1. カーネル法への招待

正定値カーネルによるデータ解析
ー カーネル法の基礎と展開 ー

福水健次
統計数理研究所／総合研究大学院大学

統計数理研究所　公開講座
2011年1月13,14日

概要

- カーネル法の基本
  - 線形データ解析と非線形データ解析
  - カーネル法の原理

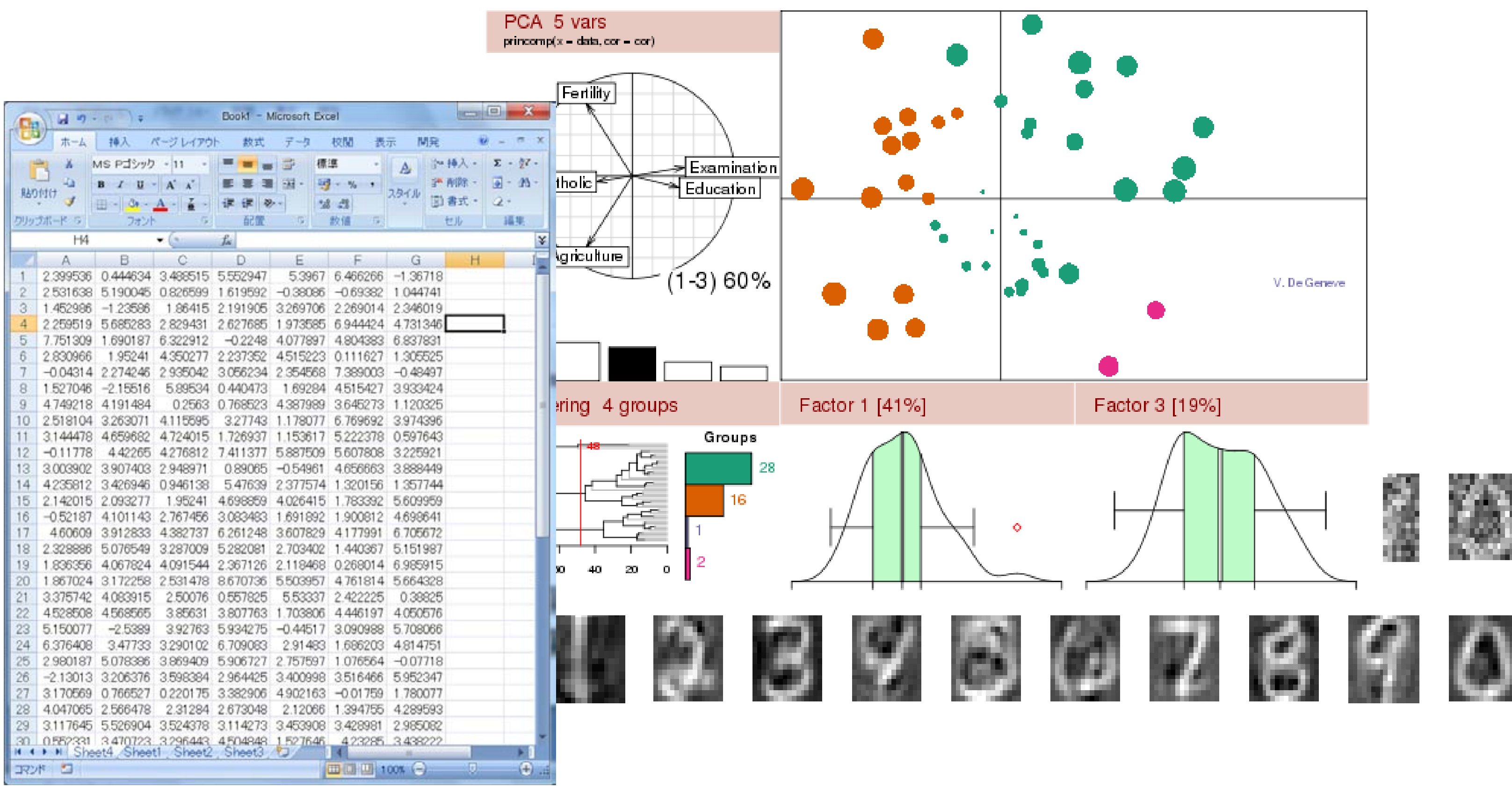
- カーネル法の2つの例
  - カーネル主成分分析: PCAの非線形拡張
  - リッジ回帰とそのカーネル化

