k−平均法
教師なし学習モデルの基礎

13T4054L　根岸　睦

目次

- k−平均法とは
- k−平均法によるクラスタリングの例
- 画像データへの応用
- k−平均法の数学的根拠
- 怠惰学習モデルとしてのk−近傍法

k−平均法とは

- クラスタリング手法のひとつ
- 教師なし学習である
- データを適当に分けて、それからより上手く分かれるように調節する

k−平均法によるクラスタリングの例

・トレーニングセットとして、(x,y)平面上 の多数の点

$(x_n, y_n)_{n=1}^{N}$　　が与えられたとする

・トレーニングセットに含まれるデータは

$\mathbf{x}_n = (x_n, y_n)^T$ と表記する

k−平均法を用いてこれらのデータを2つのクラスタに分ける

k−平均法によるクラスタリングの例

1.

・代表点　$\mu_k{}_{k=1}^{2}$ をランダムに決定する

・各点と代表点との距離　$|\mathbf{x}_n - \mu_k|$ を計算して距離が短い方の代表点に属するものとする

・代表点に属するかどうかを示す変数を定義

$$r_{nk} = \begin{cases} 1 & \mathbf{x}_n \text{が} k \text{番目の代表点に属する場合} \\ 0 & \text{それ以外の場合} \end{cases} \tag{6.1}$$

k−平均法によるクラスタリングの例

2.

・1で分けられたクラスターの重心を求める

$$\mu_k = \frac{\sum \mathbf{x}_n}{N_k}(k = 1, 2) \tag{6.2}$$

(6.1)より

$$\mu_k = \frac{\sum_{n=1}^{N} r_{nk}\mathbf{x}_n}{\sum_{n=1}^{N} r_{nk}} \tag{6.3}$$

k−平均法によるクラスタリングの例

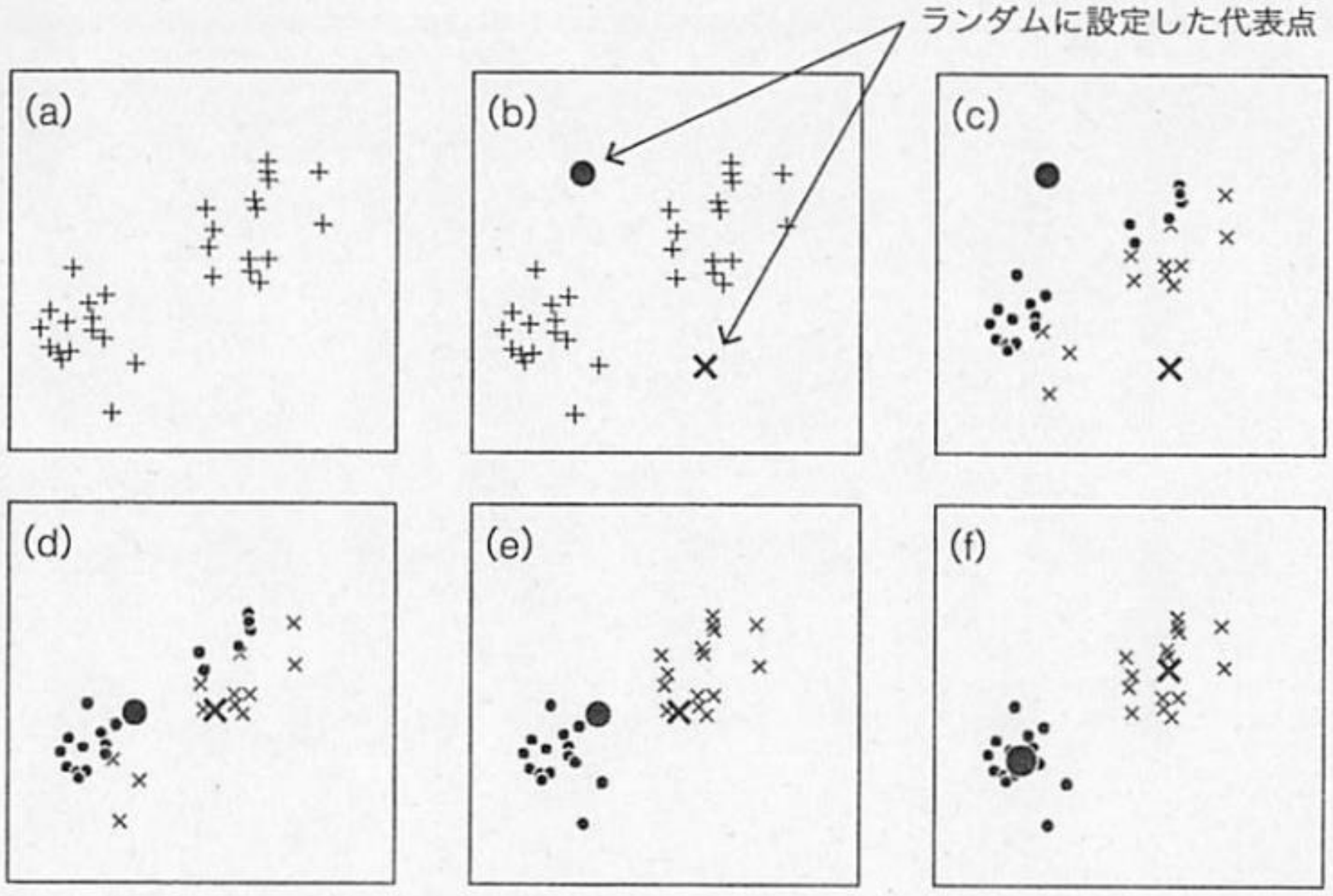
3.

