カーネル多変量解析

第3 章固有値問題を用いたカーネル多変量解析
3.3 クラスタリング

茨城大学工学部情報工学科
國井慎也

クラスタリング

- クラスタリングとは

  -データの集まりをデータ間の類似度に従って、いくつかのグループに分けること

- クラスタリングのアルゴリズム

  -k-平均(means)法

  -スペクトラルクラスタリング

カーネルk-平均法(1)

- サンプル点集合:
$$x^{(1)}, x^{(2)}, \ldots, x^{(n)}$$
- その特徴ベクトル:
$$\varphi\left(x^{(1)}\right), \varphi\left(x^{(2)}\right), \ldots, \varphi\left(x^{(n)}\right)$$
- 代表点:
$$\mu_1, \mu_2, \ldots, \mu_c$$
-代表点はあらかじめ決めておく

カーネルk-平均法(2)

- それぞれのサンプル点は、その点に最も近い代表点のグループに入れる。

  式で書くと

$$N_i = \{ x^{(l)} \mid \mu_i = \arg\min_{\mu_j} \| \varphi(x^{(l)}) - \mu_j \|^2 \}$$

- 代表点はグループに属するサンプル点の重心を取る。

  式で書くと

$$\mu_i = \frac{1}{|N_i|} \sum_{x^{(j)} \in N_i} \varphi(x^{(j)})$$

カーネルk-平均法(3)

- 解を求める

  -代表点からメンバーへの二乗距離の総和が小さくなるように代表点を決める。

  式で書くと

$$L=\sum_{i=1}^{c}\sum_{x^{(j)}\in N_i}\|\varphi\left(x^{(j)}\right)-\mu_i\|^2$$

  を最小化するような N_i, μ_i を決める問題になる

最適化の方法

- k-平均法では、代表点とグループを両方一度に最適化することは難しい。

  -適当な初期値からスタートし、一方を固定し他方を最適化を行う。

距離の計算

- どのグループに属しているかを判定する
  -特徴ベクトルと代表ベクトルの距離を計算する

$$\|\varphi(x^{(j)})-\mu_i\|^2=\|\varphi(x^{(j)})-\frac{1}{|N_i|}\sum_{x^{(l)}\in N_i}\varphi(x^{(l)})\|^2$$

$$=k(x^{(j)},x^{(j)})-\frac{2}{|N_i|}\sum_{x^{(l)}\in N_i}k(x^{(j)},x^{(l)})+\frac{1}{|N_i|^2}\sum_{x^{(l)}\in N_i}\sum_{x^{(m)}\in N_i}k(x^{(l)},x^{(m)})$$

- 最終的に、カーネル関数を使って表される。

カーネルk-平均法(4)

- カーネルk-平均法

  [1]サンプルを適当にc個に分け、N_iを初期化する

  [2]式（3.55）に基づいてN_iを更新

  [3]グループ分けが収束するまで[2]を繰り返す

- k-平均法では、目的関数Lは単調減少するので、局所最適解に収束する。しかし、大域的最適化に収束するとは限らない。

k-平均法の欠点

- K-平均法の欠点

  -反復演算が必要

  -収束解が必ず目的関数を最適にするものではない

- 解決策

  -クラスタリングの問題を個有値問題として定式化する

二値変数におけるクラスタリング

- 簡単のために二つのグループの場合とする

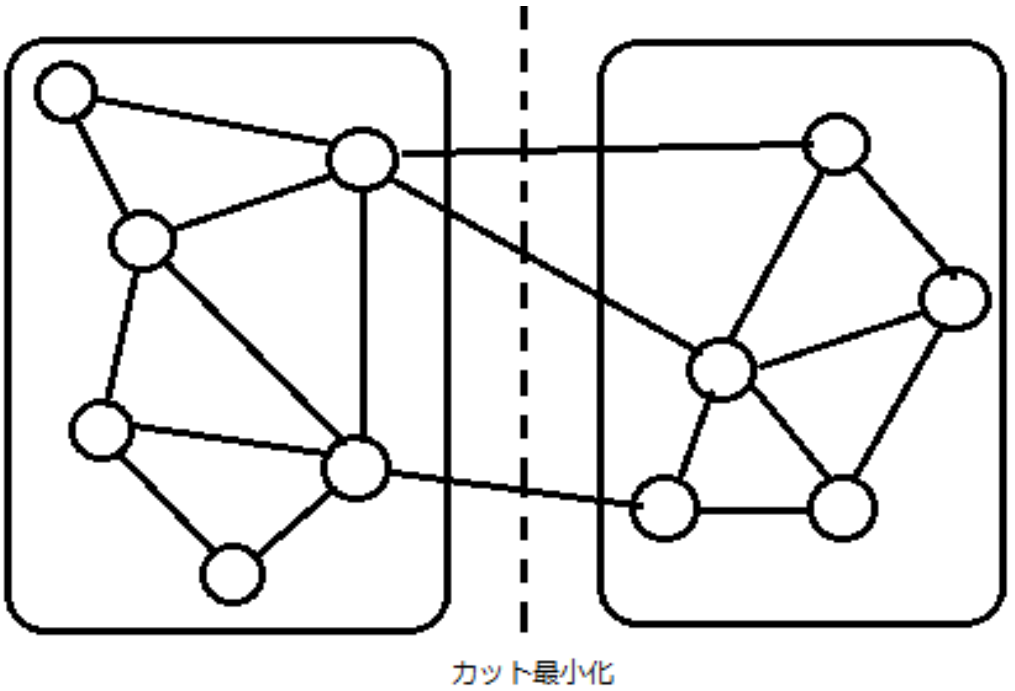
  -クラスタリングは、各サンプル$x^{(i)}$に対してグループに対応するラベル$\beta_i=\pm1$を割り当てる問題となる

