データ解析基礎

第14回　多次元データの取扱い

多次元データの扱い

- データ解析の基本はデータの分布を目で見ること.
  - データの次元数が低いときはそのまま図示できる.
  - 次元数が高いときは図示が難しい.
- 高次元データを表示するには次元を減らすことが必要.
  - どうやって次元を減らすか？

次元圧縮の方法

3

- **主成分分析** (principal component analysis: PCA)
  - **データのばらつきが大きい方向を取り出す．**
  - **互いに相関のない特徴量を取り出す．**
  - **パターン認識で特徴抽出でよく使われる．**
- **多次元尺度法** (multi-dimensional scaling: MDS)
  - **データの距離の関係を低次元の空間で表す．類似データを近く，そうでないものを遠くに付置．**
  - **何を類似性を定める要因になっているかを知る．**

主成分分析の考え方

4

- **問題**
  - **多次元データ** x_i **が多数ある（平均は** 0 **とする）．**
  - x_i **をある単位ベクトル** y **の実数倍** $(a_i y)$ **で表したとき，近似誤差の総和が最も小さくなるような** y **を求めたい．**
- **定式化**
  - **次の評価関数を最小化する** a_i **，** y **を求めればよい．**

$$E = \Sigma_i \| x_i - a_i y \|^2$$

主成分分析のステップ 1

5

- $\boldsymbol{y}$ を固定したとき E を最小化する a_i は $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{x}_i$ となる．
  - 誤差関数を

$$E = \Sigma_i \{ \| \boldsymbol{x}_i \|^2 - 2\, a_i\, \boldsymbol{y}^{\mathrm{T}}\, \boldsymbol{x}_i + a_i^2 \}$$

  と展開し，a_i で偏微分すればわかる．

- 上の結果を評価関数に代入すると

$$E = \Sigma_i \| \boldsymbol{x}_i \|^2 - \Sigma_i\, a_i{}^2$$

  - よって，E の最小化は $\Sigma_i\, a_i^2$ の最大化と同値．

主成分分析のステップ 2

6

- さらに，$a_i = \boldsymbol{y}^{\mathrm{T}}\boldsymbol{x}_i$ より最大化すべきものは

$$\boldsymbol{y}^{\mathrm{T}} (\Sigma_i\, \boldsymbol{x}_i\, \boldsymbol{x}_i{}^{\mathrm{T}}) \boldsymbol{y}$$

- $\Sigma_i\, \boldsymbol{x}_i\, \boldsymbol{x}_i{}^{\mathrm{T}}$ は分散行列（対称行列）なので，直交行列 Qを使って対角化が可能．

$$Q^{\mathrm{T}} (\Sigma_i\, \boldsymbol{x}_i\, \boldsymbol{x}_i{}^{\mathrm{T}})\, Q = \mathrm{diag}\, (\lambda_1, \lambda_2, ..., \lambda_n)$$

- $\boldsymbol{y}^{\mathrm{T}} (\Sigma_i\, \boldsymbol{x}_i\, \boldsymbol{x}_i{}^{\mathrm{T}}) \boldsymbol{y}$ を最大化する $\boldsymbol{y}$ は，λ_j の最大値に対応する行列 Q の列ベクトルとなる．

複数の成分の合成による表現

7

- 問題
  - 多数のデータ x_i を二つの単位ベクトル y_1, y_2 の荷重和 $(a_{1i}y_1 + a_{2i}y_2)$ で近似したとき，近似誤差の総和が最小化されるような y_1, y_2 を求めたい．
- 解法
  - 主成分分析を行なって y_1 を求める．
  - データから x_i から y_1 の成分を差し引く．
  - 残った分について主成分分析をすると y_2 が求まる．

複数の成分の合成による表現

8

- 別の考え方
  - 二つの単位ベクトル y_1, y_2 として，$\Sigma_i \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}}$ の固有ベクトルのうち，対応する固有値が最大のものと2番目に大きなものをとればよい．
- 対称行列の固有ベクトルは互いに直交することから，y_1, y_2 は互いに直交する．
- 同様にして三つの単位ベクトルの場合は，上位三つの固有値に対応する固有ベクトルをとればよい．

直交する特徴量の和による表現

9

- 同じことを繰り返すと，直交する単位ベクトル $\{y_k\}$ を用いて，データ x は次のように表せることになる．

$$x = \Sigma_k a_k y_k$$

（a_k を特徴量と呼ぶことがある）

- y_kのうち番号が若いものほど重要度（エネルギー）が大きい．
  - これが「主成分」分析と呼ばれる理由．

主成分分析を行なうときの注意

10

- 事前にデータを標準化しておく．
  - 平均が0，分散が1になるように正規化する．
  - これをしないと，固有ベクトルが分散が大きな方向を向いてしまう．
- 標準化すると，

