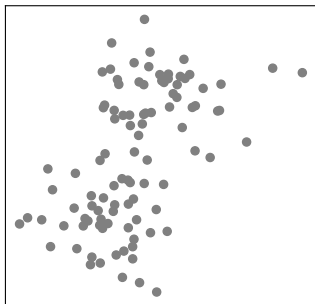
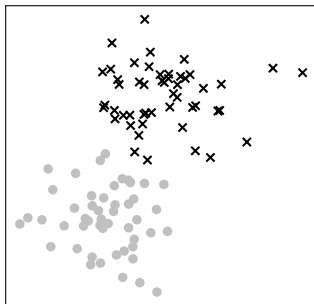


# 11. モデル推定



(a) 正解情報のないデータ



(b) クラスタリング結果の例



(c) 確率密度推定結果の例

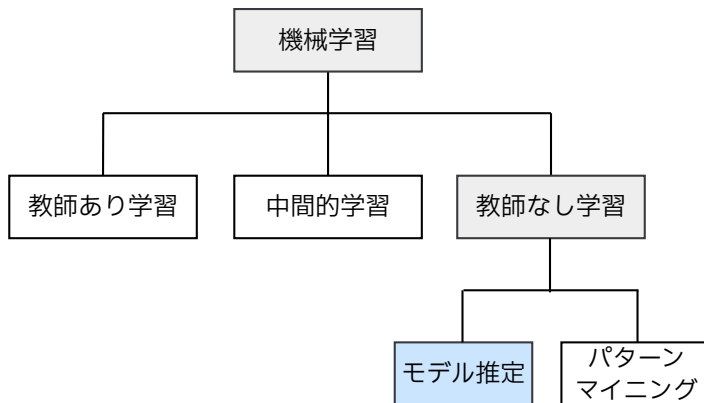
- 11.1 「教師なし・モデル推定」問題の定義
- 11.2 次元削減
- 11.3 クラスタリング
- 11.4 確率密度推定



- 荒木雅弘：『Pythonではじめる機械学習』（森北出版、2025年）
- スライドとコード

## 11.1 「教師なし・モデル推定」問題の定義 (1/3)

- 問題設定
  - 正解なし数値ベクトル → クラスモデル
  - データ全体を説明するモデルを見つける



- 応用例
  - データ可視化, 顧客セグメンテーション, 異常検知

## 11.1 「教師なし・モデル推定」問題の定義 (2/3)

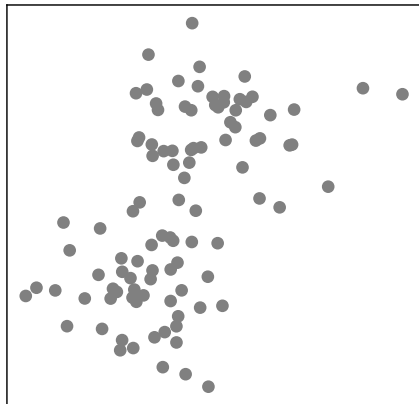
- データセット（正解なし）
  - （密な） $d$  次元数値ベクトルの集合

$$\{\boldsymbol{x}_i\} \quad (i = 1, \dots, N)$$

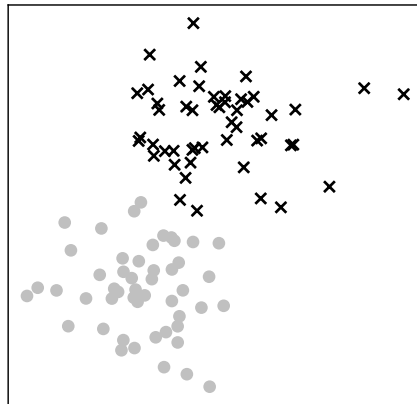
- モデル推定とは
  - 次元削減
    - 高次元空間上のデータを低次元空間に写像して、データの本質的な構造を捉える
  - クラスタリング
    - 個々のデータを生じさせた共通の性質をもつクラスを見つける
  - 確率密度推定
    - データが複数のクラスに分られると仮定した場合の統計的性質を推定する