・求まった重心を代表点としてまた新たに重心を求める

4.

・重心が変化しなくなったら終了

・最終的に得られた代表点が各クラスターの代表となる



▲図6.1 k平均法によるクラスタリングの例

画像データへの応用

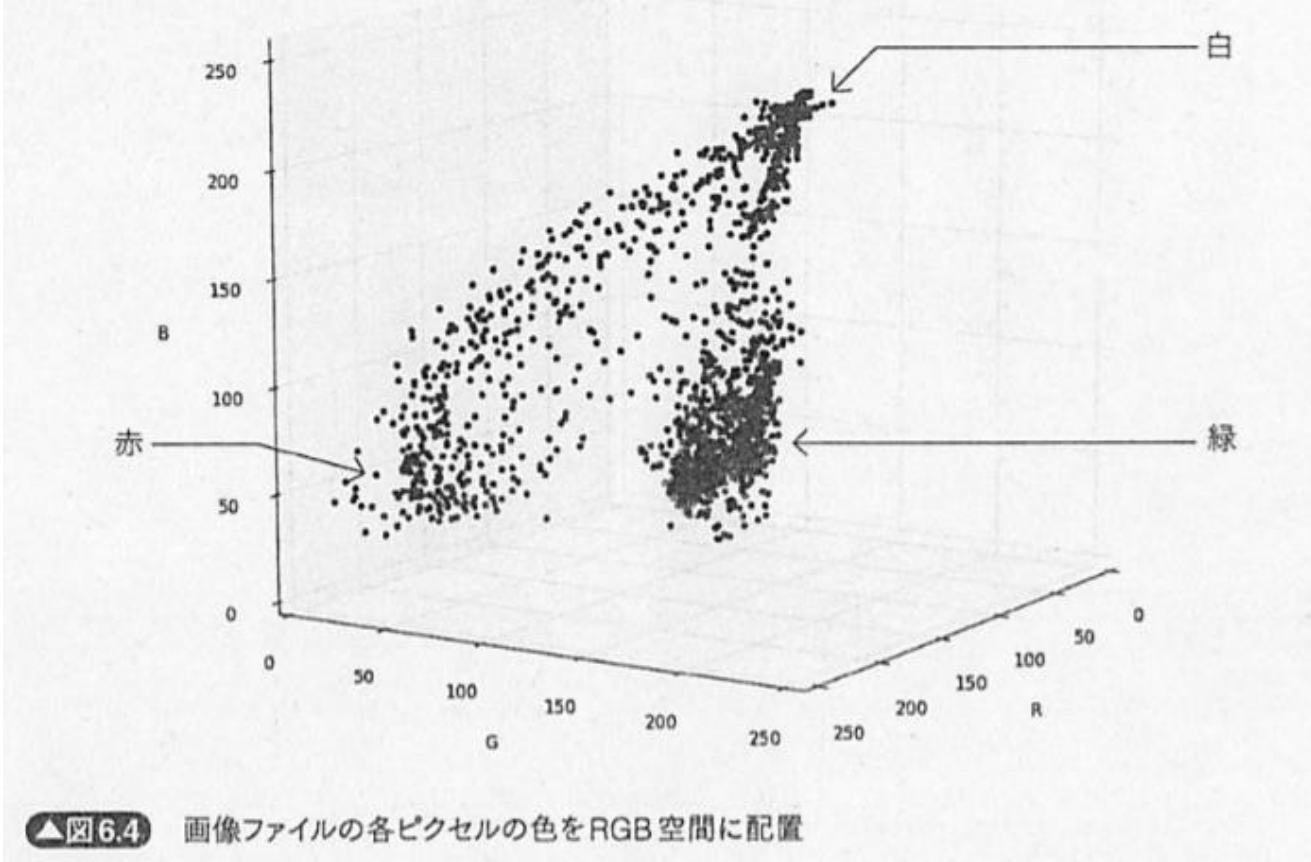
・k−平均法を用いて画像ファイルの減色処理を行う

・右の画像から指定された
　数の「代表色」を抽出する

画像データへの応用

・画像ファイルの
各ピクセルにおける
RGBの値を三次元空間で表す

・空間上の２点間の距離により色の類似性を判別する



▲図 6.4 画像ファイルの各ピクセルの色をRGB空間に配置

画像データへの応用

・k−平均法を適用する際に定める分類するクラスタの数に応じて抽出する代表色の数を変えることができる

・画像ファイルの各ピクセルを代表色で置き換えることで、画像の減色処理を行える

・クラスタの数がK＝2, 3, 5, 16の場合のクラスタリング処理を行う

画像データへの応用

K=2　1回目

```
==========================
Number of clusters: K=2
Initial centers: [[115, 211, 232], [149, 250, 83]]
==========================
[[234, 219, 224], [113, 97, 64]]
Distortion: J=4666062700
[[234, 219, 224], [113, 96, 64]]
Distortion: J=898840622
[[234, 219, 224], [113, 96, 64]]
Distortion: J=898908644
```

画像データへの応用

K=2　2回目

```
==========
Number of clusters: K=2
Initial centers: [[99, 64, 69], [124, 105, 3]]
==========
[[165, 146, 133], [106, 128, 50]]
Distortion: J=5680153782
[[219, 190, 194], [94, 93, 50]]
Distortion: J=2152282248
[[229, 208, 213], [106, 94, 58]]
Distortion: J=1039812376
[[233, 215, 220], [110, 95, 62]]
