カーネル法入門
1. カーネル法へのイントロダクション

福水健次

統計数理研究所／総合研究大学院大学

大阪大学大阪大学大学院基礎工学研究科・集中講義

2014 September

カーネル法:

近年 (1990年代半ばごろから) 発展したデータ解析の方法論. 非線形な情報や高次モーメントの扱いが容易.

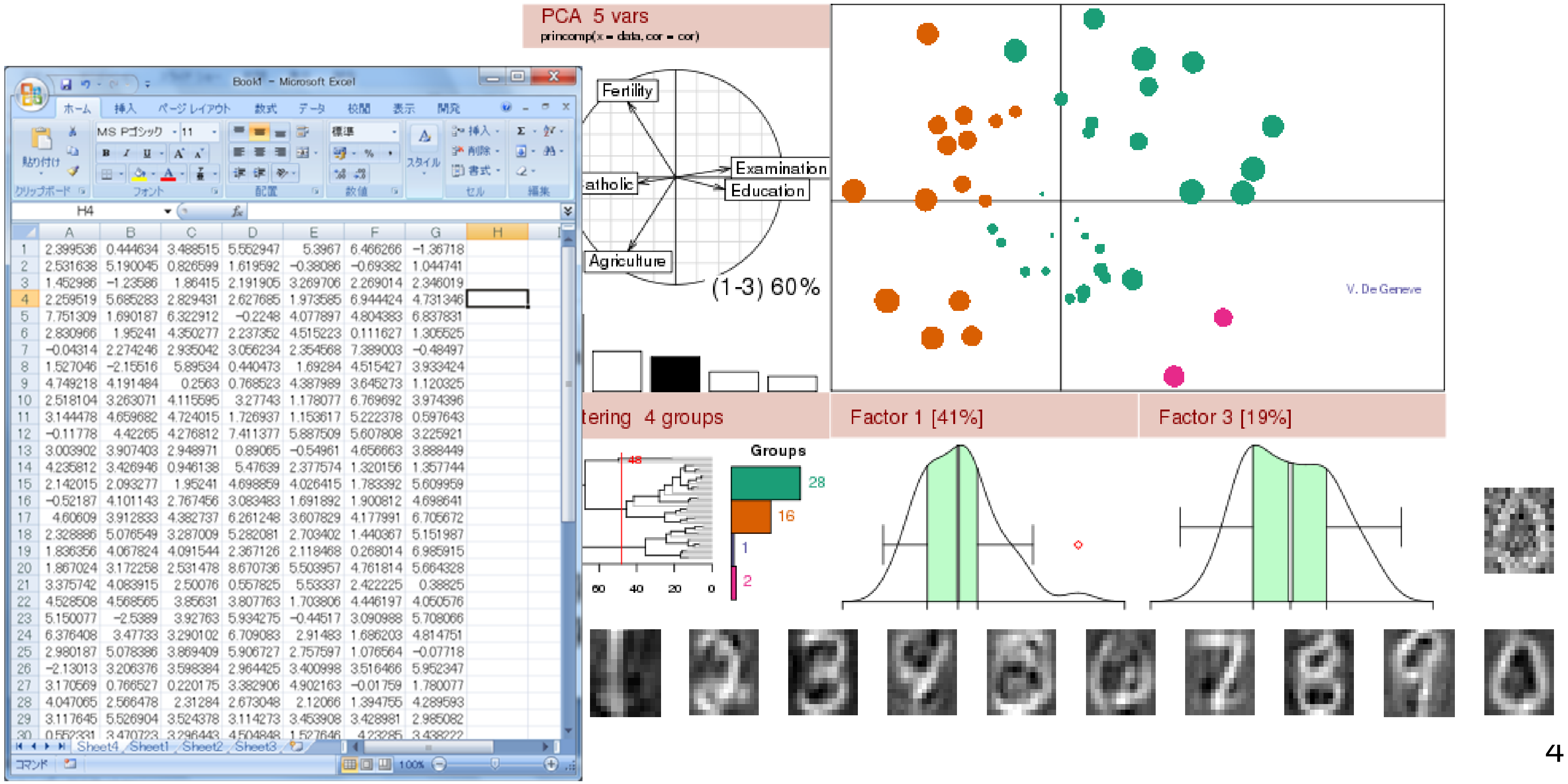
サポートベクターマシンの提案が発端となった.

線形なデータ解析，非線形なデータ解析

データ解析とは?

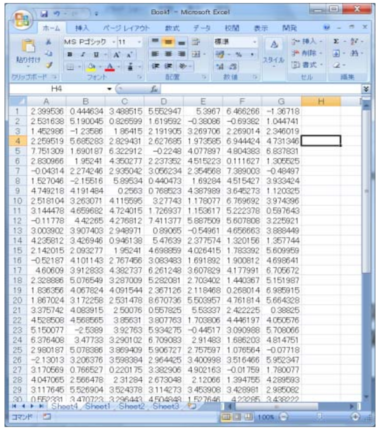
Analysis of data is a process of inspecting, cleaning, transforming, and modeling data with the goal of highlighting useful information, suggesting conclusions, and supporting decision making.

