# クラスター分析

# クラスター分析 …

以たものを集める

クラスター分析 … 行方向にグルーピング

主成分、因子 … 列方向にグルーピング

因子分析

| id | use | design | price |
|---|---|---|---|
| 1 | 3 | 2 | 3 |
| 2 | 3 | 5 | 1 |
| 3 | 4 | 1 | 4 |
| ⋮ | ⋮ | ⋮ | ⋮ |

$a_1 f_1 + a_2 f_2 + \varepsilon$

クラスター分析

| id | use | design | price | |
|---|---|---|---|---|
| 1 | 3 | 2 | 3 | $x_1$ |
| 2 | 3 | 5 | 1 | |
| 3 | 4 | 1 | 4 | $x_3$ |
| ⋮ | ⋮ | ⋮ | ⋮ | |

| 都道府県 | 人口 | 面積 | 人口密度 |
|---|---|---|---|
| 東京 | ; | ; | ; |
| 大阪 | ; | ; | ; |
| 北海道 | ; | ; | ; |

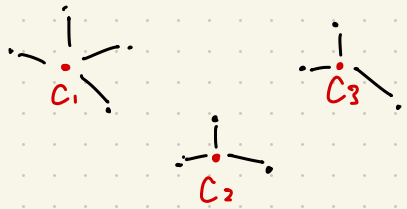
主成分
… 発展度, 山地.
とかでデータ説明したい

ある量 $L$ を最小にする

centroid
クラスターの重心
計算で出る

$$L(C_1, \cdots C_k) = \sum_{k}^{k} \sum_{x \in C_k} \| x - C_k \|_2^2$$

図にする




・3クラスター … $k = 3$

・$C_k$ とデータ点の距離をはかる.

ex. 東京 … $C_1$ クラスター

$$d = \sqrt{(\text{東京人口} - C_1\text{人口})^2 + (\text{東京面} - C_1\text{面})^2 + (\text{東京密} - C_1\text{密})^2}$$

※ ユークリッド距離

$d = \| x - C_k \|_2$ … 小さいほど, 距離が近い

・$L(C_1, \cdots C_k)$ は各データとそのクラスターの重心との距離を合計したもの

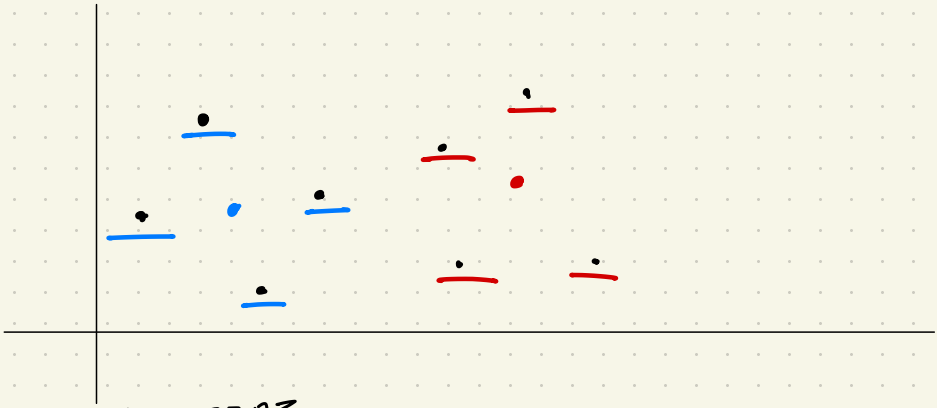
・$L(C_1, \cdots C_k)$ が最小になるクラスター分けが良さそう.

・厳密に計算するのは難しい, なるべく小さくなるように

$L(c_1 \cdots c_k)$ を小さくしたい．

k-mean 法

1. クラスター所属をランダムに決める
2. 各クラスターの centroid が決まる → 平均
3. 各データについて，最も近い centroid を見つけ，クラスターを再設定 （$\|x - c_k\|$）
4. 2〜3 を繰り返す （→ iter. max）
5. 再設定が最もなくなると終る．



k-mean 法の問題

→ 初期値がランダムなので，おかしなことに!!

もうちょっとマシに!!

k-mean ++ … クラスターの重心って離れてるほうがいい!

1. 1つめの重心をデータ点の中からランダムに決める … $g_1$

2. 2つめの重心は1つめの重心から遠いほど選ばれやすくする.