Distortion: J=920696718
[[234, 218, 222], [112, 96, 63]]
Distortion: J=902490943
[[234, 219, 223], [112, 96, 64]]
Distortion: J=899540054
[[234, 219, 223], [112, 96, 64]]
Distortion: J=899089824
```

画像データへの応用

K=2　3回目

```
==========
Number of clusters: K=2
Initial centers: [[248, 211, 183], [124, 27, 115]]
==========
[[229, 215, 217], [110, 93, 61]]
Distortion: J=1838523044
[[233, 217, 222], [111, 96, 63]]
Distortion: J=907283622
[[234, 219, 223], [112, 96, 63]]
Distortion: J=900193420
[[234, 219, 223], [112, 96, 64]]
Distortion: J=899234811
[[234, 219, 223], [112, 96, 64]]
Distortion: J=899089824
```

画像データへの応用

K=2

画像データへの応用

K=3　1回目

```
==========================
Number of clusters: K=3
Initial centers: [[158, 225, 25], [226, 20, 68], [124, 100, 51]]
==========================
[[237, 232, 232], [204, 60, 98], [97, 112, 65]]
Distortion: J=3844441740
[[236, 227, 229], [196, 61, 91], [88, 111, 59]]
Distortion: J=592524964
[[235, 226, 228], [193, 61, 89], [86, 111, 58]]
Distortion: J=576240864
[[235, 225, 227], [192, 61, 89], [86, 111, 58]]
Distortion: J=575711240
```