– *Wikipedia*



線形なデータ解析

– 数値の表



行列 表現

$$\mathbf{X} = \begin{pmatrix} X_1^{(1)} & \Lambda & X_m^{(1)} \\ X_1^{(2)} & \Lambda & X_m^{(2)} \\ \mathrm{M} & & \mathrm{M} \\ X_1^{(N)} & \Lambda & X_m^{(N)} \end{pmatrix}$$

m 次元
N 個のデータ

– 線形代数を使ってデータ解析を行う.

- 相関,
- 主成分分析(Principal component analysis, PCA),
- 正準相関分析(Canonical correlation analysis, CCA), etc.
- 線形回帰,
- 線形判別分析
- ロジスティック回帰

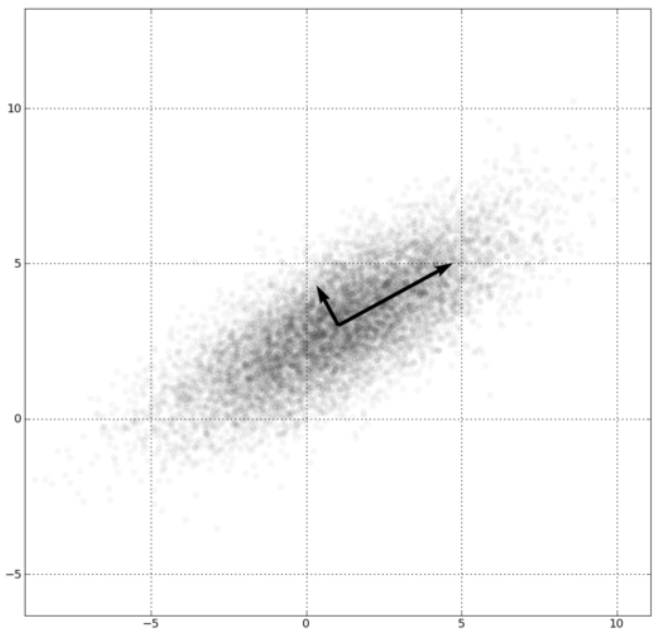
■ 例1: 主成分分析(Principal component analysis, PCA)

PCA: 分散が最大となる低次元部分空間にデータを射影する. .

1st direction = $\mathrm{argmax}_{\|a\|=1} \mathrm{Var}[a^T X]$

$$\mathrm{Var}[a^T X] = \frac{1}{N}\sum_{i=1}^{N}\left\{a^T\left(X^{(i)} - \frac{1}{N}\sum_{j=1}^{N} X^{(j)}\right)\right\}^2$$

$$= a^T V_{XX} a.$$



$$V_{XX} = \frac{1}{N}\sum_{i=1}^{N}\left(X^{(i)} - \frac{1}{N}\sum_{j=1}^{N} X^{(j)}\right)\left(X^{(i)} - \frac{1}{N}\sum_{j=1}^{N} X^{(j)}\right)^T$$

--- X の分散共分散行列

– 第p主成分方向

$= u_p$ ： V_{XX} の第p最大固有値に対する単位固有ベクトル

PCA ⇨ 行列 $V_{XX} = XX^T - \bar{X}\bar{X}^T$ の固有値問題

■ 例2: 線形識別(判別)

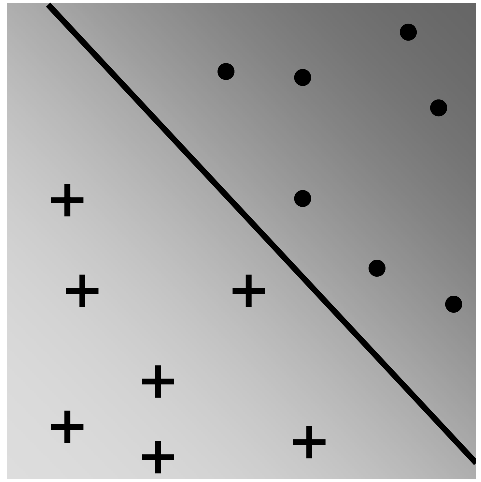
– ２値識別

入力

$$\mathbf{X} = \begin{pmatrix} X_1^{(1)} & \Lambda & X_m^{(1)} \\ X_1^{(2)} & \Lambda & X_m^{(2)} \\ \mathrm{M} & & \mathrm{M} \\ X_1^{(N)} & \Lambda & X_m^{(N)} \end{pmatrix}$$

クラスラベル

$$Y = \begin{pmatrix} Y^{(1)} \\ Y^{(2)} \\ \mathrm{M} \\ Y^{(N)} \end{pmatrix} \in \{\pm 1\}^N$$