$\Sigma_i x_i x_i^{\mathrm{T}}$ は対角成分が1の共分散行列になる．

  - これは相関係数を並べたものなので，「相関行列」とよぶ．

主成分分析のまとめ

11

- 相関行列を対角化し，固有ベクトルを求める．
- 固有値が大きなものから順に選ぶ（エネルギー重視）．
- 特徴量を与えるベクトルは互いに直交．
- 特徴出力は無相関．つまり，$\Sigma_j a_{kj} a_{lj} = \delta_{kl}$

  データ: $\boldsymbol{x} = \Sigma_k a_k \boldsymbol{y}_k$ （特徴量の荷重和）

  相関行列: $C = \Sigma_j \boldsymbol{x}_j \boldsymbol{x}^{\mathrm{t}}_j = Y \Lambda Y^{-1}$ （対角化）

多次元尺度法 (multi-dimensional scaling)

12

- 多次元尺度法
  - データの類似度が得られたときに，類似度が高いデータを近くに，低いデータを遠くに付置するような座標を見いだす方法．
  - データ間の類似性を規定している要因を探る．
- 計量MDS
  - 類似度をそのまま用いて分析する．
- 非計量MDS
  - 類似度の大小関係を用いて分析する．

準備

13

- 平均 0 の n 個のm次元データx_iがあるとする．
- データ行列 X
  - x_iを並べてできるm×n行列

$$X = [\, x_1 x_2 \ \ldots \ x_n \,]$$

- 距離行列 D^2
  - 異なるデータの間の距離を

$$d_{ij}^2 = \| x_i - x_j \|^2$$

と決め，これを並べてn×n行列としたもの.

MDSの問題設定

14

- データx_iは知らない．
- しかし，距離行列は知っている．
- このとき，距離行列からデータを復元するにはどうすればよいか？
  - つまり，距離関係を保存するようなデータの値を見つけるにはどうすればよいか？

計量MDS

15

- 距離行列は以下のように分解できる．

$$D^2 = \mathrm{diag}(X^{\mathrm{T}}X)\,\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}} - 2X^{\mathrm{T}}X + \mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}\,\mathrm{diag}(X^{\mathrm{T}}X)$$

  - ここで，$\mathbf{1}_n$ は全要素が1であるn次元ベクトル．

    （$\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$ は全要素が1であるn×n行列になる）

- 中心化行列と呼ばれる行列 J_n を次のように決める．

$$J_n = I_n - (1/\mathrm{n})\,\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}$$

（I_n はn×n単位行列）

計量MDS

16

- D^2 とJ_n を組み合わせて以下の行列 P を作る．

$$P = -1/2 \quad J_n D^2 \; J_n$$

（これをYoung-Householder 変換という）

- このPは以下の性質をもつ．

$$P = \; J_n X^{\mathrm{T}}X\, J_n = X^{\mathrm{T}}X$$

（2番目の等号は x_i の平均が0のときのみ成立）

  - Pの階数はデータの次元数 mに一致する．

計量MDS

17

- P を固有展開すると

  $P = X^{\mathrm{T}}X = Q^{\mathrm{T}}\Lambda Q$　(ただし，Pの rankは m)

  Λ：m×m 行列　Q：m×n 行列

- したがって，Pを標準化してΛ, Qを求めれば，

  $X = \Lambda^{1/2} Q$ として求められることになる．

- ただし，X は一意には決まらない．

  - Xが解であれば，それに回転行列をかけたものも解になる．

非計量MDS

18

- 距離と類似度が線形の関係にない場合を考える．

  - 類似度が順序尺度で与えられていればよい．

- 類似度に変換を加えた上で，座標値 X を求める．

  o_{ij}：与えられた類似度

  δ_{ij}：o_{ij} を変換して得た距離　$\delta_{ij} = F(o_{ij})$

  X：各データに対応するベクトルからなる行列

  $d_{ij}(X)$：X から計算される距離行列

非計量MDS

19

- 以下の評価関数が最小になる変換Fと行列Xを求める．

$$\phi = \Sigma_{i<j} (\delta_{ij} - d_{ij}^2(X))^2 / \Sigma_{i<j} d_{ij}^2(X)$$

where $\delta_{ij} = F(o_{ij})$

- 以下の二つに分けて最適値を求める．
  - X を固定したときに，ϕを最小化するδ_{ij}を求める．
    → Kruscal の単調回帰法
  - X に関して ϕ を最小化する．
    → 最急降下法

最終レポートについて

20

- 締切は2月12日（木）
  - 以降一切受け付けない．
- 本日分が1問：
  - これは必ず解く．
- これまでの復習分が 3 問：
  - それぞれに選択問題が2問ずつある．
- 成績はこれまでのレポートをすべて考慮し判断する．