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

3.2 次元圧縮とデータ依存カーネル

茨城大学工学部情報工学科
倉持辰洋

はじめに

- データに依存したカーネルを使った次元圧縮手法
  - いくつかの手法を紹介する
  - いずれの方法も固有値問題を解くことに帰着する
- 入力の空間の構造などをカーネル関数の設計にうまく生かし妥当な結果を得ようとするものである

次元圧縮について

- 「データを次元圧縮する」の意味
  - 与えられたサンプルの低次元表現を求める
    ⇒与えられたサンプルのみ対応
  - 高次元空間から低次元空間への写像を求める
    ⇒サンプル以外の新規データも対応

  微妙に異なる２つの場合がある

両者の等価性(1)

- 実は両者は同じ固有値問題に基づく
- 平均0の場合のカーネル主成分の解は

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_{ji} (\boldsymbol{x}^{(i)}, \boldsymbol{x}), \qquad K \boldsymbol{\alpha}_j = \lambda_j \boldsymbol{\alpha}_j$$

で求められた

これを使ってサンプル $\boldsymbol{x}^{(i)}$ の低次元表現を計算する

両者の等価性(3)

- 古典的なMDSで求める低次元表現と本質的に同じ
- 一方、固有ベクトルの成分を重みとしカーネル関数の重みつきの和を計算すれば低次元空間への写像が得られる

↓　したがって

同じ固有値問題を解くことでどちらの解も求められる

グラフ上の物理モデルに基づく次元圧縮

- ラプラシアン固有マップ法
  - –サンプルからグラム行列を推定して次元圧縮をする最も基本なアルゴリズム
  - –サンプルデータがグラフの頂点
  - –頂点を結ぶ枝が二つのデータの関係

グラフの枝への重みつけ

- 互いに近いデータは大きな、遠いデータは小さな重みとなるようにする

  –データ空間Xが実数ベクトル空間ならガウスカーネルを重みとして取れる

- Iとjを結ぶ枝の重み K_{ij} を成分とする行列をKとする

サンプルを1次元の値に縮約

- サンプルを一次元の値に縮約して表現する

  –i番目のサンプルの表現 β_i を決めるためにデータ間の重みつきの差を小さくすることを考える

  つまり

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

  を解く
- これにより K_{ij} が大きく互いに近いサンプルは近くに配置される

- β_iは定数倍しても本質的に等価である
  その定数倍の自由度を除くため制約
  $$\boldsymbol{\beta}^T \Lambda \boldsymbol{\beta} = 1 \quad \text{をおく}$$
  するとラグランジュ関数が
  $$L(\boldsymbol{\beta}) = \boldsymbol{\beta}^T P \boldsymbol{\beta} - \lambda(\boldsymbol{\beta}^T \Lambda \boldsymbol{\beta} - 1)$$
  となるような最適化問題となる

- $\boldsymbol{\beta}$ で微分して０とおくと

$$P\boldsymbol{\beta} = \lambda \Lambda \boldsymbol{\beta}$$

という一般化固有値問題の最小固有値に対応する固有ベクトルを求めることに帰着する

- 最小の固有値ベクトルから順番に $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots$ と取っていけばよい
- 各ベクトルのl番目の成分を集めたのがラプラシアン固有マップ法におけるサンプルの低次元表現となる

ラプラシアン固有マップほうと カーネル主成分分析の関係

- $P = \Lambda - K$より

$$K\boldsymbol{\beta} = \lambda' \Lambda \boldsymbol{\beta}, \qquad \lambda' = 1 - \lambda$$

λの最小化はこちらの固有値問題では固有値λ'の最大化となる

また、この式はカーネル主成分析の式

$$K\boldsymbol{\alpha} = \lambda \boldsymbol{\alpha}$$

と類似してる

ラプラシアン固有マップ法の特徴

- ラプラシアン固有マップ法はカーネル主成分分析をグラフ上の熱伝導と関連付け補正する
  - この補正により比較的カーネル関数のパラメータに敏感ではなくなってる
- サンプルに対する低次元表現は得られるが写像は得られない
  - 固有値問題にΛが入るため
  - サンプル以外の入力へはトランザクションなどの手法を利用する