識別器

$$h(x) = \mathrm{sgn}(a^T x + b)$$

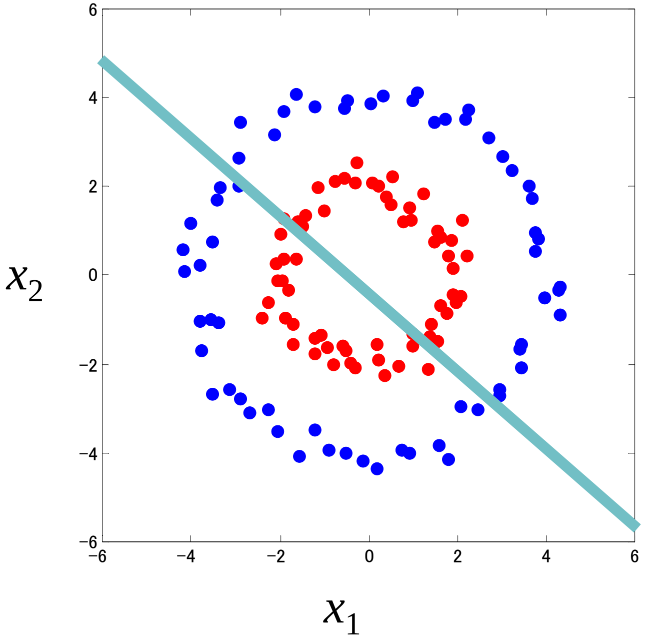
を次にようにに構成する

$$h(X^{(i)}) = Y^{(i)} \qquad \text{for all (or most) } i.$$

– 例: Fisherの線形判別分析, 線形サポートベクターマシン, etc.

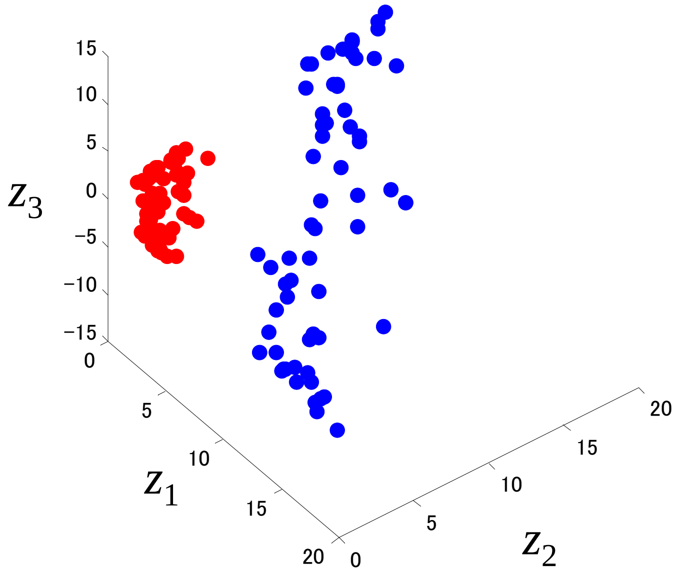
線形で十分か?

線形識別不能



transform

線形識別可能

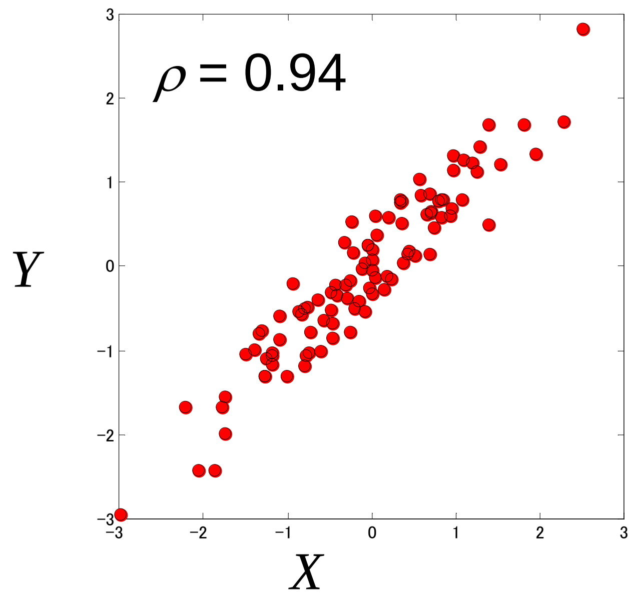


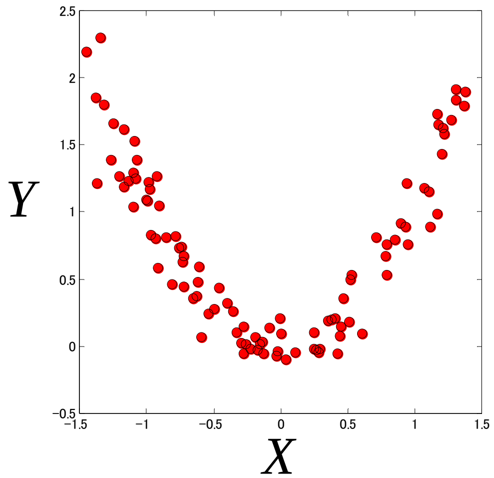
$$(z_1, z_2, z_3) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