画像データへの応用

K=3　2回目

```
==========================
Number of clusters: K=3
Initial centers: [[210, 255, 236], [70, 39, 156], [46, 52, 103]]
==========================
[[232, 219, 221], [200, 79, 110], [95, 98, 54]]
Distortion: J=1841156306
[[236, 227, 229], [194, 79, 102], [86, 105, 54]]
Distortion: J=608007521
[[236, 228, 230], [190, 79, 99], [84, 107, 55]]
Distortion: J=578707450
[[236, 228, 230], [189, 78, 97], [83, 108, 55]]
Distortion: J=576465514
[[236, 227, 230], [189, 77, 97], [83, 108, 55]]
Distortion: J=576177150
```

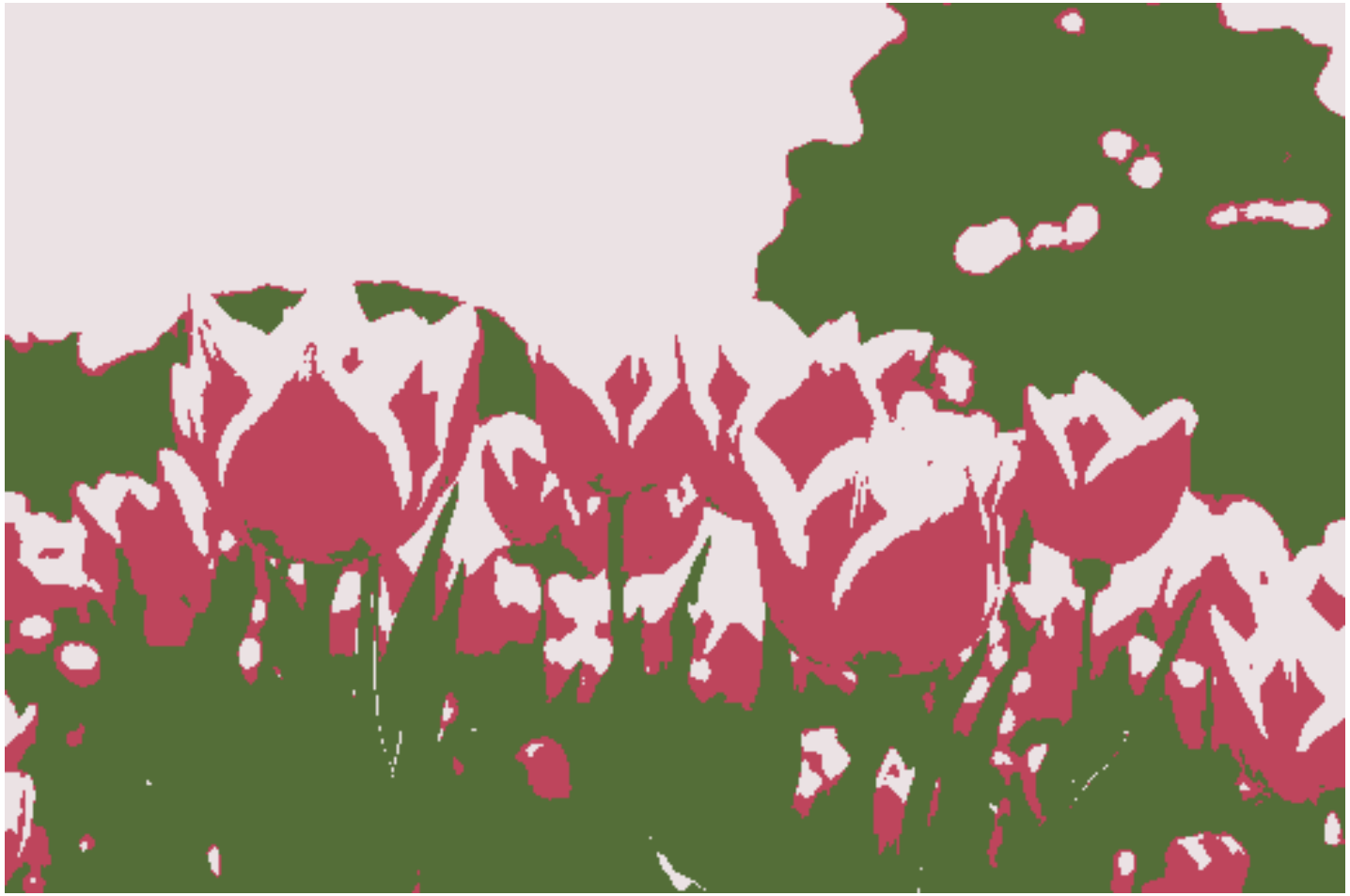
画像データへの応用

K=3　3回目

```
========================
Number of clusters: K=3
Initial centers: [[125, 222, 89], [152, 192, 176], [149, 134, 191]]
========================
[[83, 110, 54], [230, 225, 224], [190, 69, 96]]
Distortion: J=3017139446
[[84, 110, 56], [235, 226, 228], [190, 69, 92]]
Distortion: J=580667686
[[84, 110, 56], [235, 226, 228], [190, 69, 92]]
Distortion: J=576048962
[[84, 110, 56], [235, 226, 228], [190, 69, 92]]
Distortion: J=576048962
```

画像データへの応用

K=3



画像データへの応用

K=5　1回目

```
==========
Number of clusters: K=5
Initial centers: [[130, 83, 128], [146, 229, 150], [4, 181, 83], [75,
172, 161], [208, 86, 186]]
==========
[[118, 95, 62], [230, 230, 226], [46, 83, 28], [113, 154, 107], [225,
121, 157]]
Distortion: J=2130606766

↓↓↓

[[184, 34, 64], [242, 239, 240], [53, 80, 31], [108, 132, 76], [205,
149, 161]]
Distortion: J=254027399
[[184, 34, 64], [242, 239, 240], [53, 79, 31], [108, 132, 76], [205,
150, 161]]
Distortion: J=253907716
```

画像データへの応用

K=5　2回目

```
========================
Number of clusters: K=5
Initial centers: [[19, 187, 66], [76, 242, 103], [165, 253, 91], [128,
90, 30], [36, 122, 34]]
========================
