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
3.1 カーネル主成分分析

茨城大学工学部

菊池裕紀

はじめに

- 主成分分析(PCA)

  ー代表的な多変量解析の手法

高次元空間のデータを圧縮するために、そのデータを低次元空間に射影する。
データの低次元構造を抽出することができる。

↓　拡張

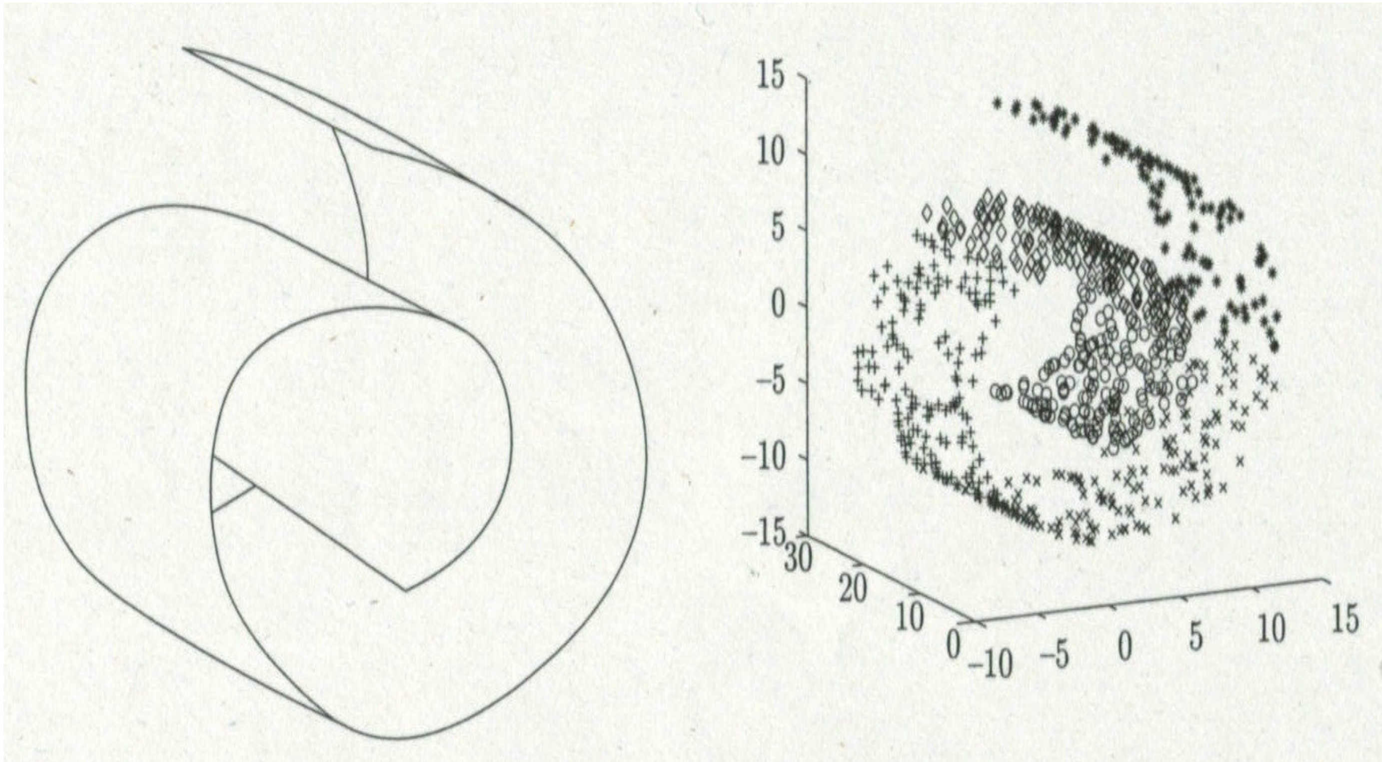
カーネル法の枠組みに！！

低次元構造の抽出と情報量

- 主成分分析を行う際の規準
  - ー低次元に射影したとき、ばらつき（分散）ができるだけ大きくなるようにする。
  - ー縮小したデータの近似誤差（二乗誤差）ができるだけ小さくなるようにする。