## 11.1 「教師なし・モデル推定」問題の定義 (3/3)

- 正解なしデータ，クラスタリング結果，確率密度推定結果



(a) 正解情報のないデータ



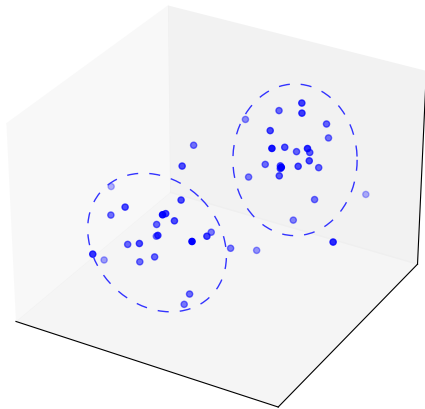
(b) クラスタリング結果の例



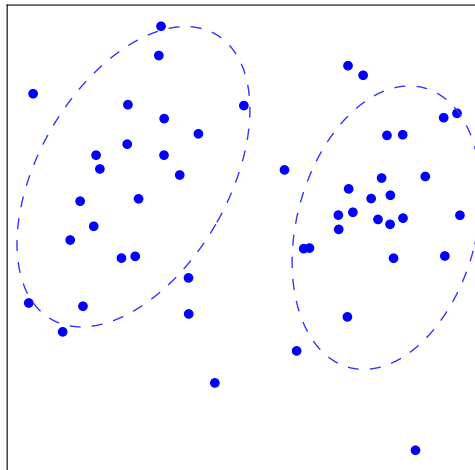
(c) 確率密度推定結果の例

## 11.2 次元削減 (1/3)

- 次元削減の目的
  - 高次元空間上のデータを低次元空間へ写像することで、データの本質的な構造を捉える
  - データの可視化に有効



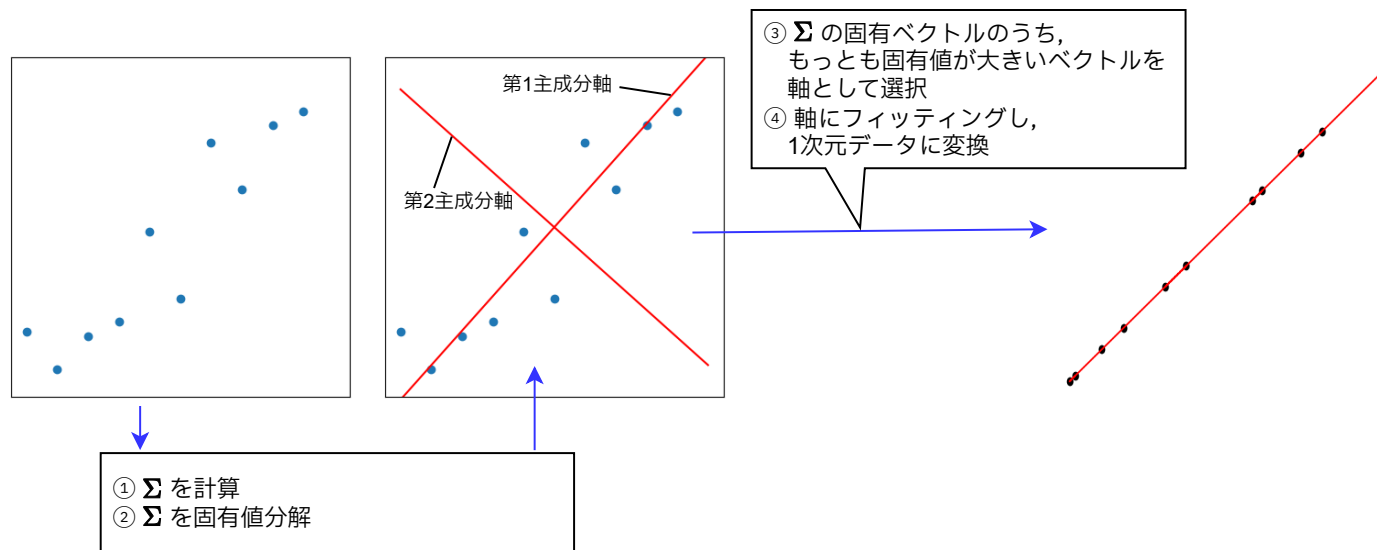
(a) 高次元空間上のデータ



(b) 次元削減後のデータ

## 11.2 次元削減 (2/3)

- 主成分分析 (PCA)
  - 高次元空間上のデータの散らばり方をできるだけ保存する低次元空間への写像を求める
  - データの共分散行列  $\Sigma$  を固有値分解し, 固有値の大きい順に  $k$  個の固有ベクトルからなる部分空間へデータを射影