[[93, 149, 101], [119, 158, 112], [229, 218, 219], [146, 92, 75], [58,
90, 39]]
Distortion: J=2532055640

↓↓↓

[[107, 131, 74], [203, 149, 160], [242, 239, 240], [184, 34, 64], [52,
78, 30]]
Distortion: J=254547378
[[107, 132, 74], [204, 150, 160], [242, 239, 240], [184, 34, 64], [52,
78, 30]]
Distortion: J=254361799
```

画像データへの応用

K=5　3回目

```
========================
Number of clusters: K=5
Initial centers: [[156, 47, 67], [112, 34, 29], [99, 37, 199], [153,
193, 248], [155, 102, 125]]
========================
[[186, 36, 64], [67, 83, 35], [0, 0, 0], [237, 230, 232], [143, 133,
103]]
Distortion: J=1318927725

↓↓↓

[[184, 34, 64], [106, 131, 74], [52, 78, 30], [242, 239, 240], [203,
149, 160]]
Distortion: J=254602212
[[184, 34, 64], [107, 132, 74], [52, 78, 30], [242, 239, 240], [204,
150, 160]]
Distortion: J=254435266
```

画像データへの応用

K=5

画像データへの応用

K=16　1回目

```
======================
Number of clusters: K=16
Initial centers: [[43, 62, 134], [122, 126, 69], [74, 93, 57], [228,
246, 43], [114, 236, 92], [161, 131, 43], [46, 189, 212], [118, 109,
100], [107, 43, 178], [247, 207, 238], [112, 101, 158], [106, 89, 195],
[137, 139, 252], [1, 194, 236], [20, 126, 238], [153, 134, 35]]
======================
[[0, 0, 0], [112, 137, 73], [66, 82, 36], [204, 193, 131], [160, 184,
132], [191, 39, 64], [0, 0, 0], [157, 111, 109], [0, 0, 0], [237, 225,
229], [198, 130, 151], [0, 0, 0], [171, 182, 174], [0, 0, 0], [0, 0, 0],
[129, 139, 42]]
Distortion: J=692590576

↓↓↓

[[46, 71, 25], [108, 133, 75], [63, 91, 39], [177, 174, 157], [135, 150,
107], [166, 19, 40], [29, 55, 18], [204, 52, 92], [18, 42, 12], [248,
248, 248], [222, 114, 151], [11, 27, 10], [222, 200, 207], [0, 0, 0],
[0, 0, 0], [86, 115, 51]]
Distortion: J=99866797
```