Watch the movie! *https://www.youtube.com/watch?v=3liCbRZPrZA*

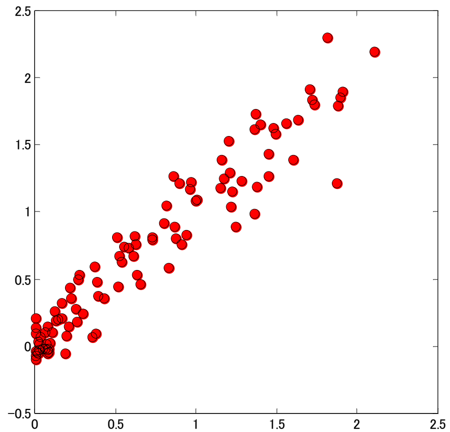
■ Another example: correlation

$$\rho_{XY} = \frac{Cov[X,Y]}{\sqrt{Var[X]Var[Y]}} = \frac{E\left[\left(X - E[X]\right)\left(Y - E[Y]\right)\right]}{\sqrt{E\left[\left(X - E[X]\right)^2\right]E\left[\left(Y - E[Y]\right)^2\right]}}$$





$\rho(X, Y)$
$= 0.17$



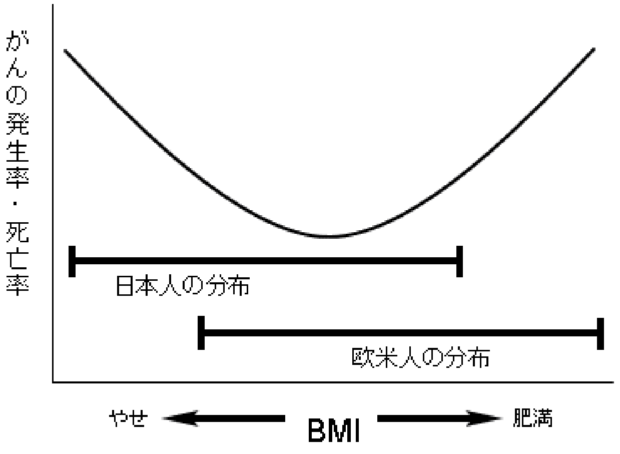
$\rho(X^2, Y)$
$= 0.96$

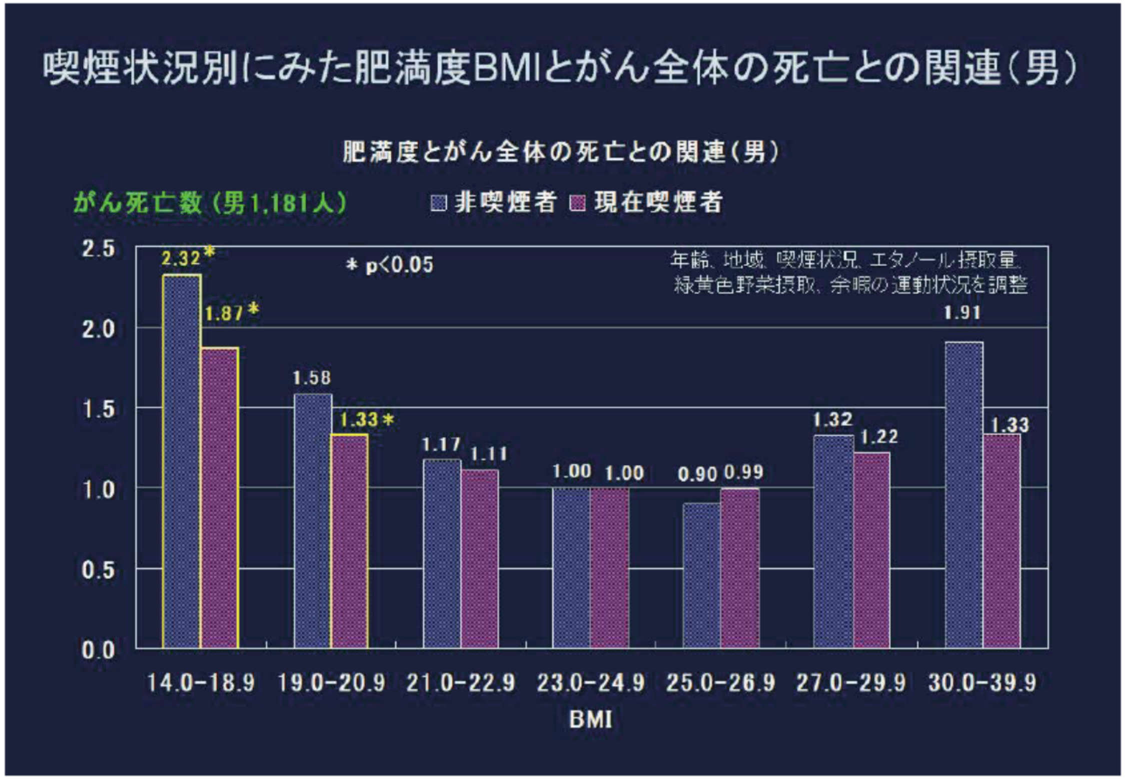
(X, Y)

transform

(X^2, Y)

肥満とがんとの関係





がん研　多目的コホート研究（JPHC Study）
肥満度（BMI）のがん全体の罹患に与える影響

非線形変換は有望

Analysis of data is a process of inspecting, cleaning, transforming, and modeling data with the goal of highlighting useful information, suggesting conclusions, and supporting decision making.