多様体上の距離に基づく次元圧縮

- 多様体とは

  –高次元空間の中でデータによくあてはまる曲線や空間といった低次元の部分空間

  ⇒多様体を見つけることが次元圧縮の目的である

ISOMAP

- 多様体の上の距離に着目
  - 距離とカーネルには関係があるので距離の問題をカーネルの問題としてとくことができる
- 抽出した部分空間上で与えられたサンプル点間の距離を利用して元の空間を復元する
  - 点の距離はユークリッド空間を伸縮しなければ変化しない

    ⇒点の距離を保つことができるようなデータの空間配置を求めることで復元できる

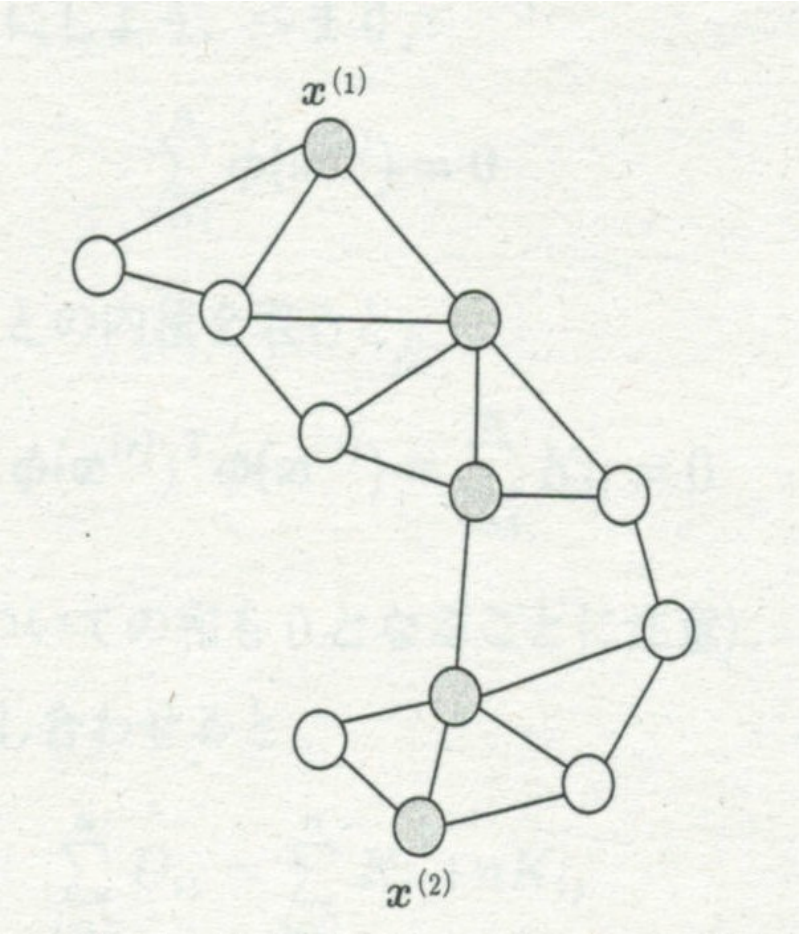
近傍グラフの作成

- 与えられてるのは多様体上のサンプル点のみ
  - サンプル点を端点とする近傍グラフを作成し多様体を近似した骨組み構造とする
- 近傍グラフ作成の条件
  - 前もってサンプル点同士のユーグリッド距離を測る
  - あらかじめ決めたしきい値ε以下
  - k個以下の近傍

    ⇒これらについて点と点を枝で結ぶ

近傍グラフ

- サンプル点から作成するグラフは右図のようになる
- 枝はサンプル点間のユーグリッド距離をあらわす
- 離れた点の多様体上の距離はグラフの最短距離で近似する



距離からカーネルへの変換

- カーネル法は距離と対をなす類似度に基づく方法
  - なので距離を類似度へ変換する必要がある

    ⇒距離と類似度の関係を知る必要がある

線形モデルの貼り合わせによる次元圧縮

- 多様体はどんなに変形していても狭い範囲で見れば線形空間として見れる

  –この性質を利用し次の２ステップからなるアルゴリズムで多様体のあてはめをする

  [1]狭い範囲の点だけを使い低次元の線形モデルをあてはめる

  [2]そのような線形空間をなめらかにつなぎ全体のの多様性を推定する
- この方法を局所線埋め込み法という