概要

- カーネル法の基本
  - 線形データ解析と非線形データ解析
  - カーネル法の原理
- カーネル法の2つの例
  - カーネル主成分分析: PCAの非線形拡張
  - リッジ回帰とそのカーネル化

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*



線形データ解析

– データは数値の「テーブル」として与えられることが多い.
　→ 行列表現

$$\mathbf{X} = \begin{pmatrix} X_1^{(1)} & \cdots & X_m^{(1)} \\ X_1^{(2)} & \cdots & X_m^{(2)} \\ \vdots & & \vdots \\ X_1^{(N)} & \cdots & X_m^{(N)} \end{pmatrix}$$

m dimensional, N data

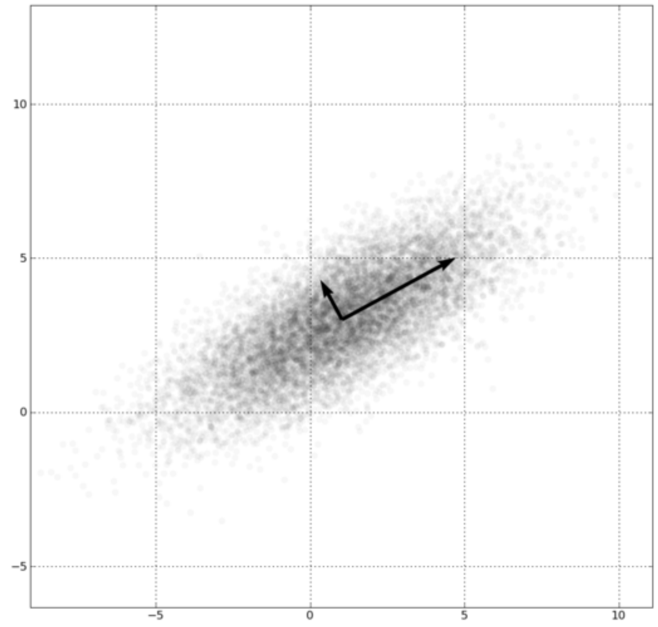
– データ解析には線形代数が主な数学的道具

- 相関 Correlation,
- 線形回帰 Linear regression analysis,
- 主成分分析 Principal component analysis,
- 正準相関分析 Canonical correlation analysis, etc.

- Example 1: Principal component analysis (PCA)

$X^{(1)}, \ldots, X^{(N)}$: m-次元のデータ

PCA: 分散が最大になるようにd-次元の部分空間へ射影する



– 第1主軸 = $\mathrm{argmax}_{\|a\|=1} \mathrm{Var}[a^T X]$

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

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

(標本)共分散行列

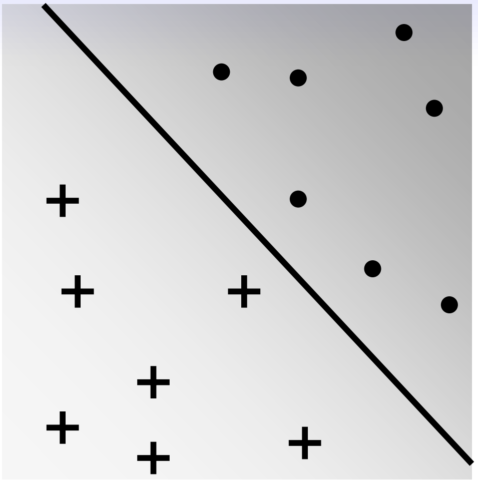
– General solution:

$u_1, \ldots, u_N$: V_{XX} の固有ベクトル(固有値の降順)