Wikipedia.

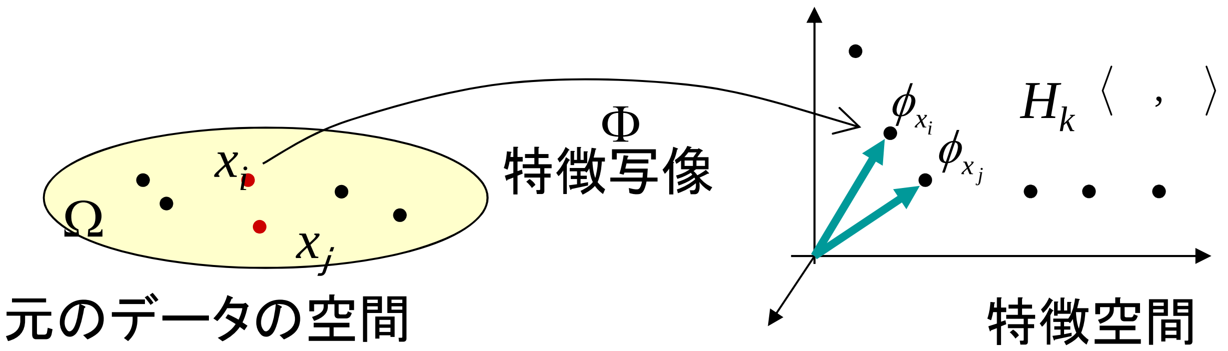
カーネル法 = データの非線形情報，高次モーメントを抽出するために，データを高次元の特徴空間に写像する方法論.

カーネル法の要点


THE KERNEL
BREWERY
LONDON
PALE ALE
CENTENNIAL
22.12.09
5.7% ABV

カーネル法の概略

– カーネル法の概念図



特徴空間で線形データ解析を施す!

e.g. SVM

– 特徴空間として望まれる性質:

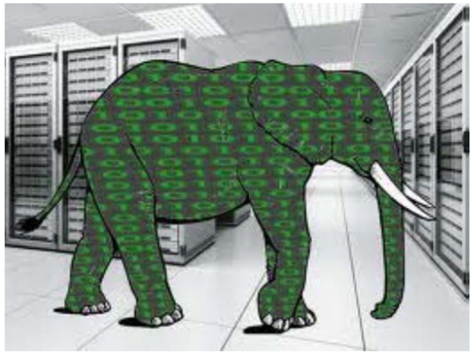
- データのさまざまな非線形特徴を有していること
- 内積計算が容易にできること.
  多くの線形データ解析の計算は内積に依拠している.

■ 計算の問題

– 高次情報の抽出

$$(X, Y, Z) \rightarrow (X, Y, Z, X^2, Y^2, Z^2, XY, YZ, ZX, \ldots)$$

– 元の空間の次元が高いと 計算は実現できない!



e.g. 10000 次元のデータ,
２次までの特徴

$${}_{10000}C_1 + {}_{10000}C_2 = 50{,}005{,}000$$

– 計算量爆発.
より効率的な方法が必要 → カーネル法

特徴空間と正定値カーネル

– 特徴写像:　元の空間から特徴空間への写像

$$\Phi:\ \Omega \to H, \qquad x \mapsto \Phi(x)$$

– 特別な特徴空間（再生核ヒルベルト空間）を用いると，特徴ベクトルの内積計算が関数値 (正定値カーネル) $k(x,y)$の評価に置き換えられる

$$\langle \Phi(X_i), \Phi(X_j) \rangle = k(X_i, X_j)$$

kernel trick

– 内積計算さえできれば，特徴ベクトル Φ(X).の陽な形は知らなくてもよい.

正定値カーネル

<u>定義.</u>

Ω: 集合

カーネル $k: \Omega \times \Omega \to R$ が正定値であるとは

1) (対称性) $$k(x, y) = k(y, x)$$

2) (正値性) 任意の点 $x_1, \ldots, x_n \in \Omega$ $(\forall n)$ に対し,

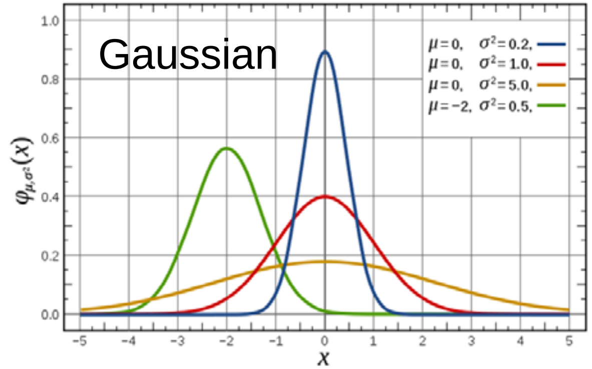
(Gram行列) $$\begin{pmatrix} k(x_1, x_1) & \Lambda & k(x_1, x_n) \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ k(x_n, x_1) & \Lambda & k(x_n, x_n) \end{pmatrix}$$ が半正定値

i.e., $\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$ for any $c_i \in \mathbf{R}$

