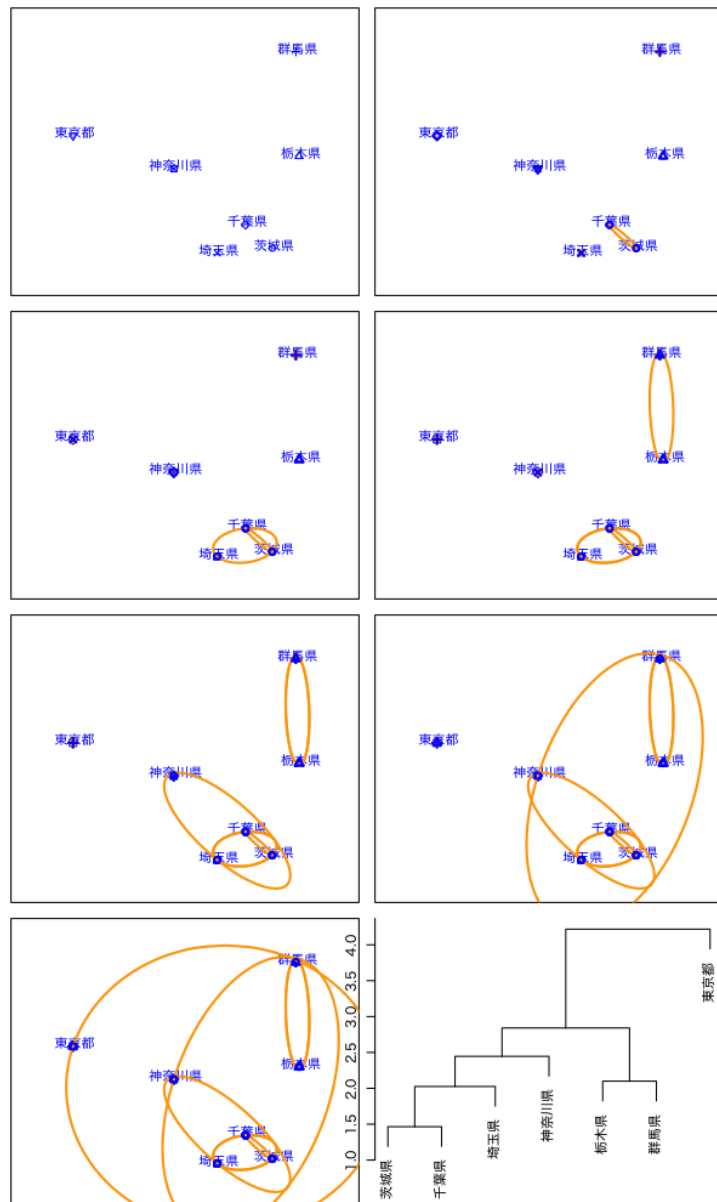


図 1: 凝集的手続きの例

非階層的方法

非階層的方法の手続き

- 対象の変数: $X = (X_1, X_2, \dots, X_d)^T$ (d 次元)
- 観測データ: n 個の個体の組

$$\{\mathbf{x}_i\}_{i=1}^n = \{(x_{i1}, x_{i2}, \dots, x_{id})^T\}_{i=1}^n$$

- 個体とクラスタの対応 C を推定

$C(i)$ = (個体 i が属するクラスタ番号)

- 対応 C の **全体の良さ** を評価する損失関数を設定
- 観測データ $\{\mathbf{x}_i\}_{i=1}^n$ に最適な対応 $\{C(i)\}_{i=1}^n$ を決定

k-平均法の損失関数

- クラスタの個数 k を指定
- 2つの個体 i, i' の **近さ=損失** を距離の二乗で評価

$$\|\mathbf{x}_i - \mathbf{x}_{i'}\|^2 = \sum_{j=1}^d (x_{ij} - x_{i'j})^2$$

- 損失関数 $W(C)$: クラスタ内の平均の近さを評価

$$W(C) = \sum_{l=1}^k \frac{1}{n_l} \sum_{i: C(i)=l} \sum_{i': C(i')=l} \|\mathbf{x}_i - \mathbf{x}_{i'}\|^2$$

k-平均法の性質

- クラスタ l に属する個体の平均

$$\bar{\mathbf{x}}_l = \frac{1}{n_l} \sum_{i: C(i)=l} \mathbf{x}_i$$

- 損失関数 $W(C)$ の等価な表現

$$W(C) = 2 \sum_{l=1}^k \sum_{i: C(i)=l} \|\mathbf{x}_i - \bar{\mathbf{x}}_l\|^2$$