- 第*p*主軸 = u_p
- $X^{(i)}$の第*p* 主成分 = $u_p^{\mathrm{T}} X^{(\mathrm{i})}$

–「PCA → 固有値分解 (線形代数)」

- Example 2: Linear classification
  - 2値識別



Input data

$$\mathbf{X} = \begin{pmatrix} X_1^{(1)} & \cdots & X_m^{(1)} \\ X_1^{(2)} & \cdots & X_m^{(2)} \\ \vdots & & \vdots \\ X_1^{(N)} & \cdots & X_m^{(N)} \end{pmatrix}$$

Class label

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

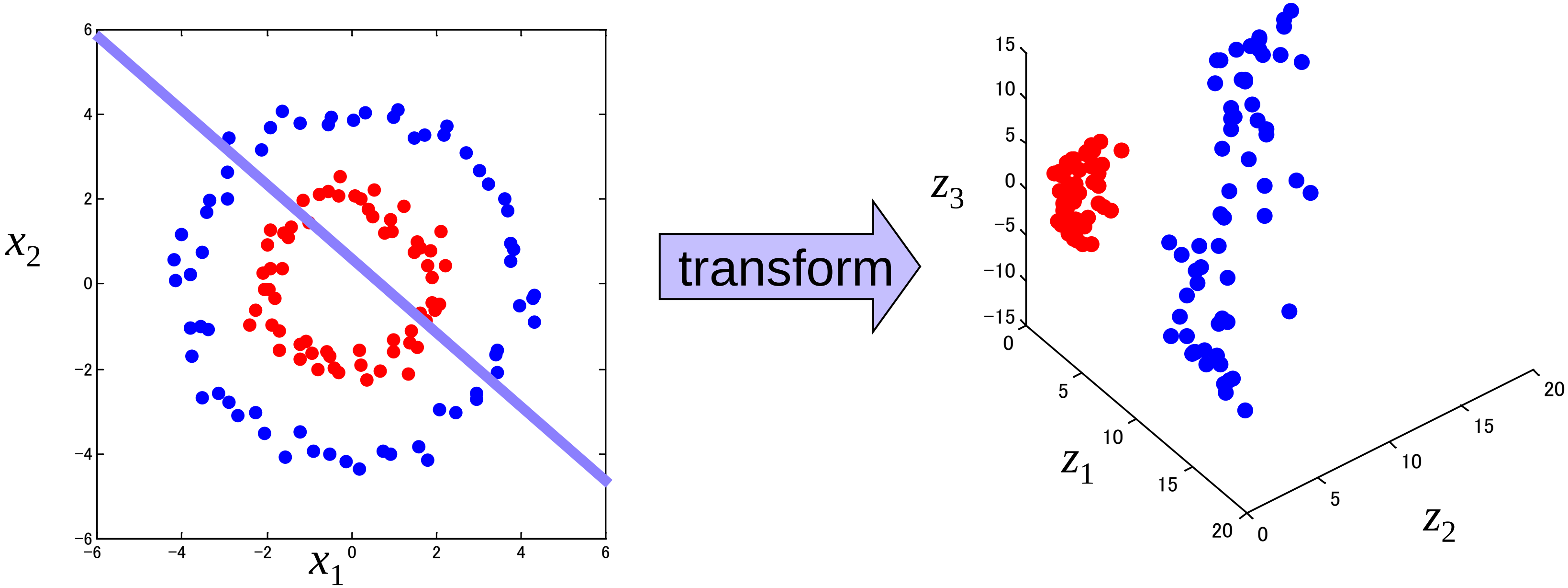
線形識別関数による方法: $h(x) = \mathrm{sgn}\left(a^T x + b\right)$

so that $h(X^{(i)}) = Y^{(i)}$ for all (or most) i.

