# 5章　k-means

**5月9日（金）**

**豊川　幸秀**

# k-meansとは？

- 非階層的手法の一種。
  - ー階層的手法と対極にある手法。
  - ーあらかじめクラスタ数を固定しておいてから分類を行う

# アルゴリズム（1）

入力：分割後のクラスタ数K及びデータセット

出力：以下の評価関数を最小化するデータの
クラスタへの割り当て

$$\sum_{i=1}^{K} \sum_{x \in C_i} \| x - c_i \|^2$$

x：データ　$c_i$：クラスタiにおける重心

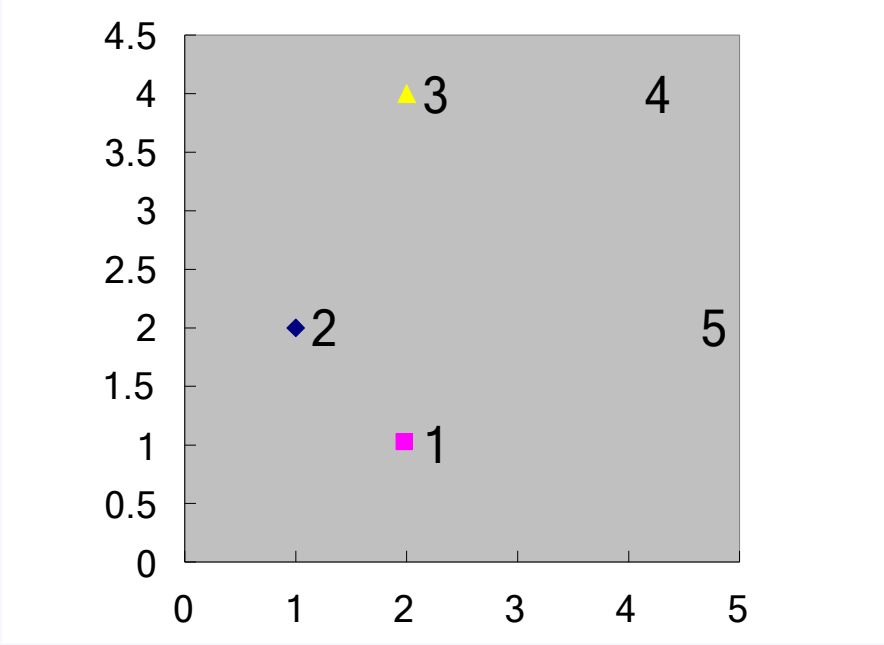
# アルゴリズム（２）

①目標クラスタ数であるK個のクラスタに対する代表点$c_1, c_2, \cdots, c_k$を作成

②各xに対し、各代表点との距離を測り、最も距離の近い$c_i$のクラスタをxのクラスタに設定

③②において各xのクラスタが変化しなければ終了。変化した場合、代表点$c_1, c_2, \cdots, c_k$を各クラスタの重心に更新し、②に戻る。

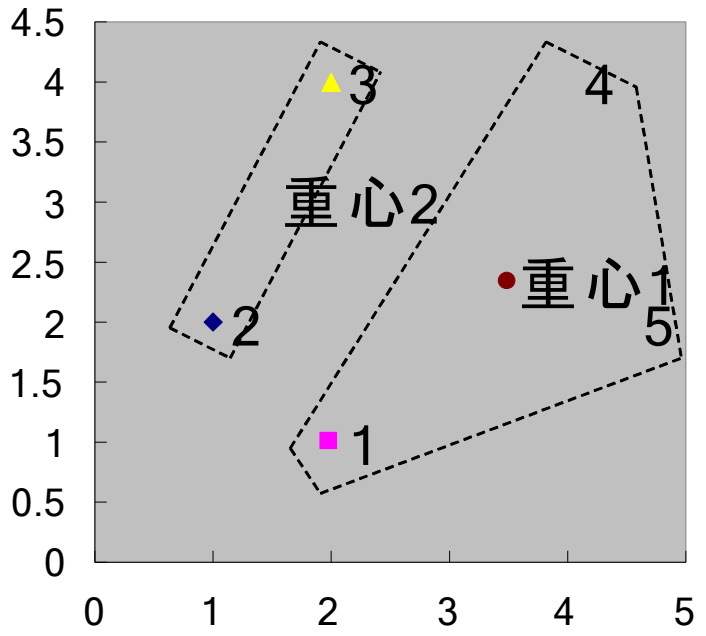
# 数値例(1)－1




| データ | 重心1 | 重心2 | クラスタ |
|---|---|---|---|
| データ1 | 0 | 1.4142 | 1 |
| データ2 | 1.4142 | 0 | 2 |
| データ3 | 3.0000 | 2.2361 | 2 |
| データ4 | 3.6056 | 3.6056 | 1 |
| データ5 | 2.6926 | 3.5000 | 1 |

