統計的機械学習論

6. k平均法

夏季集中講義(2019/09/02-13)

@九州工業大学

(飯塚キャンパス 大学院セミナー室７F)

このクラスのねらいと達成目標

ねらい

教師なしクラスタリングの基礎であるk平均法のアルゴリズムを理解する。

達成目標

- **代表点の更新手続きを理解する**
- **k平均法の応用例を知る**
- **k近傍法との違いを把握する**

クラスの進行

6 k平均法

類似するデータをグループ化するアルゴリズム。

類似するk近傍法（怠惰学習による分類アルゴリズム）についても紹介。

6.1 k平均法によるクラスタリングと応用例

画像の「色」グループ化

文書のカテゴリー判定

6.1.1 教師なし学習としてのクラスタリング

教師なしクラスタリングの基礎を学ぼう

例：データセット：$\mathbf{x}_n = (x_n, y_n)^{\mathrm{T}}$，代表点：$\{\mu_k\}_{k=1}^{2}$

代表点との距離 $\|\mathbf{x}_n - \boldsymbol{\mu}_n\|$ を計算

↓

距離が短い方の代表点に所属する

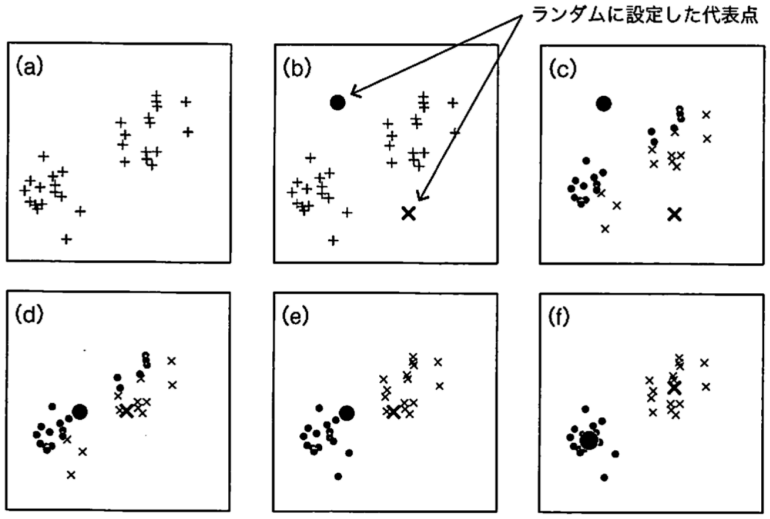
↓

変数 r_{nk} に属性を代入する

$r_{nk} = 1 \rightarrow x_n$ はk番目の代表点に属する

$r_{nk} = 0 \rightarrow x_n$ はk番目の代表点に属さない

6.1.1 教師なし学習としてのクラスタリング



6.1.2 k平均法としてのクラスタリング

現在のクラスターが「最適」な分類ではないとき

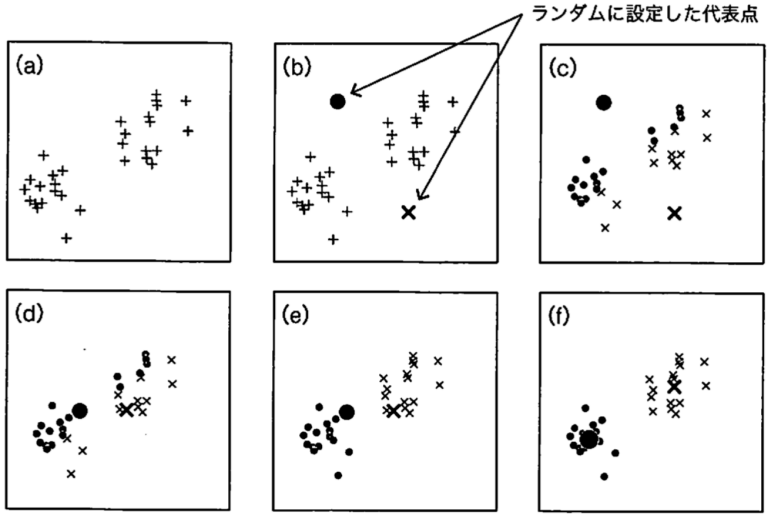
改めて代表点を取り直す。この時の代表点は既存の代表点を元に分類したクラスターの「重心」を新たな代表点にする。

$$\boldsymbol{\mu}_k = \frac{\sum \mathbf{x}_n}{N_k} \qquad (k = 1, 2)$$

この時の重心はk番目の代表点に属する点についてのみ求める。

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} r_{nk}\mathbf{x}_n}{\sum_{n=1}^{N} r_{nk}} \qquad (k = 1, 2)$$