この規準を元に、もとのデータの情報量をできるだけ保つような低次元構造を抽出していく。

カーネル主成分分析と固有値問題



適切な特徴抽出により十分高い次元に移せば
線形の構造として捉えることができる。

例えば……

$x=(x_1, x_2)^T$という2次元の変数で作られる

$$a_1 x_1 + a_2 x_2^2 + a_3 x_2 + a_4 = 0$$

これは放物線を表している。ここで、

$$\varphi_1(\boldsymbol{x}) = x_1, \varphi_2(\boldsymbol{x}) = x_2^2, \varphi_3(\boldsymbol{x}) = x_2$$

という変換を考えると、

$$a_1 \varphi_1(\boldsymbol{x}) + a_2 \varphi_2(\boldsymbol{x}) + a_3 \varphi_3(\boldsymbol{x}) + a_4 = 0$$

３次元空間の平面の方程式に落としこめた。

高次元の特徴ベクトルに変換してから通常の主成分分析を行う。

高次元の特徴ベクトルに変換してから、通常の主成分分析を行って、低次元の線形部分空間を求める。

ーカーネル主成分分析

- 平均0の場合

特徴ベクトル $\varphi(\boldsymbol{x})$ を1次元の直線上に射影

関数　$f(\boldsymbol{x}) = \boldsymbol{w}^T \varphi(\boldsymbol{x})$ を得る。

制約：**w**は単位ベクトルであり $\|\boldsymbol{w}\|^2 = 1$ を満たす。

$$E_n[\varphi(\boldsymbol{x})] = \frac{1}{n}\sum_{i=1}^{n} \varphi(\boldsymbol{x}^{(i)}) = 0$$

簡単のために $E_n = 0$ とする。

ここで、射影した点の分散は…

$$Var_n[f(\boldsymbol{x})] = \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{w}^{\boldsymbol{T}}\varphi\left(\boldsymbol{x}^{(\boldsymbol{i})}\right)\right)^2$$

これを制約下で最大化する問題を解けばよい。

ーラグランジュ未定乗数法

$$L(\boldsymbol{w}) = -Var_n[f(\boldsymbol{x})] + \lambda\left(\|\boldsymbol{w}\|^2 - 1\right)$$

↓ wで微分

$$-\frac{2}{n}\sum_{i=1}^{n}\left(\boldsymbol{w}^{\boldsymbol{T}}\varphi\left(\boldsymbol{x}^{(\boldsymbol{i})}\right)\right)\varphi\left(\boldsymbol{x}^{(\boldsymbol{i})}\right) + 2\lambda\boldsymbol{w} = 0$$

ここで $\lambda \neq 0$ ならば、$\alpha_i = \boldsymbol{w}^T \varphi(\boldsymbol{x}^{(i)})/(n\lambda)$ とおいて…

$$\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i \varphi(\boldsymbol{x}^{(i)})$$

すなわち…

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i \varphi(\boldsymbol{x}^{(i)})^T \varphi(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}^{(i)}, \boldsymbol{x})$$

分散は導入したαを用いて、

$$Var_n[f(\boldsymbol{x})] = \frac{1}{n} \sum_{l=1}^{n} \left(\sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(l)}) \right)^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \alpha_i \alpha_j k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(l)}) k(\boldsymbol{x}^{(j)}, \boldsymbol{x}^{(l)})$$

$$= \frac{1}{n} \alpha^T K^2 \alpha$$

L(w)をαを用いて表すと、

$$L(\boldsymbol{\alpha}) = -\frac{1}{n}\boldsymbol{\alpha}^{T}K^{2}\boldsymbol{\alpha} + \lambda\left(\boldsymbol{\alpha}^{T}K\boldsymbol{\alpha} - 1\right)$$