画像データへの応用

K=16　2回目

```
==========
Number of clusters: K=16
Initial centers: [[229, 29, 161], [245, 232, 205], [63, 79, 164], [83,
86, 86], [209, 227, 247], [194, 33, 74], [107, 93, 159], [132, 171, 70],
[17, 184, 126], [16, 50, 152], [252, 255, 173], [198, 194, 30], [45,
146, 30], [179, 119, 203], [171, 109, 0], [233, 194, 219]]
==========
[[226, 80, 133], [240, 236, 235], [0, 0, 0], [73, 94, 50], [248, 248,
249], [183, 34, 62], [135, 133, 126], [124, 146, 89], [0, 0, 0], [0, 0,
0], [203, 196, 134], [0, 0, 0], [58, 96, 30], [202, 140, 157], [142, 96,
47], [219, 192, 200]]
Distortion: J=488932745

↓↓↓

[[206, 42, 86], [225, 217, 220], [48, 72, 26], [84, 114, 49], [250, 250,
250], [163, 17, 38], [124, 149, 100], [104, 131, 71], [33, 59, 20], [23,
48, 14], [164, 167, 141], [13, 35, 10], [63, 91, 39], [222, 104, 144],
[164, 96, 85], [217, 175, 188]]
Distortion: J=83605636
```

画像データへの応用

K=16　3回目

```
========================
Number of clusters: K=16
Initial centers: [[185, 195, 223], [170, 49, 9], [69, 3, 5], [122, 26,
78], [176, 51, 75], [102, 87, 147], [30, 86, 16], [11, 248, 176], [185,
214, 113], [184, 225, 152], [100, 10, 125], [159, 251, 228], [194, 172,
136], [235, 103, 217], [160, 50, 140], [247, 180, 2]]
========================
[[239, 234, 236], [161, 15, 30], [97, 20, 20], [119, 51, 47], [187, 55,
76], [106, 134, 85], [64, 93, 37], [0, 0, 0], [133, 163, 82], [182, 195,
172], [0, 0, 0], [232, 251, 246], [169, 154, 131], [233, 123, 169],
[216, 81, 129], [0, 0, 0]]
Distortion: J=815162041

↓↓↓

[[228, 218, 222], [185, 19, 48], [144, 17, 30], [159, 95, 81], [207, 43,
88], [98, 125, 63], [78, 108, 45], [57, 83, 35], [117, 144, 91], [207,
189, 191], [41, 66, 22], [250, 250, 250], [154, 162, 130], [226, 139,
170], [218, 89, 131], [23, 49, 15]]
Distortion: J=73942934
```

画像データへの応用

K=16

k−平均法の数学的根拠

・今回のk-平均法では「二乗歪み」というグループ分けの基準を用いている。

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} |\mathbf{x}_n - \boldsymbol{\mu}_k|^2 \tag{6.4}$$

k平均法の手続きによって、この「二乗歪み」が最終的に極小値に達することを証明する。

k−平均法の数学的根拠

- $\mathbf{x}_n{}_{n=1}^{N}$: 任意の特定次元ベクトル。トレーニングセット
  として扱う。
- $\mu_k{}_{k=1}^{K}$: K個のクラスターに分類するとしたときの
  代表点
- 各データが属するクラスター

$$r_{nk} = \begin{cases} 1 & \mathbf{x}_n \text{が} k \text{番目の代表点に属する場合} \\ 0 & \text{それ以外の場合} \end{cases} \tag{6.5}$$

k−平均法の数学的根拠

・二重歪みの値を表す式

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} |\mathbf{x}_n - \mu_k|^2 \qquad (6.6)$$