- 最適な対応 C : クラスタ内変動の総和が最小

演習

問題

- 以下の問に答えなさい
 - 損失関数 $W(C)$ の等価な表現を示しなさい

$$\begin{aligned} W(C) &= \sum_{l=1}^k \frac{1}{n_l} \sum_{i: C(i)=l} \sum_{i': C(i')=l} \|x_i - x_{i'}\|^2 \\ &= 2 \sum_{l=1}^k \sum_{i: C(i)=l} \|x_i - \bar{x}_l\|^2 \end{aligned}$$

- 以下の $\hat{\mu}$ を求めなさい

$$\hat{\mu} = \arg \min_{\mu} \sum_{i: C(i)=l} \|x_i - \mu\|^2$$

近似的な最適化

クラスタ対応の最適化

- 最適化: 損失関数 $W(C)$ を最小とする C を決定
- 貪欲な C の探索
 - 原理的には全ての値を計算すればよい
 - 可能な C の数: k^n 通り (有限個のパターン)
 - サンプル数 n が小さくない限り実時間での実行は不可能
- 近似的な C の探索
 - いくつかのアルゴリズムが提案されている
 - 基本的な考え方: **Lloyd-Forgy のアルゴリズム**
標本平均と変動の平方和の性質を利用

$$\bar{x}_l = \arg \min_{\mu} \sum_{i: C(i)=l} \|x_i - \mu\|^2 \quad (\text{クラスタ } l \text{ の標本平均})$$

Lloyd-Forgy のアルゴリズム

1. クラスタ中心の初期値 $\mu_1, \mu_2, \dots, \mu_k$ を与える
2. 各データの所属クラスタ番号 $C(i)$ を求める

$$C(i) = \arg \min_l \|x_i - \mu_l\|$$

3. 各クラスタ中心 μ_l ($l = 1, 2, \dots, k$) を更新する

$$\mu_l = \frac{1}{n_l} \sum_{i: C(i)=l} x_i, \quad n_l = |\{x_i | C(i) = l\}|$$

4. 中心が変化しなくなるまで 2, 3 を繰り返す

アルゴリズムの性質

- 結果は **確率的**
 - 初期値 $\mu_1, \mu_2, \dots, \mu_k$ に依存
 - アルゴリズムの成否は確率的
(最適解が得られない場合もある)
- 一般には複数の初期値をランダムに試して損失を最小とする解を採用する
- 平均の代わりにメドイド (medoid; 中心にある観測値) を用いる方法もある