6.1.1 教師なし学習としてのクラスタリング



6.1.2 k平均法としてのクラスタリング

現実問題ではトレーニングセットは複雑で多数のクラスターについて分類しなければならないこともある。

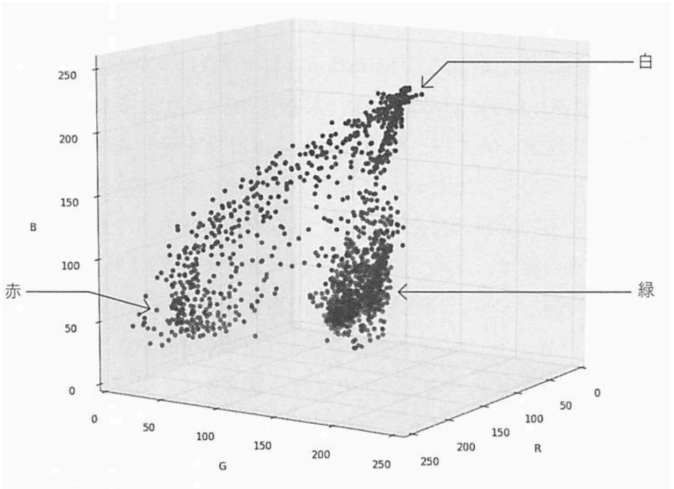
代表点の初期位置によって結果が異なる時もある。

計算を繰り返してより適切と思われるクラスターを発見しよう。

グループ分けを判定する「二乗歪み」という基準がある。

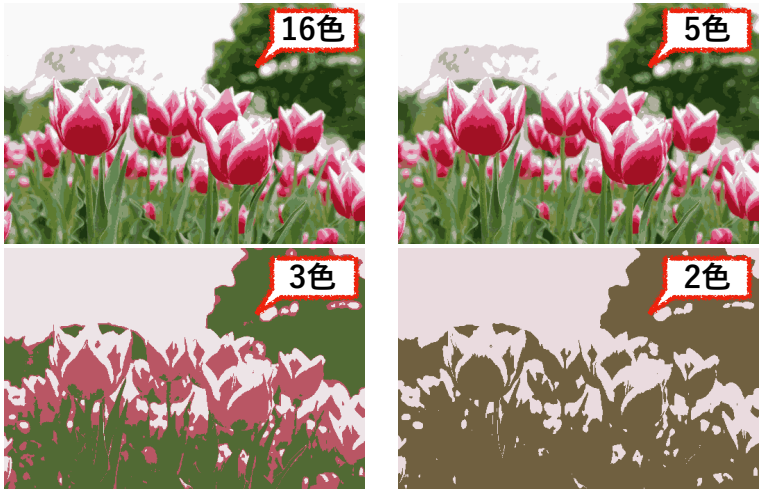
6.1.3 画像データへの応用

代表色を決め、色空間における座標データからデータを分類

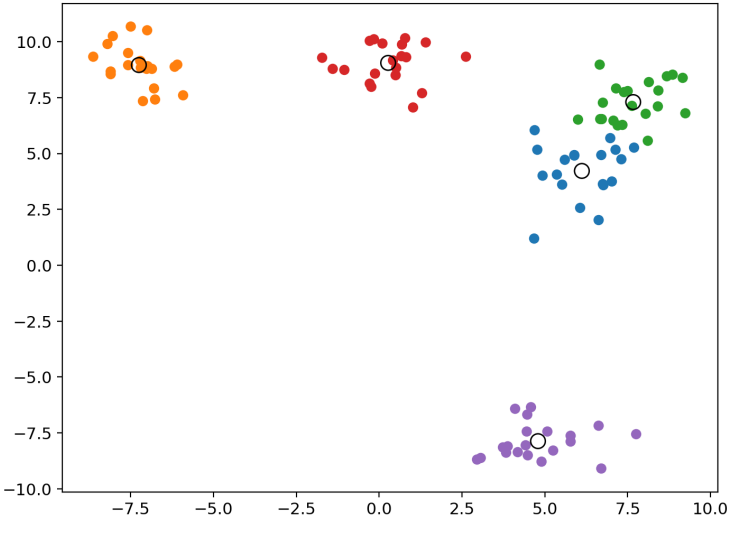


6.1.3 画像データへの応用

代表色を決め、色空間における座標データからデータを分類



6.1.4 サンプルコードによる確認



6.1.5 k平均法の数学的根拠

グループ分けに対する歪みの値を計算

↓

「二乗歪み」

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

代表点を更新して「二乗歪み」をできるだけ小さくしていく

↓

代表点近くにデータを集めるように分類

$r_{nk} = 1$　　$k = argmin\|\mathbf{x}_n - \boldsymbol{\mu}_n\|$の場合

$r_{nk} = 0$　　それ以外の場合

6.1.5 k平均法の数学的根拠

$\|\mathbf{x}_n - \boldsymbol{\mu}_n\|^2$が最小になるように属性を決める