以降、k平均法の手続きに従ってr_{nk}とμ_kを修正していくとJの値は減少し、極小値に達することを示す。

k−平均法の数学的根拠

・まず、各データが属するクラスタを選択しなおす。

この時、各データx_nについて、代表点からの距離$|x_n - \mu_k|$が最も小さいクラスターを選択する。するとr_{nk}を以下の条件で再定義できる。

$$r_{nk} = \begin{cases} 1 & k = \arg\min_{k'} |\mathbf{x}_n - \mu_{k'}| \\ 0 & \text{それ以外の場合} \end{cases} \tag{6.7}$$

k−平均法の数学的根拠

・次に、現在の分類状態において、各クラスターの代表点を取り直す。

・(6.6)を最小にするμ_kを選択する。

・(6.6)をμ_kについてみると下に凸な二次関数となっている。
→　これは偏微分係数が0になるとき最小化するということ

k−平均法の数学的根拠

・Jを成分表記すると以下のようになる。

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} \left\{ r_{nk} \sum_{i} ([\mathbf{x}_n]i - [\mu_k]i)^2 \right\} \tag{6.8}$$

k−平均法の数学的根拠

・kとiが特定の成分なので、k=k′ , i=i′ として、Jを以下のように変形する。

$$J' = \sum_{n=1}^{N} r_{nk'}([\mathbf{x}_n]i' - [\mu_{k'}]i')^2$$

$$= \sum_{n=1}^{N} r_{nk'}([\mathbf{x}_n]i'^2 - 2[\mathbf{x}_n]i'[\mu_{k'}]i' + [\mu_{k'}]i'^2)$$

k−平均法の数学的根拠

・$[\mu_{k'}]_{i'}$で偏微分すると

$$\frac{\partial J}{\partial [\mu_{k'}]_{i'}} = -2 \sum_{n=1}^{N} r_{nk'} ([\mathbf{x}_n]_{i'} - [\mu_{k'}]_{i'})$$

・元々は特定の成分$[\mu_k]_i$なので書き直すと

$$\frac{\partial J}{\partial [\mu_k]_i} = -2 \sum_{n=1}^{N} r_{nk} ([\mathbf{x}_n]_i - [\mu_k]_i) \qquad (6.9)$$

k−平均法の数学的根拠

・偏微分係数が0になるという条件から

$$-2\sum_{n=1}^{N} r_{nk}([\mathbf{x}_n]_i - [\mu_k]_i) = 0$$

$$-2\sum_{n=1}^{N} r_{nk}[\mathbf{x}_n]_i + 2\sum_{n=1}^{N} r_{nk}[\mu_k]_i = 0$$

$$\sum_{n=1}^{N} r_{nk}[\mu_k]_i = \sum_{n=1}^{N} r_{nk}[\mathbf{x}_n]_i$$

$$[\mu_k]_i = \frac{\sum_{n=1}^{N} r_{nk}[\mathbf{x}_n]_i}{\sum_{n=1}^{N} r_{nk}} \tag{6.10}$$

k−平均法の数学的根拠

・成分表記からベクトル表記に戻すと、以下の結果が得られる。

$$\mu_k = \frac{\sum_{n=1}^{N} r_{nk}\mathbf{x}_n}{\sum_{n=1}^{N} r_{nk}} \qquad (6.11)$$

・これはクラスターの重心を新たな代表点にとるという(6.3)の手続きと同じものになっている。したがって、(6.3)の手続きによって、Jが大きくなることはあり得ない。