$$x_l^{\text{medoid}} = \arg \min_{x_i} \sum_{i': C(i')=l} \|x_i - x_{i'}\|^2$$

事例

- 都道府県別好きなおむすびの具 (一部) での例

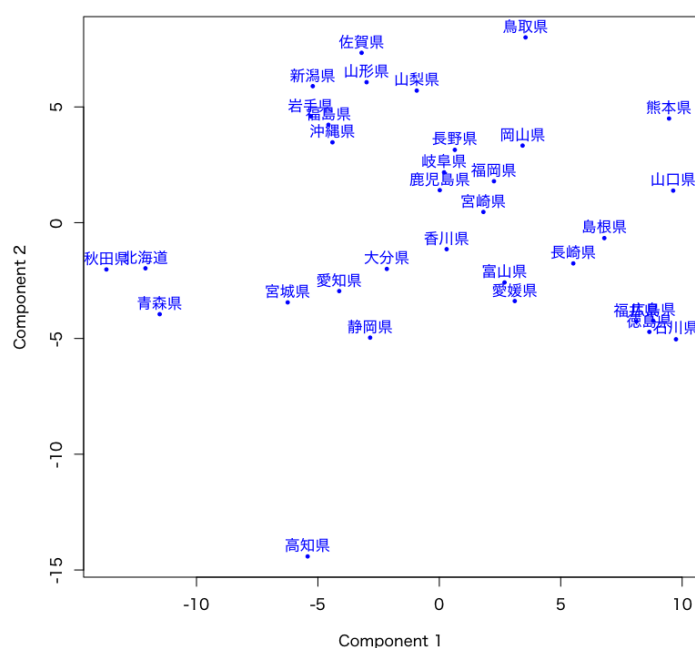


図 2: 非階層的クラスタリング

解析事例

都道府県別の社会生活統計指標

- データの属性

Forest : 森林面積割合 (%) 2014 年

Agri : 就業者 1 人当たり農業産出額 (販売農家) (万円) 2014 年

Ratio : 全国総人口に占める人口割合 (%) 2015 年

Land : 土地生産性 (耕地面積 1 ヘクタール当たり) (万円) 2014 年

Goods : 商業年間商品販売額 [卸売業 + 小売業] (事業所当たり) (百万円) 2013 年

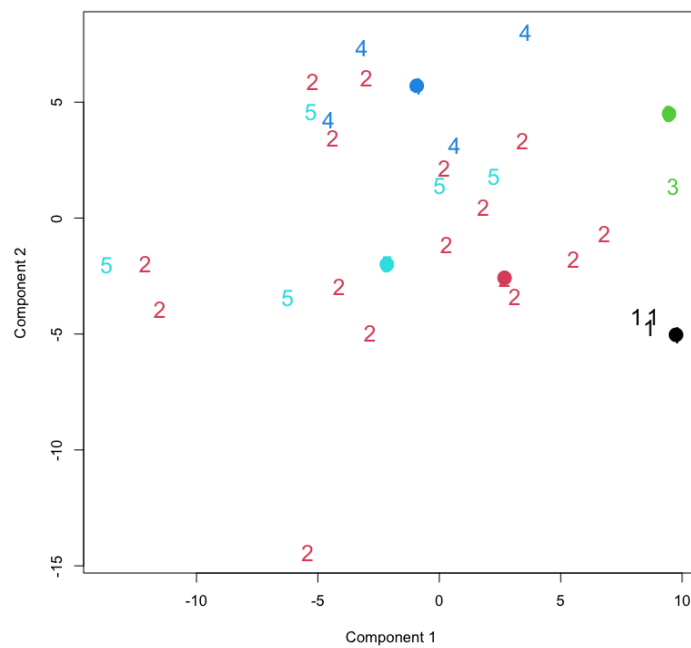


図 3: Lloyd-Forgy のアルゴリズム (その 1)

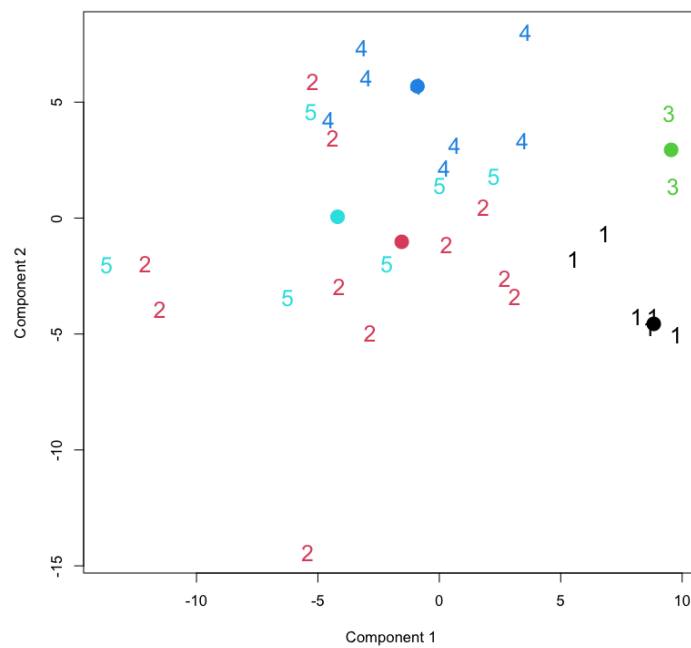


図 4: Lloyd-Forgy のアルゴリズム (その 2)