重心1:1　重心2:2

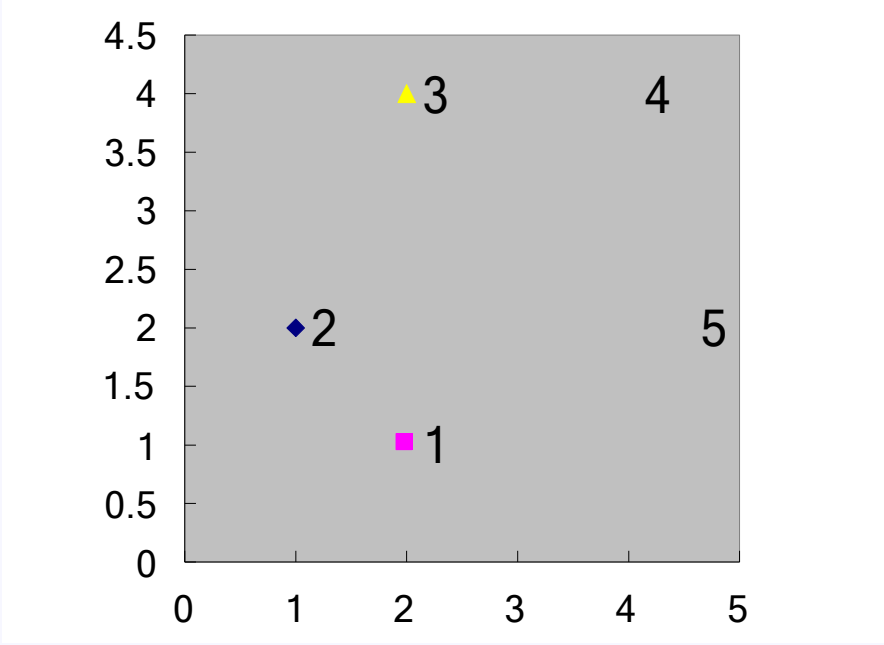
# 数値例(1)－2




| データ | 重心1 | 重心2 | クラスタ |
|---|---|---|---|
| データ1 | 2.0069 | 2.0616 | 1 |
| データ2 | 2.5221 | 1.1180 | 2 |
| データ3 | 2.2423 | 1.1180 | 2 |
| データ4 | 1.7401 | 2.6926 | 1 |
| データ5 | 1.0541 | 3.1623 | 1 |

クラスタの割り当てに変化がないため、ここで終了。

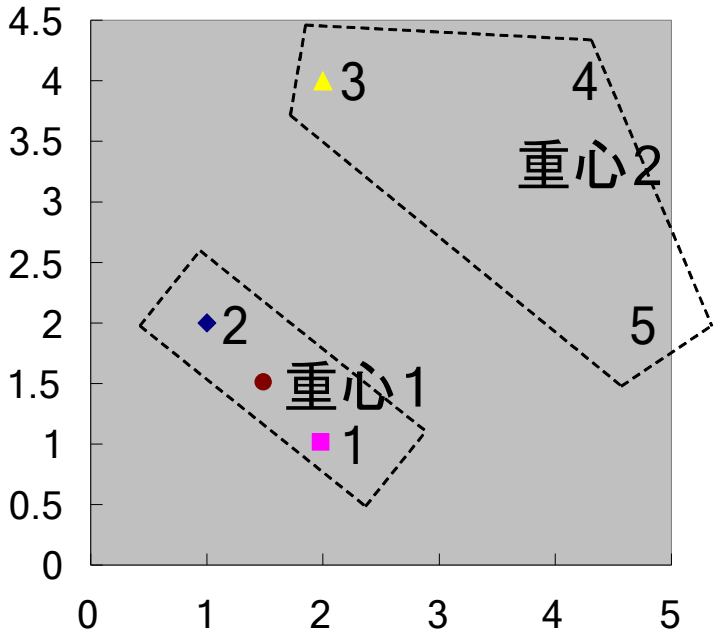
# 数値例(2)－1




| データ | 重心1 | 重心2 | クラスタ |
|---|---|---|---|
| データ1 | 0 | 3.6056 | 1 |
| データ2 | 1.4142 | 3.6056 | 1 |
| データ3 | 3.0000 | 2.0000 | 2 |
| データ4 | 3.6056 | 0 | 2 |
| データ5 | 2.6926 | 2.0616 | 2 |

重心1：1　重心2：4

# 数値例(2)－2




| データ | 重心1 | 重心2 | クラスタ |
|---|---|---|---|
| データ1 | 0.7071 | 2.7739 | 1 |
| データ2 | 0.7071 | 2.8333 | 1 |
| データ3 | 2.5495 | 1.6415 | 2 |
| データ4 | 3.5355 | 0.8333 | 2 |
| データ5 | 3.0414 | 1.6667 | 2 |

クラスタの割り当てに変化がないため、ここで終了。

# 評価関数による考察

- （1）の場合
  $2.0069^2+1.1180^2+1.1180^2+1.7401^2$
  $+1.0541^2 = 10.6667$
- （2）の場合
  $0.7071^2+0.7071^2+1.6415^2+0.8333^2$
  $+1.6667^2 = 7.1667$

→（2）のほうがより最適なクラスタリング

# 関数kmeans（2）

```
kmeans(x, centers, iter.max = 10, nstart = 1,
            algorithm = c("Hartigan-Wong", "Lloyd",
            "Forgy", "MacQueen"))
```

第1引数：データセット

第2引数：クラスタの代表点からなるベクトル

or 代表点の個数

# 関数kmeans（２）

以下省略可

第３引数：繰り返しの最大回数（デフォルトは１０回）

第４引数：ランダムに初期値を設定する場合の
パラメータ

第５引数：利用するアルゴリズム
（デフォルトは"Hartigan-Wong"）

以上です