  - Example: Fisher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?

linearly inseparable　　　　　　linearly separable



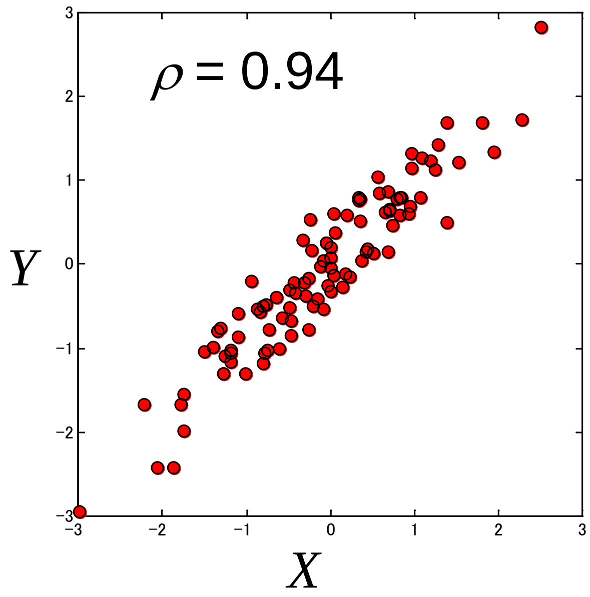
$$(z_1, z_2, z_3) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

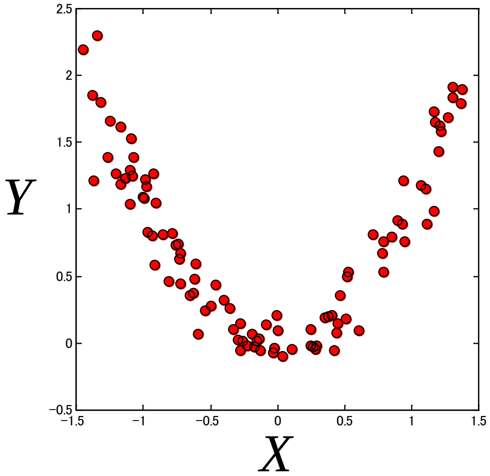
Unclear? Watch the following movie!

http://jp.youtube.com/watch?v=3liCbRZPrZA

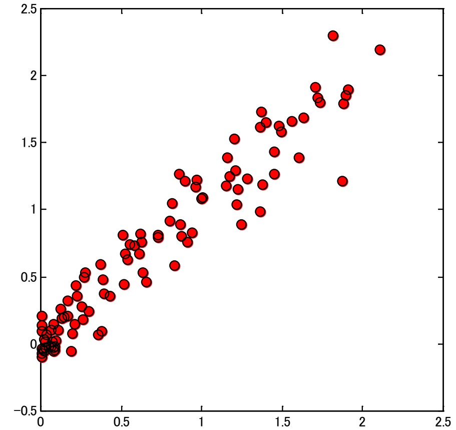
- Another example: correlation

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





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



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

(X, Y) transform (X^2, Y)

データの非線形変換によって高次モーメントを抽出するアプローチが有効そうである.

データの非線形変換

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.

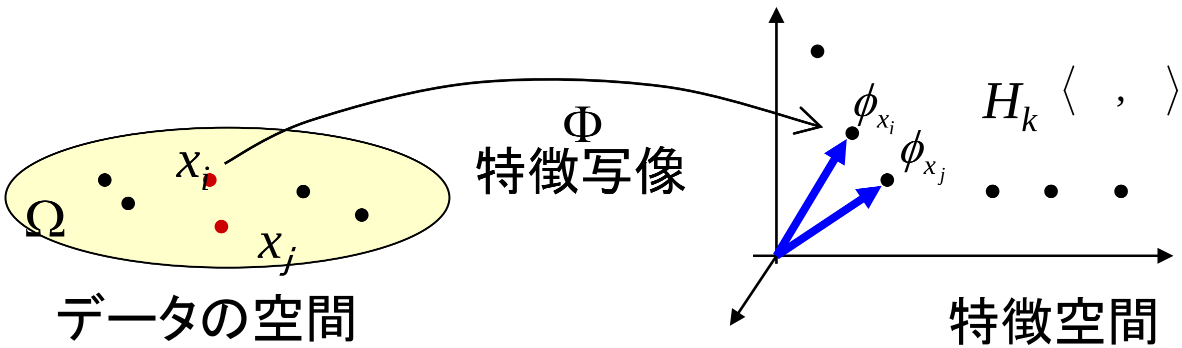
カーネル法 = データの非線形性あるいは高次の情報を扱うための非線形変換の系統的方法論.