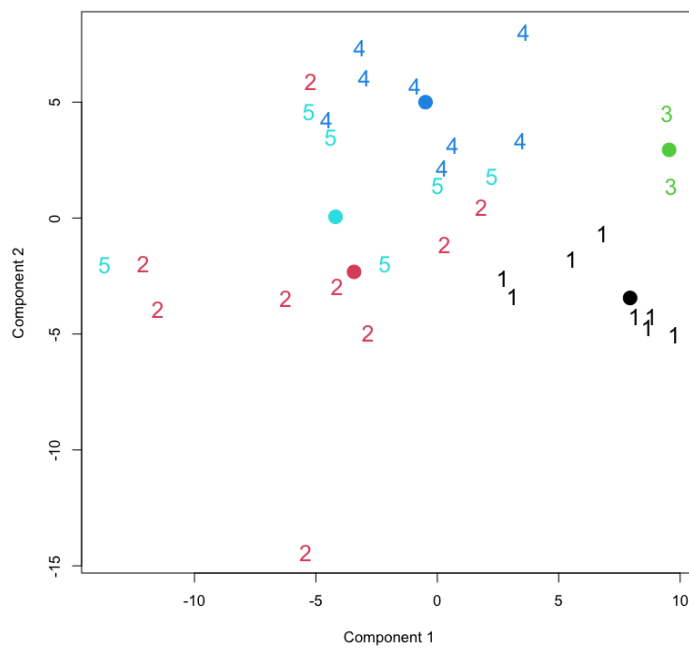


図 5: Lloyd-Forgy のアルゴリズム (その 3)

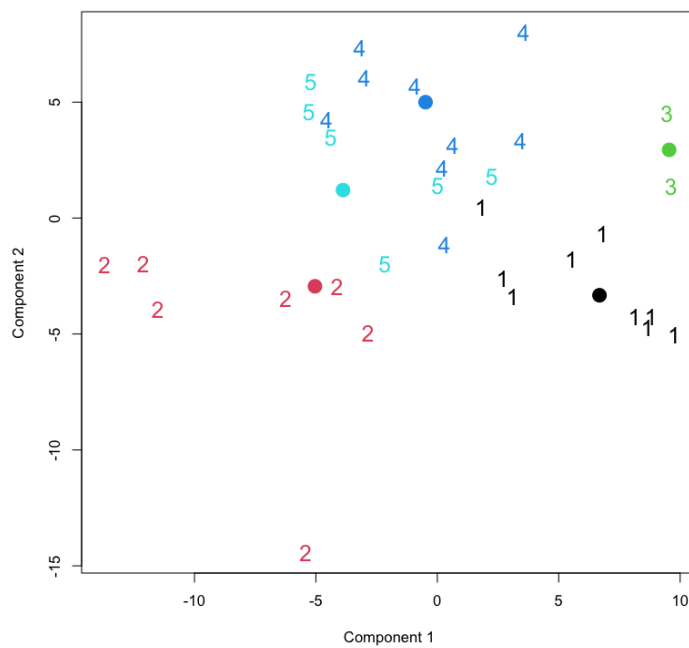


図 6: Lloyd-Forgy のアルゴリズム (その 4)

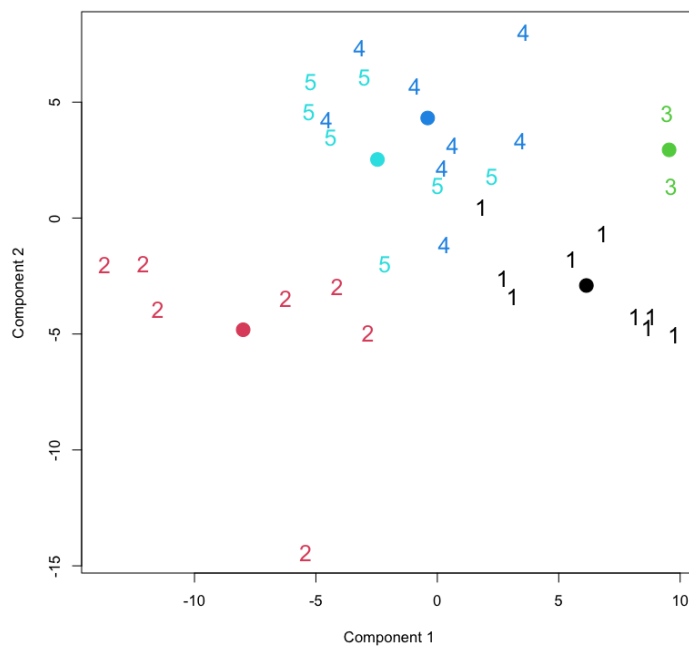


図 7: Lloyd-Forgy のアルゴリズム (その 5)