– 例: $\mathbf{R}^m$上

- Euclid内積

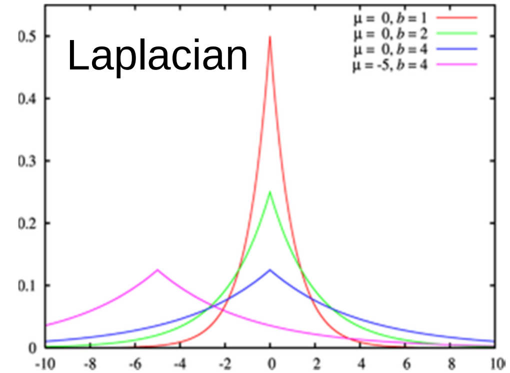
$$k(x,y) = x^T y$$



- Gaussian RBF カーネル

$$k_G(x,y) = \exp\left(-\|x-y\|^2 / \sigma^2\right) \qquad (\sigma > 0)$$

- Laplace カーネル



$$k_L(x,y) = \exp\left(-\alpha \sum_{i=1}^{m} |x_i - y_i|\right) \qquad (\alpha > 0)$$

- 多項式カーネル

$$k_P(x,y) = (c + x^T y)^d \qquad (c > 0, d \in \mathbf{N})$$

命題1.1

H を内積$\langle\cdot,\cdot\rangle$を持つベクトル空間とし，$\Phi\colon \Omega \to H$ を写像(特徴写像)とする. $k\colon \Omega \times \Omega \to \mathbf{R}$ を

$$k(x,y) = \langle \Phi(x), \Phi(y) \rangle, \qquad \text{(kernel trick)}$$

により定義すると，$k(x,y)$ は正定値である.

– カーネルトリックを成り立たせる関数は，正定値カーネルである.

*Proof)

$$\sum_{i,j=1}^{n} c_i c_j k(X_i, X_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \left\langle \Phi(X_i), \Phi(X_j) \right\rangle$$

$$= \left\langle \sum_{i=1}^{n} c_i \Phi(X_i), \sum_{j=1}^{n} c_j \Phi(X_j) \right\rangle = \left\| \sum_{i=1}^{n} c_i \Phi(X_i) \right\|^2 \geq 0$$

– 正定値性は十分でもある.

定理1.2 (Moore-Aronszajn)

Ω上の正定値カーネル k に対し，Ω上の関数からなるHilbert空間* H (再生核ヒルベルト空間, RKHS) が存在して，次が成り立つ.

1) $k(\cdot, x) \in H$ $(\forall x \in \Omega)$.

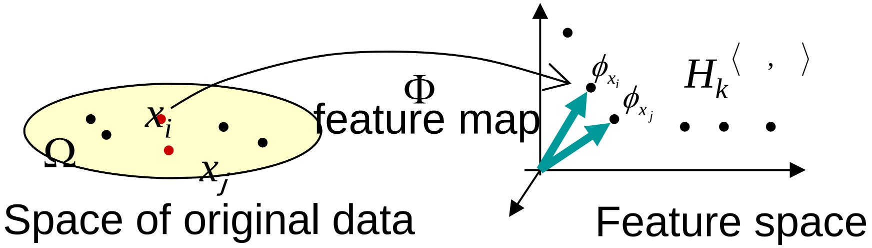
2) span $\{ k(\cdot, x) \mid x \in \Omega \}$ はHで稠密

3) (再生性)

$$\langle f, k(\cdot, x) \rangle = f(x) \qquad \text{for any } f \in H, x \in \Omega.$$

*Hilbert空間: 内積を持つベクトル空間で，内積により決まるノルムが完備であるもの.

正定値カーネルによる特徴写像



– 正定値カーネル k を用意
– 特徴空間 = RKHS
– 特徴写像:

$$\Phi: \ \Omega \to H, \qquad x \mapsto k(\cdot, x)$$

$$X_1, \dots, X_n \mapsto \ k(\cdot, X_1), \dots, k(\cdot, X_n)$$

– カーネルトリック(再生性):

$$\left\langle \Phi(X_i), \Phi(X_j) \right\rangle = k(X_i, X_j),$$

– 正定値カーネルを与えれば十分.
  - 特徴写像, 特徴ベクトルを陽に知る必要はない.
  - カーネル法の計算は, グラム行列 $\left(k(X_i, X_j)\right)_{ij}$ による計算となる.

カーネル法の例：カーネルPCA

PCAからカーネルPCAへ

- PCA: 線形な次元削減.
- カーネルPCA: 非線形な次元削減 (Schölkopf et al. 1998).