   ・各データ点と $g$ のキョリ : $\|x_i - g\|$

   ・選ばれる確率 $P(x) = \dfrac{(\|x_i - g\|)^2}{\sum(\|x_i - g\|)^2}$ … $g_2$

3. 3つめの重心は, 各データ点から $g_1$, $g_2$ の近い方を採用
   選ばれる確率を計算 … $g_1$ からも $g_2$ からも遠い点が選ばれる

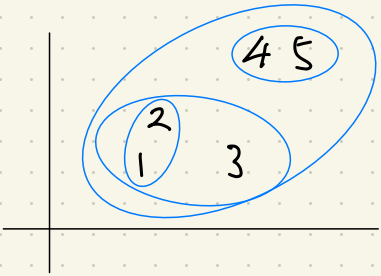
4. 3を繰り返す.

---

真の最小値 $L_{opt}$ として, 得られる $L_{++}$ とした

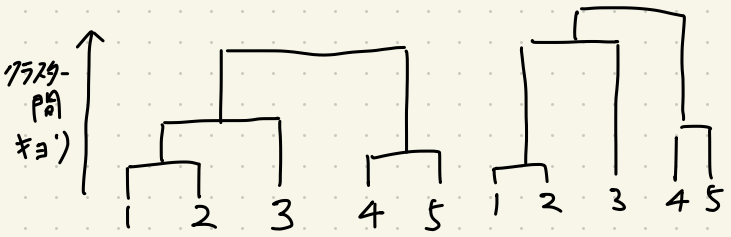
$$E(L_{++}) \leq 8(\log k + 2) L_{opt}$$

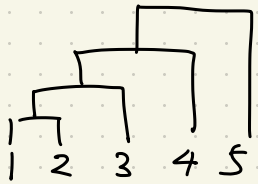
# 階層クラスター分析




○で囲われてるのがクラスター
あるクラスターが、大きなクラスターに含まれる.

失敗パターン



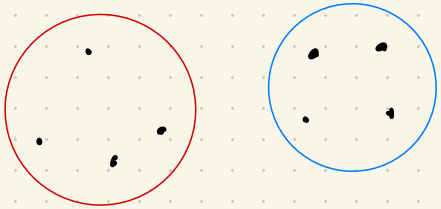




チェイニング

計算の順番

1. 各データ点を1つのクラスターと考える
2. 全てのクラスター間のキョリを計算する.
3. 以下を繰り返す
   a. 一番近いクラスターを併合
   b. データ点が1つのクラスターにまとまったら計算終了
   c. 新しく作ったクラスターと既存のクラスターのキョリを計算.

9. クラスター間のキョリ



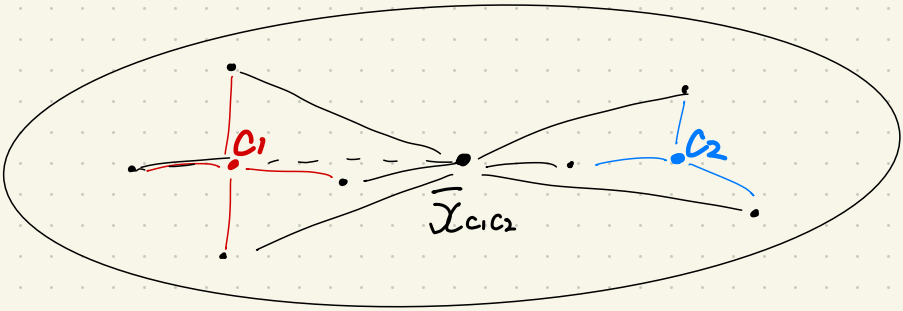
- 最短キョリ法
- 最長キョリ法
- 重心法
- Ward法

# ward 法

- クラスター $C_1$, $C_2$ 合併すると仮定.
- 合併後のクラスター内の重心からの誤りの2乗和を計算 … ①
- 合併前のクラスター内の重心からの誤りの2乗和を計算 … ②
- ① − ② が最小になるクラスターを合併する.

⇓

クラスター内のバラつきを最小にするクラスターを合併する.




$\sum \|x_i - \bar{x}_{C_1}\|^2$ $\sum \|x_i - \bar{x}_{C_2}\|^2$ before

after $\sum \|x_i - \bar{x}_{C_1C_2}\|^2$

$\sum \|x_i - \bar{x}_{C_1C_2}\|^2 - \sum \|x_i - \bar{x}_{C_1}\|^2 - \sum \|x_i - \bar{x}_{C_2}\|^2$

⇒ クラスター間の誤り $d(C_1, C_2)$

---

重心法の欠点

単調.