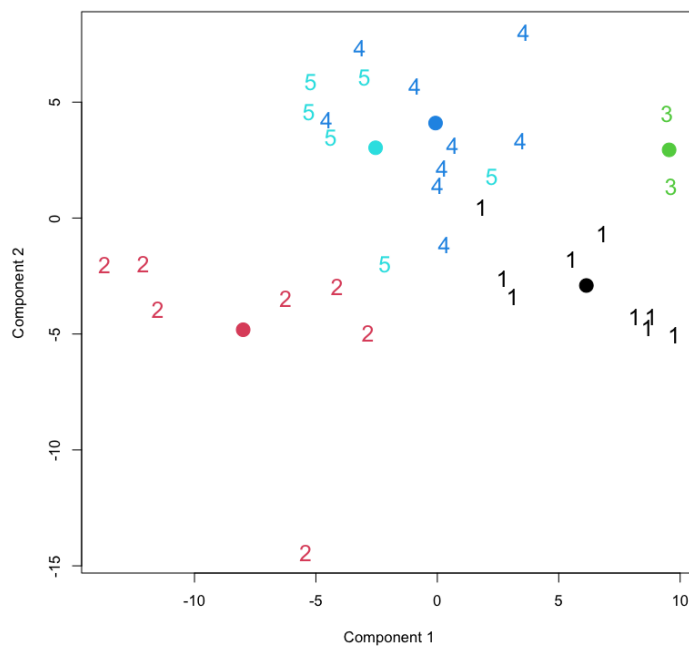


図 8: Lloyd-Forgy のアルゴリズム (その 6)