- 特徴空間でPCAを行う

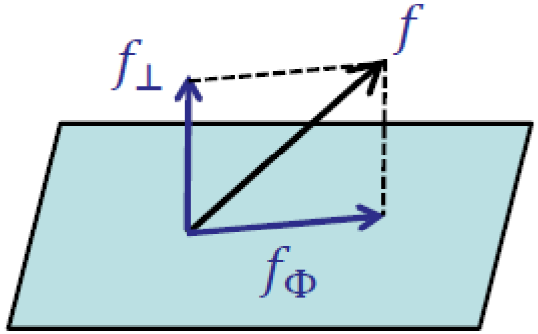
$$\max_{\|a\|=1}: \quad \mathrm{Var}[a^T X] = \frac{1}{N}\sum_{i=1}^{N}\left\{ a^T \left(X^{(i)} - \frac{1}{N}\sum_{j=1}^{N} X^{(j)} \right) \right\}^2$$

$$\Downarrow$$

$$\max_{\|f\|=1}: \quad \mathrm{Var}[\langle f, \Phi(X) \rangle] = \frac{1}{N}\sum_{i=1}^{N}\left\{ \left\langle f, \Phi(X^{(i)}) - \frac{1}{N}\sum_{j=1}^{N} \Phi(X^{(j)}) \right\rangle \right\}^2$$

次の形の f を考えれば十分

$$f=\sum_{i=1}^{N} c_i \left(\Phi(X^{(i)})-\tfrac{1}{N}\sum_{j=1}^{N}\Phi(X^{(j)})\right)$$



(直交する方向は分散に効いてこない!)　[Representer定理]

⇒ (カーネルトリックを使うと)

$$\max \quad \mathrm{Var}[\langle f, \Phi(X)\rangle] = \frac{1}{N} c^T \tilde{K}_X^2 c$$

subject to $\quad \| f \| = 1 \quad \Leftrightarrow \quad c^T \tilde{K}_X c = 1$

$$\left(\tilde{K}_X\right)_{ij} = k(X^{(i)}, X^{(j)}) - \frac{1}{N}\sum_{b=1}^{N} k(X^{(i)}, X^{(b)}) - \frac{1}{N}\sum_{a=1}^{N} k(X^{(a)}, X^{(j)}) + \frac{1}{N^2}\sum_{a,b=1}^{N} k(X^{(a)}, X^{(b)})$$

(中心化Gram行列)

– 証明

- $f = \sum_{i=1}^{N} c_i \widetilde{\Phi}(X^i)$ とすると, $\qquad [\ \widetilde{\Phi}(X^i) \equiv \Phi(X^i) - \frac{1}{N}\sum_{a=1}^{N} \Phi(X^a)\ ]$

$$\mathrm{Var}[\langle f, \Phi(\mathbf{X})\rangle] = \frac{1}{N}\sum_{s=1}^{N} \langle f, \widetilde{\Phi}(X^s)\rangle^2$$

$$= \frac{1}{N}\sum_{s=1}^{N} \left\langle \sum_i c_i \widetilde{\Phi}(X^i), \widetilde{\Phi}(X^s)\right\rangle^2$$

$$= \frac{1}{N}\sum_{s=1}^{N} \left\{\sum_i c_i \underbrace{\boxed{\left\langle \widetilde{\Phi}(X^i), \widetilde{\Phi}(X^s)\right\rangle}}_{\equiv \widetilde{K}_{is}}\right\}^2$$

$$= \frac{1}{N}\sum_{s=1}^{N} \left\{\sum_i c_i \widetilde{K}_{is}\right\}^2 = \frac{1}{N}\, c^T \widetilde{K}^2 c.$$

- $\|f\|^2 = \left\langle \sum_{i=1}^{N} c_i \widetilde{\Phi}(X^i), \sum_{j=1}^{N} c_j \widetilde{\Phi}(X^j)\right\rangle$

$$= \sum_{ij} c_i \widetilde{K}_{ij} c_j = c^T \widetilde{K} c.$$

- $\widetilde{K}_{ij} = \left\langle \left(\Phi(X^i) - \frac{1}{N}\sum_{a=1}^{N} \Phi(X^a)\right), \Phi(X^j) - \frac{1}{N}\sum_{b=1}^{N} \Phi(X^b)\right\rangle$

$$= k(X^i, X^j) - \frac{1}{N}\sum_a k(X^i, X^a) - \frac{1}{N}\sum_b k(X^j, X^b) + \frac{1}{N^2}\sum_{a,b} k(X^a, X^b)$$

■ カーネルPCAのアルゴリズム:

- 中心化Gram行列 $\widetilde{K}_X$ の計算
- $\widetilde{K}_X$ の固有分解 $\widetilde{K}_X = \sum_{i=1}^{N} \lambda_i u_i u_i^T$

$$\lambda_1 \geq \lambda_2 \geq \Lambda \geq \lambda_N \geq 0$$ eigenvalues

$$u_1, u_2, \mathrm{K}, u_N$$ unit eigenvectors

- 第p主成分方向 $f_p = \sum_j \frac{1}{\sqrt{\lambda_p}} u_{pj} \widetilde{\Phi}(X^{(j)})$,