## 11.2 次元削減 (3/3)

- t-SNE
  - 元の高次元空間における処理
    - どの範囲のデータを類似度計算の対象とみなすかをパラメータ perplexity で与える
    - データ  $\mathbf{x}_i$  と  $\mathbf{x}_j$  の類似度を,  $\mathbf{x}_i$  の近傍として  $\mathbf{x}_j$  を選択する条件付き確率  $p_{ij}$  とする
      - $p_{ij}$ : 平均を  $\mathbf{x}_i$ , 分散を perplexity に基づいて求めた  $\sigma^2$  とする正規分布に基づいて計算
  - 削減後の低次元空間における処理
    - データ  $\mathbf{y}_i$  と  $\mathbf{y}_j$  の類似度  $q_{ij}$  を, 自由度1のt分布に基づいて計算
      - t分布は正規分布よりも値の大きい範囲が広い
  - 最適化
    - $p_{ij}, q_{ij}$  両分布間のKL距離を最小化するように  $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_N\}$  の位置を逐次更新

$$KL(P, Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

## 11.3 クラスタリング

- クラスタリングとは
  - 「共通の性質をもつクラス」 = 「特徴空間上で近い値をもつデータの集まり」と考え、データのまとまりを見つける
  - まとまり：「内的結合の小ささ」と「外的分離の大きさ」が同時に満たされる集合
    - 内的結合: 同じ集合内のデータ間の距離
    - 外的分離: 異なる集合間の距離
- クラスタリング手法の分類
  - 階層的手法
    - ボトムアップ的にデータをまとめてゆく
  - 分割最適化手法
    - トップダウン的にデータ集合を分割し、最適化してゆく



## 11.3.1 階層的クラスタリング (1/3)

- 階層的クラスタリングの手順

- 入力: 正解なしデータ  $D$

- 出力: デンドログラム

1. 学習データそれぞれをクラスタの要素としたクラスタ集合  $C$  を作成

$$C \leftarrow \{c_1, c_2, \dots, c_N\}$$

2. **while**  $|C| > 1$ :

- もっとも類似度が大きいクラスタ対  $\{c_m, c_n\}$  をみつける

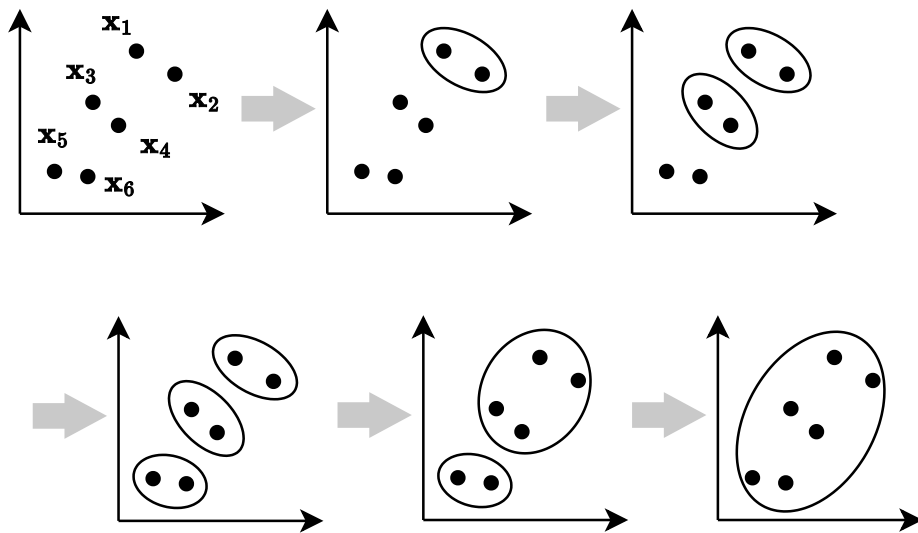
$$(c_m, c_n) \leftarrow \operatorname{argmax}_{c_i, c_j \in C} \operatorname{sim}(c_i, c_j)$$

