多変量解析

3.2　次元圧縮とデータ依存カーネル

06t4071f
林華

1. 次元圧縮とカーネル法の等価性

データを次元圧縮する
- 与えられたサンプルの低次元表現を求める
- 高次元空間xから低次元空間への写像を求める

相違点：前者はサンプルのみに着目し、その次元を小さくする
後者は新規データの場合でも低次元に移すことも可能、より一般的

共通点：固有値問題に基づくもので、等価なものである

カーネル主成分分析の解は、低次元空間への写像を次式より求められた

$$f_j(x) = \sum_{i=1}^{n} \alpha_{ji} k\left(x^{(i)}, x\right), \qquad K\alpha_j = \lambda_j \alpha_j$$

1. 次元圧縮とカーネル法の等価性—続き

上式に、サンプル $x^{(l)}$ の低次元表現を計算して見ると

$$f_j(x^{(l)}) = \sum_{i=1}^{n} \alpha_{jl} k\left(x^{(i)}, x^{(l)}\right) = (K\alpha_j)_l = \lambda_j \alpha_{jl}$$

となる。したがって、M個の固有ベクトル $\alpha_1, \alpha_2, \ldots \alpha_M$ のl番目の成分を集めてくればx(l)のM次元表現

$$\beta_l = \left(\lambda_1 \alpha_{1l}, \lambda_2 \alpha_{2l}, \ldots, \lambda_M \alpha_{Ml,}\right)^T$$

がえられる。一方、固有ベクトルの成分を重みとして
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

2 ラプラシアン固有マップ法：グラフ上の物理モデルに基づく次元圧縮

1. グラム行列とグラフ構造

- サンプルデータは有限個の点なので、グラフの頂点に対応付け
- 頂点と頂点を結ぶ枝で二つのデータ間の関連性を表現
- カーネル関数は対称なので、無向グラフで考える

2. カーネル法とグラフ構造との関係

- グラフ構造をデータとみなした時の、グラフとグラフの間のカーネル関数

2. 2　ラプラシアン固有マップ

2. ラプラシアン固有マップ法

- サンプルに対応するグラフの枝に重みをつけることを考える
- 互いに近いデータは大きな重み、遠いデータは小さい重み
- iとjを結ぶ枝の重みKijを成分とする行列Kとする

ここで、サンプルを1次元の値に縮約して表現する場合、データ間の重みつきの差を小さくすることを考え、つまり

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

という問題を解く。より、Kijが大きくて互いに近いサンプル同士は近くに配置される。上式を2次形式で書いたものを

とおく

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

2.2 ラプラシアン固有マップ一続き

ここで対角行列を

$$\Lambda_{ij} = \sum_{j=1}^{n} K_{ij}$$

とおく。各サンプルは定数倍しても
本質的に等価であるため、制約

$$\beta^T \Lambda \beta = 1$$

をおいておく。すると、ラグランジュ
関数が

$$L(\beta) = \beta^T P \beta - \lambda\left(\beta^T \Lambda \beta - 1\right)$$

となるような最適化問題なり、これを
β で微分して0とおくと

$$P\beta = \lambda \Lambda \beta$$

これは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問題
に帰着される

３　ISOMAP：多様体上の距離に基づく次元圧縮
―　次元圧縮と多様体当てはめ

- 次元圧縮の目的は、高次元空間の中でデータによく当てはまる曲線や曲面といった低次元の部分空間を見つめることである。
- 多様体はその部分空間を指す
- 多様体は狭い範囲にはユークリッド空間、広い範囲には曲がった構造
- 多様体は曲がっても、その上に適当な座標系を取ることができる
- ただし、ユークリッド空間での直交座標系とは異なり、一つの座標系で全部をカバーするのではなく、複数のユークリッド空間を貼り合わせて多様体全体を表す

３　ISOMAP：多様体上の距離に基づく次元圧縮
―　多様体上の距離

ISOMAPでは多様体の上の距離に着目して構造の抽出を行う

- 与えられたサンプルがユークリッド空間を伸縮せずに捻じ曲げたような多様体の上にあるとする
- 多様体上の点から点への距離は多様体の上を辿っていける最短経路
- 最短距離は埋め込まれた空間より、直線ではなくなる