$\widetilde{\Phi}(X^{(j)}) = \Phi(X^{(j)}) - \frac{1}{N}\sum_{b=1}^{N} \Phi(X^{(b)})$: 中心化特徴ベクトル

- $X^{(i)}$の第p主成分 $= \langle f_p, \widetilde{\Phi}(X^{(i)}) \rangle$

$$= \sum_j \frac{1}{\sqrt{\lambda_p}} u_{pj} \widetilde{K}_{ji}$$

$$= \sqrt{\lambda_p} u_{pi}$$

カーネルPCAの例

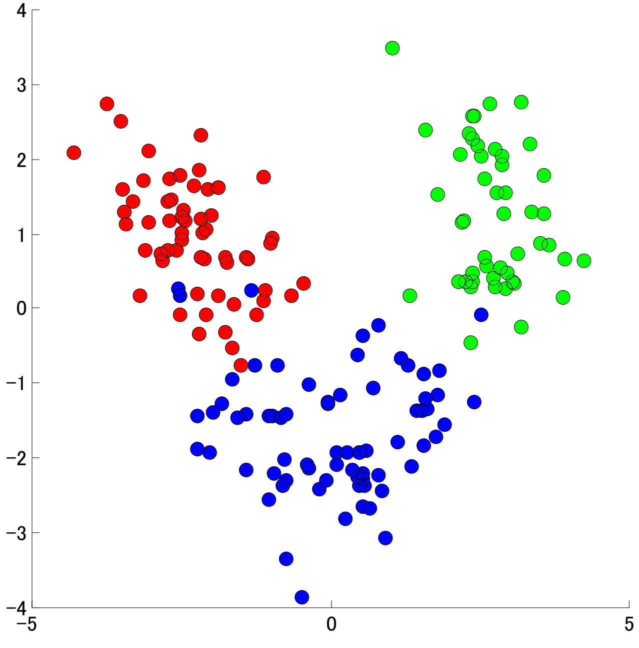
- Wine データ (UCI repository)

  3種類のイタリアワインに関する，13 次元の化学測定値

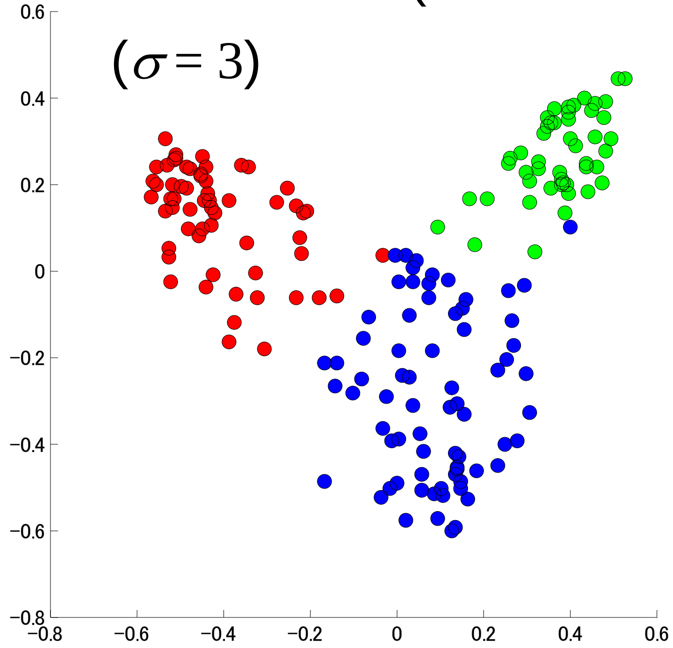
  178 データ.

  クラスの情報はカーネルPCAには用いていない

Linear PCA



Kernel PCA (Gaussian kernel)



カーネル法の構成要素

- 特徴空間上で線形データ解析を適用する. (Kernelization)
- 多くの場合，目的関数をカーネルによって書き直すことができる

$$\langle \Phi(X_i), \Phi(X_j) \rangle = k(X_i, X_j)$$

$$\langle f, \Phi(X_i) \rangle$$

- 解は以下の形で考えれば十分である（有限パラメータの問題に還元）

$$f = \sum_{i=1}^{N} c_i \Phi(X^{(i)}),$$

(Representer定理),

→ すべての量がGram行列によって表現される(サイズ = データ数).

- 元の空間が高次元でも計算量の問題が生じない．Gram行列を計算した後は，データ数のみに依存した計算量.

以上はカーネル法一般に共通の要素である.

参考文献

- 福水「カーネル法入門」1章　朝倉書店　2010.
- B. Schölkopf and A. Smola. *Learning with kernels*. MIT Press, 2002.
- 赤穂 「カーネル多変量解析 —非線形データ解析の新しい展開」 岩波書店 (2008)