- $\{c_m, c_n\}$ を融合し, デンドログラムに記録する

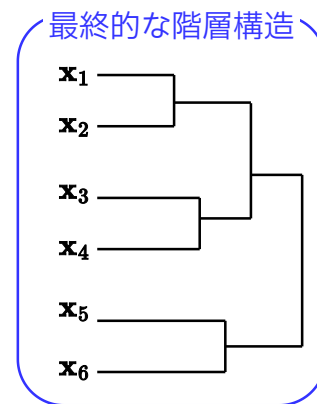
3. **return** デンドログラム

## 11.3.1 階層的クラスタリング (2/3)

- 階層的クラスタリングの手順



(a) 階層的にデータをまとめていく手順



(b) デンドログラム (樹形図)

## 11.3.1 階層的クラスタリング (3/3)

- 距離 (linkage) の定義とできるクラスタの傾向
  - 単連結法 (single) : もっとも近いデータ対の距離
    - クラスタが一方向に伸びやすくなる
  - 完全連結法 (complete) : もっとも遠いデータ対の距離
    - 直径の小さいクラスタが優先的に形成される
  - 群平均法 (average) : すべてのデータ対の距離の平均
    - 単連結と完全連結の中間的な形
  - Ward法 (ward) : 融合前後の「クラスタ内のデータと平均との距離の二乗和」の差
    - 極端な形になりにくく, よく用いられる基準

## 11.3.2 分割最適化クラスタリング (1/3)

- 分割最適化クラスタリングとは
  - データ分割の良さを評価する関数を定め，その評価関数の値が最適となる分割を求める
  - ただし，すべての可能な分割に対して評価値を求めることは，データ数  $N$  が大きくなると不可能
    - 例：2分割で  $2^N$  通り
  - 従って，適切な初期値から探索によって準最適解を求める

## 11.3.2 分割最適化クラスタリング (2/3)

- k-meansアルゴリズム

- 入力: 正解なしデータ  $D$
- 出力: クラスタ中心  $\mu_j$  ( $j = 1, \dots, k$ ) およびクラスタリング結果
  1. 入力空間上に  $k$  個の点をランダムに設定し, それらをクラスタ中心  $\mu_j$  の初期値とする
  2. 以下の処理を, すべてのクラスタ中心  $\mu_j$  が変化しなくなるまで繰り返す

**for**  $x_i \in D$ :

各クラスタ中心  $\mu_j$  との距離を計算し, もっとも近いクラスタに割りあてる

**for**  $j = 1, \dots, k$ :

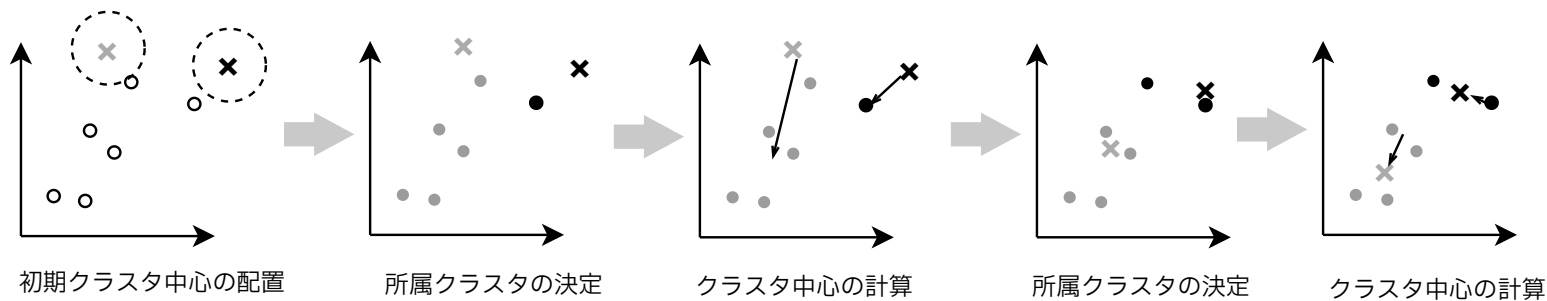
以下の式でクラスタ中心  $\mu_j$  の位置を更新する ( $N_j$  はクラスタ  $j$  のデータ数)