↓

Jについても偏微分を行う

↓

「２乗歪み」が最小になるような代表点を取り直す

Jの減少分が元のJの0.1%以下になった時に計算を終了する

6.1.6 他の分類におけるk平均法

- 文書の頻出単語（Term Frequency; TF）↓
- 珍しい単語の出現頻度（Term Frequency-Inverse Document Frequency; TF-IDF）

6.2 怠惰学習モデルとしてのk近傍法

例：データセット：$\{x_n, y_n, t_n\}_{n=1}^{N} = 0$

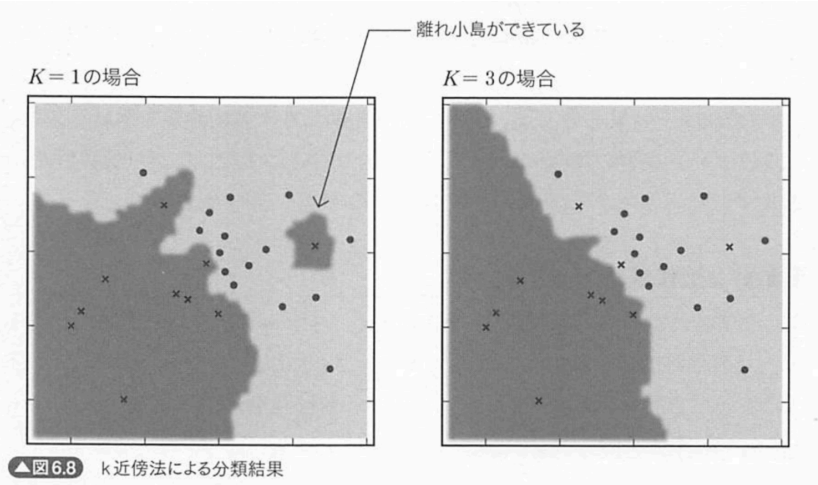
新規データに近いデータの属性から新規データの属性を判定

↓

新規データが来た時に周囲のデータの属性をもとに判定する

6.2.1 k近傍法による分類絵

教師なしクラスタリングの基礎を学ぼう



▲図6.8 k近傍法による分類結果

6.2.2 k近傍法の問題点

k近傍法の問題的：

・計算時間が長くかかる

・グループ分けが明確でない

・経験則に基づいた考え方であり、グループ分けにの基準の根拠が薄弱。

まとめ

- k平均法について理解できた
- 代表点の更新の手続きについて説明できる
- グループ分けを評価する「二乗歪み」を理解した

ミニッツペーパー

- 代表値の更新の手続きを説明してください。
  スライドの6~14までを参考に手続きの初期条件と終了条件、各データにとっては各代表点の中で最も近い代表点に所属することを決めること、新しい代表値の座標が同じ属性を持つデータ群の重心になることを記述する