↓αで微分

$$-\frac{2}{n}K^{2}\boldsymbol{\alpha} + 2\lambda \mathrm{K}\boldsymbol{\alpha} = 0$$

Kが正則であるとすれば、逆行列をかけてグラム行列の固有値問題の形にすることができる。

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

よってL(α)に上の式を代入すると…

$$L(\boldsymbol{\alpha}) = -\lambda$$

Lを最小にする

ーλができるだけ大きい値をとるとき！

– Kの最大固有値のとき

制約条件: $\|\boldsymbol{w}\|^2=1$ より $\boldsymbol{\alpha}^T K \boldsymbol{\alpha}=1$

これを満たすようにαを決めれば良い。

- 一般の場合

これまでは、サンプル平均 $E_n=0$ としてきた。

そうでない場合…

[分散=二乗平均-平均の2乗]

f(x)のサンプル平均は…

$$E_n[f(\boldsymbol{x})] = \frac{1}{n}\sum_{l=1}^{n} f(\boldsymbol{x}^{(l)}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{n} \alpha_i k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(l)}) = \frac{1}{n}\alpha^T K\mathbf{1}$$

次に平均の２乗を求める。

$$E_n[f(\boldsymbol{x})]^2 = \frac{1}{n^2}(\boldsymbol{\alpha}^T K\mathbf{1})^2 = \frac{1}{n^2}(\boldsymbol{\alpha}^T K\mathbf{1})(1^T K\boldsymbol{\alpha})$$

よって分散は、 $Var_n[f(\boldsymbol{x})] = \frac{1}{n}\boldsymbol{\alpha}^T K j_n K\boldsymbol{\alpha}$

（ただし、$J_N = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ ）

- 以下の固有値問題を解くことに帰着される。

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

- 以上をアルゴリズムの形でまとめる

[1]データ点の集合 $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \cdots, \boldsymbol{x}^{(n)}$ から、グラム行列を作る。

[2] $J_n K \boldsymbol{\alpha} = \lambda \boldsymbol{\alpha}$ という固有値問題を解く。

[3]固有値の大きい順に非線形変換を構成する。

カーネル主成分分析の問題点とデータ依存カーネル

- 問題点

  ーこれまで話てきたものはすべて特徴ベクトルの空間での話。

  ・分散や二乗誤差といっているのはもとの世界の誤差と は一般的に異なっている。

  カーネル関数はデータを超高次元に移してから低次元に抽出する。高次元に移したとき、見かけ上のばらつきを拾ってしまう。

  カーネル関数の選び方（特徴ベクトルの選び方）で

  結果が異なってくる！

サンプルと同じサイズの任意の正定値行列K
が与えられたとき…

→特徴ベクトル $\varphi\left(\boldsymbol{x}^{(i)}\right)$ が存在。

カーネル関数の定めるグラム行列Kになる。

↓勝手に正定値行列を作る。

データに応じて正定値行列を設計できる！

＊ データとは無関係にグラム行列を作ってしまうことにもなる。

(1)トランスダクション

サンプルとして与えられていない入力 $\boldsymbol{x}^{new}$ の取り扱い。

→グラム行列だけを計算するアプローチでは $k\left(\boldsymbol{x}^{new}, \boldsymbol{x}^{(i)}\right)$ の関数を計算できない。

↓対処

サンプル点以外の情報も学習時に加える。

（トランスダクション）

しかし、、、

サンプルが有限集合のときは良いが、そうでなければ学習にかかる時間が大幅に増大してしまう可能性がある。

(2)類似度と正定値行列

カーネル関数→特徴ベクトルの間の内積

特徴ベクトルどうしの類似度や近さを表している

類似度行列が正定値→カーネル関数

(＊正定値行列とカーネル関数の等価性)

・一般には…

類似度行列は正定値とは限らない。

→正定値行列を求めなければならない!!

(詳しくは5章2節で！！)