与えられた多様体上にあるサンプル点より多様体を表現し最短距離を近似的に求める

1. サンプル点を頂点とする近傍グラフを作る
2. 近くの点までの距離はその直線距離で近似
3. 遠くの点までの距離は近くの点までの最短距離を繋がる
4. グラフ上の2点間の最短経路を動的計画法より求める

３　ISOMAP：多様体上の距離に基づく次元圧縮
―　距離からカーネルへ

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、一般的に距離から類似度への変換を行わなければならない。

特徴ベクトルをユークリッド空間上の点とみなすと

$$\left\|\phi\left(x^{(i)}\right)-\phi\left(x^{(j)}\right)\right\|^2=\left\|\phi\left(x^{(i)}\right)\right\|^2+\left\|\phi\left(x^{(j)}\right)\right\|^2-2\phi\left(x^{(i)}\right)^T\phi\left(x^{(j)}\right)$$

より

$$k\left(x^{(i)},\ x^{(j)}\right)=-\frac{1}{2}\left(\left\|\phi\left(x^{(i)}\right)-\phi\left(x^{(j)}\right)\right\|^2-\left\|\phi\left(x^{(i)}\right)\right\|^2-\left\|\phi\left(x^{(j)}\right)\right\|^2\right)$$

しかし、サンプルと原点の間の距離計算できない

３　ISOMAP：多様体上の距離に基づく次元圧縮
―　距離からカーネルへの続き１

n個のデータ集合の要素の間の距離を要素としてもつ行列Dが与えられる場合、Dのi，j成分の間の距離を次のように表す

$$D_{ij} = \left\| \phi\left(x^{(i)}\right) - \phi\left(x^{(j)}\right) \right\|^2 = K_{ii} + K_{jj} - 2K_{ij}$$

特徴ベクトル全体を平行移動してもお互いの距離変化しないが、内積の値が変化する。

ここで特徴ベクトルのサンプルの平均が0になるようにする

$$\sum_{i=1}^{n} \phi\left(x^{(i)}\right) = 0$$

これとサンプルjとの内積をとると

$$\sum_{i=1}^{n} \phi\left(x^{(i)}\right)^T \phi\left(x^{(j)}\right) = \sum_{i=1}^{n} K_{ij} = 0$$

より、

$$\sum_{i=1}^{n} D_{ij} = \sum_{j=1}^{n} K_{ii} + nK_{jj}$$

３　ISOMAP：多様体上の距離に基づく次元圧縮
―　距離からカーネルへの続き２

jについても同様、よって

$$\sum_{i=1}^{n} D_{ij} = nK_{ii} + \sum_{j=1}^{n} K_{jj}$$

またすべての総和は

$$\sum_{i=1}^{n}\sum_{j=1}^{n} D_{ij} = 2n\sum_{i=1}^{n} K_{ii}$$

これらの式を用いて距離からカーネル式に

$$-D_{ij} + \frac{1}{n}\sum_{i'=1}^{n} D_{i'j} + \frac{1}{n}\sum_{j'=1}^{n} D_{ij'} - \frac{1}{n^2}\sum_{i'=1}^{n}\sum_{j'=1}^{n} D_{i'j'} = 2K_{ij}$$

4局所線形埋め込み法:
― 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も狭い範囲で見れば線形空間とみなせる

- 狭い範囲の点だけを使って低次元の線形モデルを当てはめる
- そのような線形空間を滑らかにつなぐことより全体の多様体を推定する

最初のステップ:近傍の点の間の線形関係を見付ける

$$\min_{W} \left\| x^{(i)} - \sum_{j \in \mathrm{N}_i} W_{ij} x^{(j)} \right\|^2$$

という最小化問題を解いてWijを求める
次に各1次元にあるすべての近傍系について最小化をする

$$\min_{\beta} \sum_{i=1}^{n} \left(\beta_i - \sum_{j \in \mathrm{N}_i} W_{ij} \beta_j \right)^2$$

4局所線形埋め込み法:
ー　線形モデルの貼り合わせによる次元圧縮ー続き

上式のラグランジュ関数は

$$\|(I-W)\beta\|^2-\lambda\left(\|\beta\|^2-1\right)$$

とかける。β で微分すると

$$(I-W)^T(I-W)\beta=\lambda\beta$$

という固有値問題に帰着される
この固有値の最小化は

$$\left(W+W^T-W^TW\right)\beta=(1-\lambda)\beta$$

と書けるので

$$\tilde{K}=W+W^T-W^TW$$

いう行列の固有値最大化と等価に
なる