- ここで、ラプラシアン固有マップ法で導入したサンプル点から作るグラフ構造を考える

  -各頂点がサンプル点

  -枝にはサンプル点どうしの近さを表す重みがつく

グラフ構造



- このグラフは、近いもの同士が集まっているようになる。
- 二値変数のクラスタリングは、このグラフを二つに分けることと同じ意味
- 最適な分け方は、カットが最小となるように分ける

スペクトラルクラスタリング(1)

- 式で書くと

$$\min_{\beta} \sum_{i,\,j} K_{ij}\left(\beta_i - \beta_j\right)^2 = 2\,\boldsymbol{\beta}^{\boldsymbol{T}}\, P\, \boldsymbol{\beta}$$

ここで、Pはラプラシアン固有マップ法のときの式(3.29)と同じになる。

対角行列Λを $\Lambda_{ii} = \sum_{j=1}^{n} K_{ij}$ と定義すればP=Λ-Kと書ける。

スペクトラルクラスタリング(2)

- βは2値ベクトルであるので、一般的に解くのが困難

  → βを任意の実数ベクトルとする
- このとき、βはいくらでも小さい値をとるので、次の制約をかける

$$\boldsymbol{\beta}^T \Lambda \boldsymbol{\beta} = 1$$

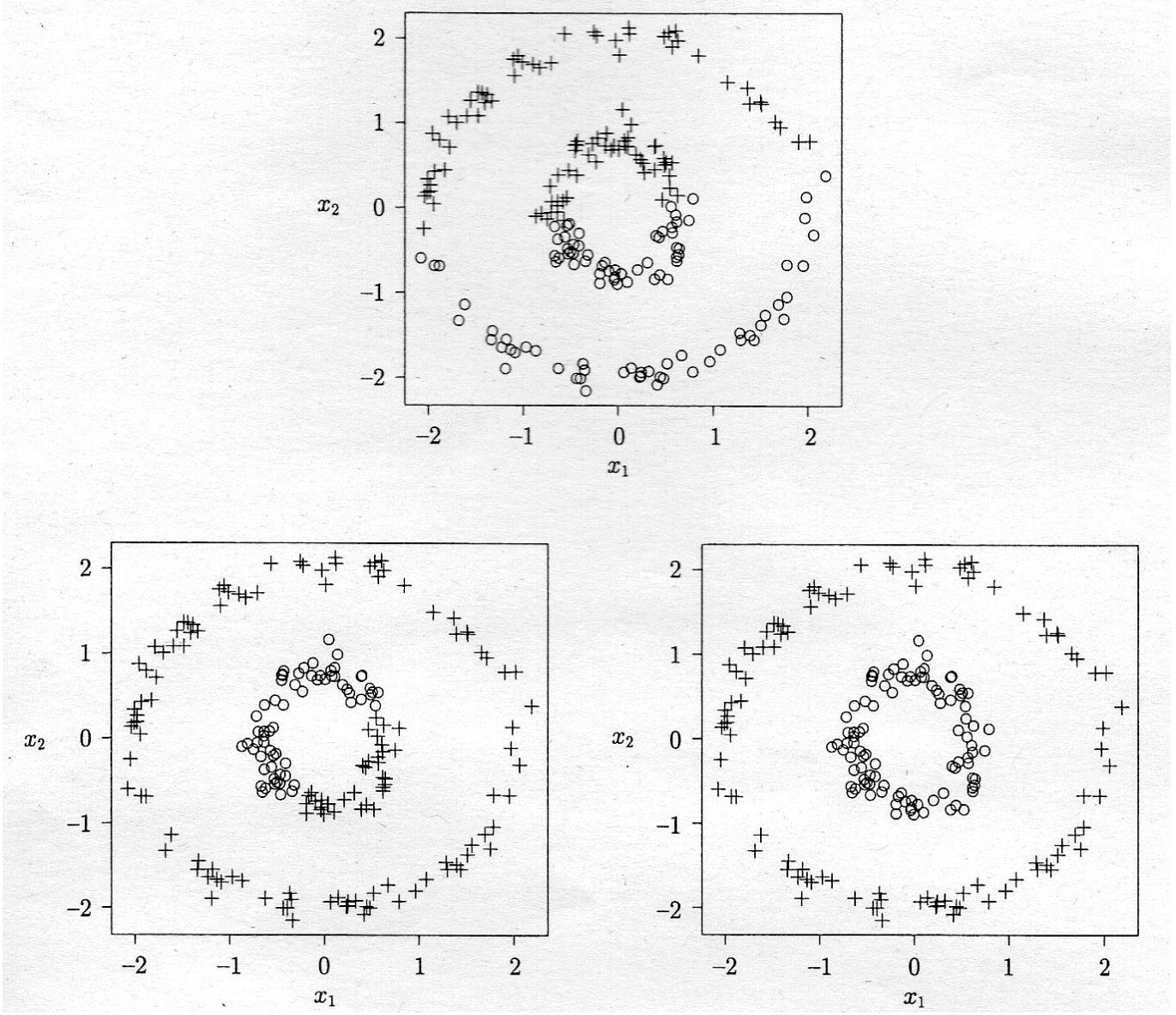
- これは、ラプラシアン固有マップ法と完全に等価な最小化問題となる
- このように,固有値問題を解くことによってクラスタリングを行う手法をスペクトラルクラスタリングという

離散化

- スペクトラルクラスタリングでは、最終的に得られた実数ベクトルを離散化する必要がある。

- 例えば、固有ベクトルの成分$\beta 1,\ldots,\beta n$を並べて、ある閾値で切って、閾値以上と以下のグループに分ける方法がある。

実行例



- 上図：通常のk-平均法
- 左図：カーネルk-平均法
- 右図:スペクトラルクラスタリング