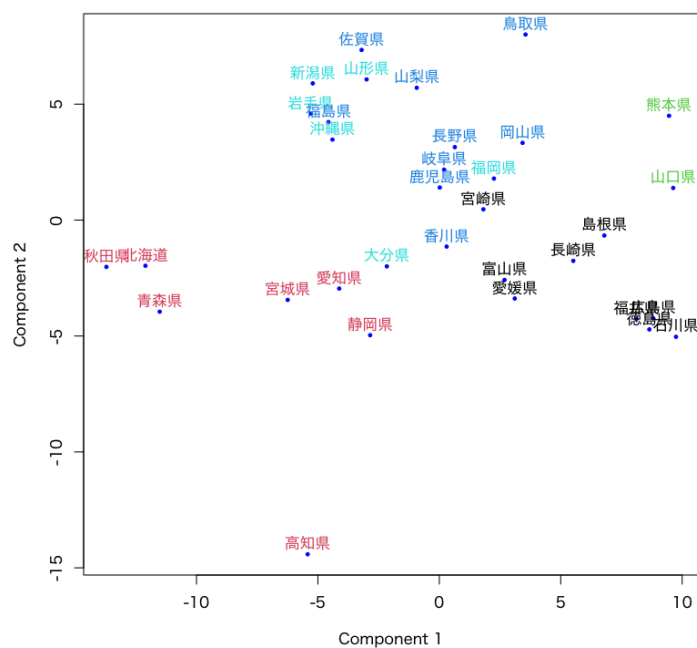


図 9: クラスターリングの結果

- 平均 0, 分散 1 に正規化して解析
- ユークリッド距離 + k-平均法
- ユークリッド距離 + k-メドイド法

都道府県別好きなおむすびの具

- データの属性

Q2. おむすびの具では何が一番好きですか？

A. 梅 B. 鮭 C. 昆布 D. かつお E. 明太子 F. たらこ G. ツナ H. その他

【回答者数】

男性 9,702 人 32.0%

女性 20,616 人 68.0%

総数 30,318 人 100.0%

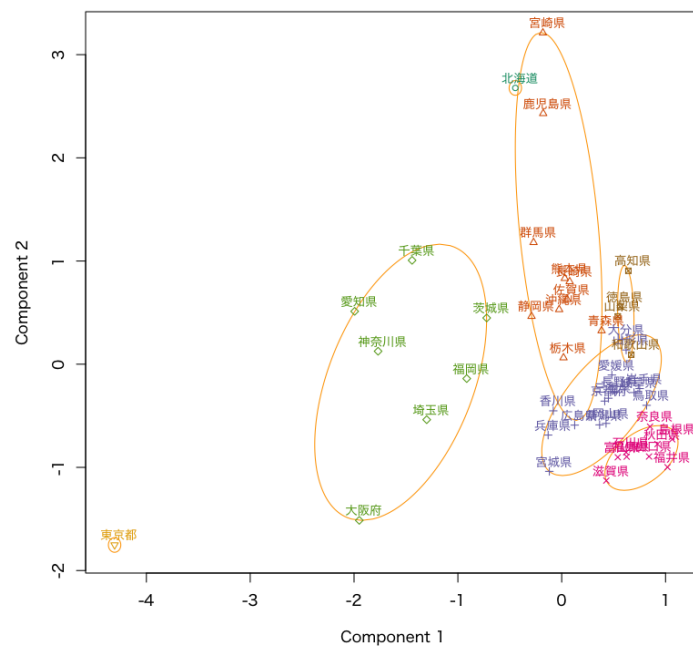
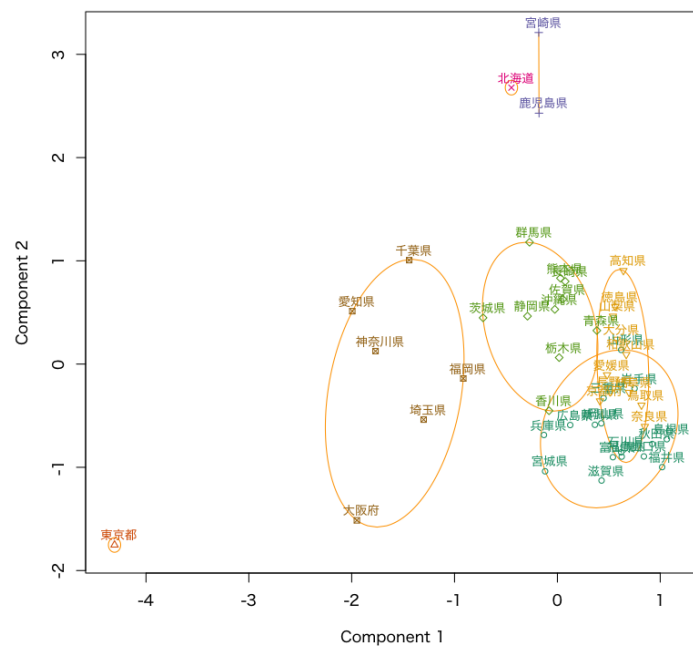
に対する回答を県別に集計

- Hellinger 距離を利用

p, q を確率ベクトルとして定義される確率分布の距離

$$d_{\text{hel}}(p, q) = \frac{1}{\sqrt{2}} d_{\text{euc}}(\sqrt{p}, \sqrt{q})$$

- Hellinger 距離 + k-メドイド法



非階層的方法の評価

- 評価の対象
 - x_i を含むクラスタ C^1 と x_i の距離

$$d_i^1 = D(x_i, C^1 \setminus x_i)$$

- 一番近いクラスタ C^2 と x_i の距離

$$d_i^2 = D(x_i, C^2)$$

- シルエット係数 (silhouette coefficient)

$$S_i = \frac{d_i^2 - d_i^1}{\max(d_i^1, d_i^2)}$$

シルエット係数の性質

- 定義より
 - $-1 \leq S_i \leq 1$
 - 1 に近いほど適切なクラスタリング
- 全体の良さを評価するには S_i の平均を用いる
- 距離の計算を適切に行えば階層的方法でも利用可

演習

問題

- 以下の問に答えなさい
 - 群平均法において凝集係数が以下を満たすことを示しなさい

$$0 \leq AC \leq 1$$

- シルエット係数が以下を満たすことを示しなさい

$$-1 \leq S_i \leq 1$$

次回の予定

- 第1日：時系列の基本モデル
- 第2日：モデルの推定と予測