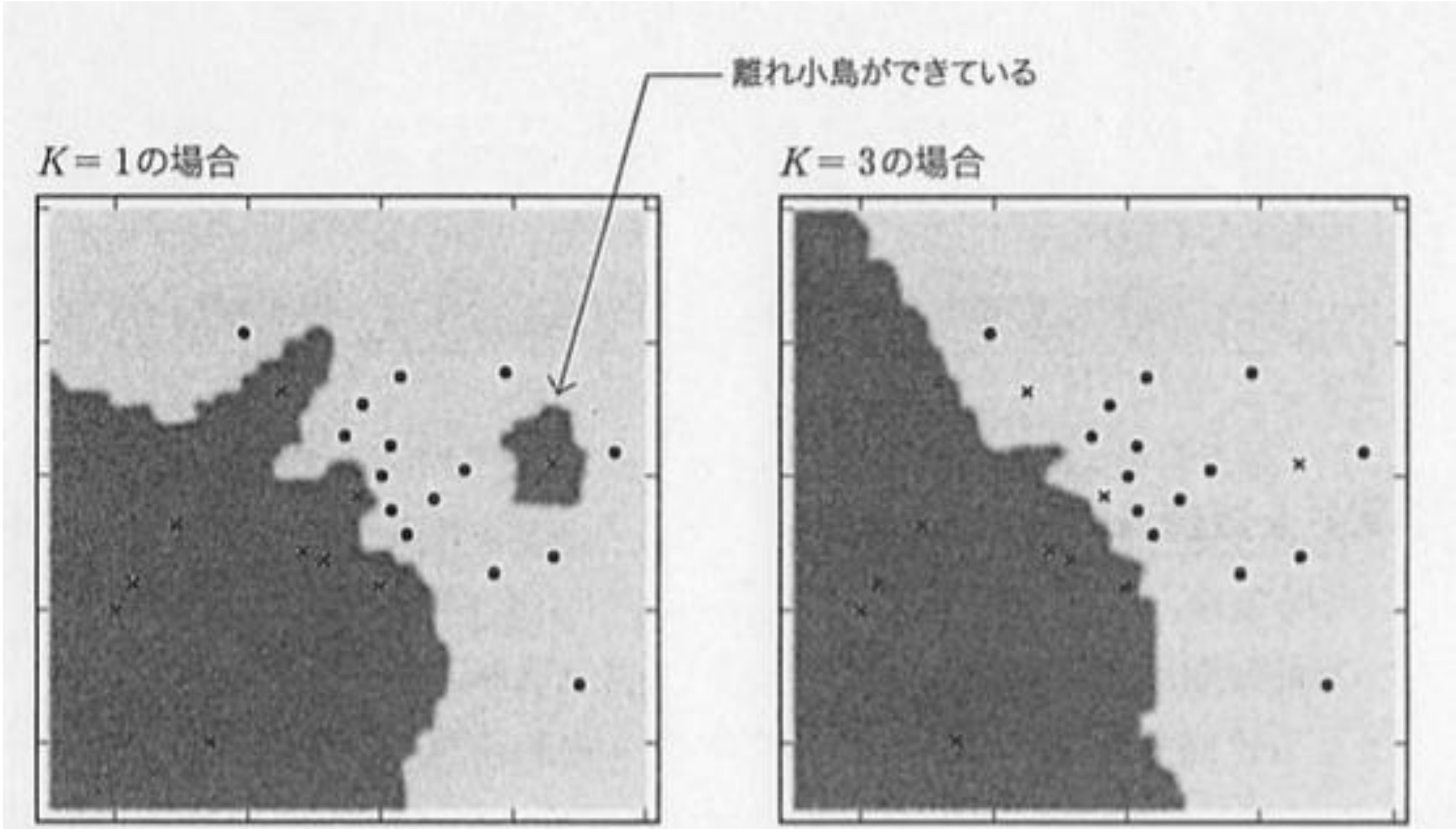
・以上により、k平均法の操作を繰り返すと、Jの値は必ず小さくなるか、それ以上変化しない極小値に達する。

怠惰学習モデルとしてのk−近傍法

k−近傍法とは

・教師あり学習に分類される。

・新たなデータ(x,y)が与えられた際に、その周りのデータを見て、自分の近くにあるデータの目的変数の値から、自分自身の目的変数を推定するということを行います。



▲図6.8 k近傍法による分類結果

怠惰学習モデルとしてのk−近傍法

k−近傍法の問題点

1. 新たなデータの分類に時間がかかる。
   - データが与えられた場合、すべてのデータについて参照して計算しなおす必要がある。

2. 分析のモデルが明確ではない。
   - 与えられたデータからたまたまそのように分類できた、というだけ。
   - 仮説と実証ではなく、単純に事実からの判断でしかないため将来に生かせない。