$$\mu_j \leftarrow \frac{1}{N_j} \sum_{x_i \in \text{クラスタ}_j} x_i$$

3. **return** クラスタ中心  $\mu_j$  ( $j = 1, \dots, k$ ) およびクラスタリング結果

## 11.3.2 分割最適化クラスタリング (3/3)

- k-meansアルゴリズム

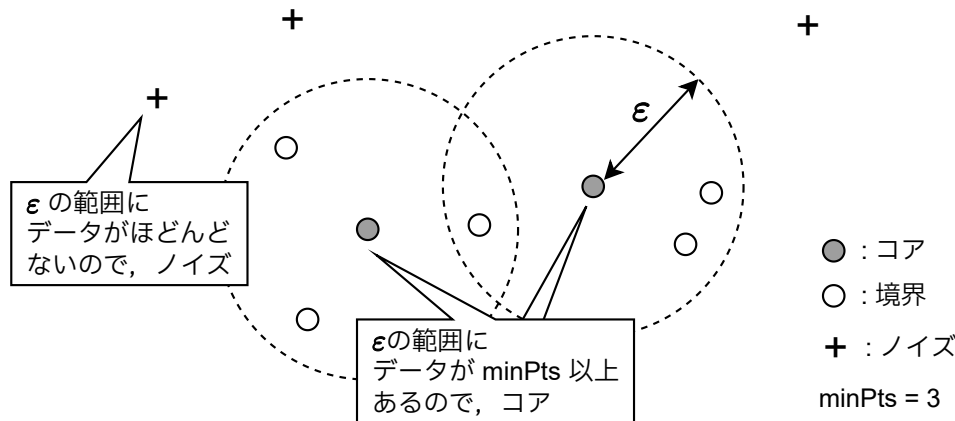


## 11.3.3 自動でクラスタ数を決定するクラスタリング (1/3)

- 分割最適化クラスタリングの問題点
  - クラスタ数  $k$  を予め決めなければならない
- クラスタ数自動決定法のアイデア
  - クラスタ数  $k$  を変化させながらクラスタリングを行い、クラスタリングの妥当性を評価する指標を用いて最適な  $k$  を決定する
    - $k$  を順に大きくして、データとクラスタ中心との平均二乗距離が急激に減少しなくなる点を選ぶ（例：k-means におけるエルボーメソッド）
    - モデルの複雑さと適合度のバランスを考慮した情報量規準を最小化する  $k$  を選ぶ（例：確率密度推定におけるBIC 規準）
  - データの密度に基づく方法 → DBSCAN

## 11.3.3 自動でクラスタ数を決定するクラスタリング (2/3)

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
  - データの密度が高い部分をクラスタ, データの密度が低い部分を外れ値とする手法
  - パラメータ
    - $\epsilon$  (epsilon) : あるデータについて, どれくらいの範囲までを近いとみなすかを定める距離
    - minPts : 近いとみなした範囲内にデータポイントがいくつあれば密集しているとみなすかの数





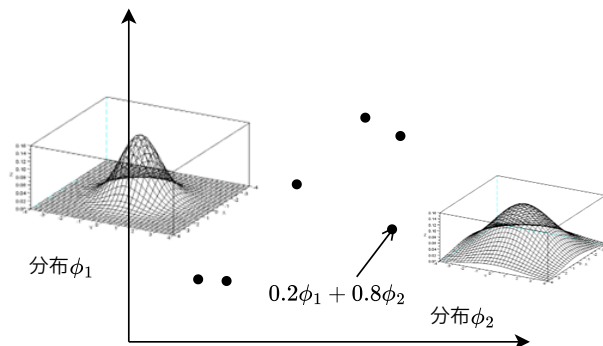
## 11.3.3 自動でクラスタ数を決定するクラスタリング (3/3)

- 手順

1. 近傍点を探索 個々のデータについて, 距離  $\epsilon$  の範囲内にある他のデータ (近傍点) を調べる.
2. データの分類 各データを, 以下の基準で分類する.
  - コア:  $\epsilon$  の範囲内に minPts 個以上のデータがある
  - 境界:  $\epsilon$  の範囲内に他のデータはあるが, minPts 未満のデータしかない
  - ノイズ:  $\epsilon$  の範囲内にほとんど点がない
3. クラスタ形成 近傍にコアがあれば, それをとり込んでクラスタを拡張する.

## 11.4 確率密度推定 (1/5)

- 教師なし学習で識別器を作る問題
  - クラスタリング結果からは、1クラス1プロトタイプの単純な識別器しかできない
  - 各クラスの事前確率や確率密度関数も推定したい
- ガウス混合分布モデル
  - データの広がりを経数の正規分布の混合で表す
  - $k$  個の初期分布を与え、EMアルゴリズムで最適化してゆく



## 11.4 確率密度推定 (2/5)

- k-means法からガウス混合分布モデルへ (EMアルゴリズム)
  - $k$  個のクラスタ中心を乱数で決める
    - ⇒  $k$  個の正規分布を乱数で決める
  - クラスタ中心との距離を基準に各データをいずれかのクラスタに所属させる
    - ⇒ 正規分布から求められる確率に基づき、データを各クラスタに緩やかに所属させる
  - 所属させたデータをもとにクラスタ中心を再計算
    - ⇒ データのクラスタへの所属度を重みとして、分布のパラメータを再計算

## 11.4 確率密度推定 (3/5)

- EM アルゴリズム

- 入力 : 正解なしデータ  $D$

- 出力 : 各クラスを表す確率密度関数のパラメータ

1. 入力空間上に  $k$  個のクラス  $c_j$   $j = 1, \dots, k$  を, ランダムに分布  $\phi_j$  を設定して配置する

2. 以下を, 分布のパラメータの変化量が閾値以下になるか, 一定回数になるまで繰り返す

- /\* Eステップ \*/

- **for**  $\mathbf{x}_i \in D$ :

- $\phi_j$  を用いて確率  $p(c_j \mid \mathbf{x}_i)$  ( $j = 1, \dots, k$ ) を計算

- 確率の和が1になるように正規化

- /\* Mステップ \*/

- Eステップで求めた  $p(c_j \mid \mathbf{x}_i)$  をデータの重みとして, 分布  $\phi_j$  のパラメータを再計算

3. **return** 分布  $\phi_j$  のパラメータ

## 11.4 確率密度推定 (4/5)

- E (Expectation) ステップ：確率計算

$$p(c_m | \mathbf{x}_i) = \frac{p(c_m)p(\mathbf{x}_i | c_m)}{p(\mathbf{x}_i)} = \frac{p(c_m)p(\mathbf{x}_i | c_m)}{\sum_{j=1}^k p(c_j)p(\mathbf{x}_i | c_j)} = \frac{p(c_m)\phi(\mathbf{x}_i; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}{\sum_{j=1}^k p(c_j)\phi(\mathbf{x}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

- M (Maximization) ステップ：分布の最尤推定

$$\boldsymbol{\mu}_m = \frac{1}{|D|} \sum_{\mathbf{x}_i \in D} p(c_m | \mathbf{x}_i) \mathbf{x}_i$$

$$\boldsymbol{\Sigma}_m = \frac{1}{|D|} \sum_{\mathbf{x}_i \in D} p(c_m | \mathbf{x}_i) (\mathbf{x}_i - \boldsymbol{\mu}_m)(\mathbf{x}_i - \boldsymbol{\mu}_m)^\top$$

## 11.4 確率密度推定 (5/5)

- ガウス混合分布モデルの問題点
  - 分割数  $k$  を予め決めなければならない
- 情報量規準の最小化
  - 2分割から始めて、分割数を適応的に決定する
  - 分割の妥当性の判断：BIC (Bayesian Information Criterion)が小さくなれば分割を継続

$$BIC = -2 \log L + q \log N$$

- $L$ : モデルの尤度
- $q$ : モデルのパラメータ数
- $N$ : データ数

## まとめ

- モデル推定
  - データのまとまりを発見するプロセス
- 次元削減
  - 高次元空間上のデータを低次元空間に写像して，データの本質的な構造を捉える
- クラスタリング
  - 階層的クラスタリングは，類似度に基づいてボトムアップにデータをまとめてゆく
  - 分割最適化クラスタリングは，トップダウンでのデータの分割を最適化
- 確率密度推定
  - 分割最適化クラスタリングの一般化