概要

- カーネル法の基本
  - 線形データ解析と非線形データ解析
  - カーネル法の原理

- カーネル法の2つの例
  - カーネル主成分分析: PCAの非線形拡張
  - リッジ回帰とそのカーネル化

カーネル法の概観



特徴空間で線形のデータ解析を行う!

e.g. SVM

– どのような変換(特徴写像)がよいか?
  - 元のデータのさまざまな非線形性が抽出できる.
  - 特徴空間で, 内積が計算しやすい.
    多くの線形データ解析手法は, 内積計算に拠っている.

- 計算論的な問題
  - もちろん高次項を並べてもよいが，，，

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

  - 元のデータが高次元だと計算量爆発!

    e.g. 元のデータが100 次元のとき，3次まで取ると，特徴ベクトルの次元は

    $${}_{100}C_1 + {}_{100}C_2 + {}_{100}C_3 = 166750.$$

  - 現在の計算機を持ってしても，行列演算は困難.
    より効率的な方法 → カーネル法.

正定値カーネルによる内積計算

– 特徴写像:

$$\Phi: \quad \Omega \quad \to \quad H$$

– 特徴写像をうまく選択すると，特徴空間での内積が正定値カーネル $k(x, y)$ により与えられる

$$\langle \Phi(X_i), \Phi(X_j) \rangle = k(X_i, X_j) \qquad \text{kernel trick.}$$

– 多くの線形データ解析手法では，内積の値のみが必要で，特徴ベクトル $\Phi(X)$ の形は知らなくてもよい.
(*e.g*. PCA. 後述)

概要

- カーネル法の基本
  - 線形データ解析と非線形データ解析
  - カーネル法の原理

- カーネル法の2つの例
  - カーネル主成分分析: PCAの非線形拡張
  - リッジ回帰とそのカーネル化

PCA から Kernel PCAへ

- PCA: 線形の次元削減法.
- Kernel PCA: 非線形な次元削減法 (Schölkopf et al. 1998).

- Review of PCA

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

- PCAの計算
  - 方向ベクトルa とデータの内積
  - 目的関数の最適化(固有値問題に還元される)

– 特徴空間でのPCA

- 特徴ベクトル: $X^{(1)},\ldots,X^{(N)} \quad \rightarrow \quad \Phi(X^{(1)}),\ldots,\Phi(X^{(N)})$

- 仮定:
  – 特徴空間 H は内積 $\langle\ ,\ \rangle$ を持ち，特徴ベクトルの内積が

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

により計算可能

- 目的関数:

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

f: 特徴空間 H 内での方向ベクトル

ただし $$\tilde{\Phi}(X^{(i)}) = \Phi(X^{(i)}) - \frac{1}{N}\sum_{j=1}^{N}\Phi(X^{(j)})$$

– 解は次の形で十分

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

∵） 空間 H を特徴ベクトル $\tilde{\Phi}(X^{(1)}),\ldots,\tilde{\Phi}(X^{(N)})$ の張る部分空間 H_0 と，その直交補空間 $H_0^{\perp}$ に直交分解：

$$H = H_0 \oplus H_0^{\perp}.$$

方向ベクトル f を

$$f = g + h, \qquad (g \in H_0, h \in H_0^{\perp})$$

と表わすと，目的関数は

$$\max_{\|f\|^2=1} \frac{1}{N}\sum_{i=1}^{N}\left\{\left\langle f, \tilde{\Phi}(X^{(i)})\right\rangle\right\}^2 = \max_{\|g\|^2+\|h\|^2=1} \frac{1}{N}\sum_{i=1}^{N}\left\{\left\langle g, \tilde{\Phi}(X^{(i)})\right\rangle\right\}^2$$

$h=0$ の時が最適.

– 内積計算： $f = \sum_{i=1}^{N} c_i \tilde{\Phi}(X^{(i)})$

- ノルム^2　$\| f \|^2 = c^T \tilde{K} c$
- 分散　$\sum_i \left\langle f, \tilde{\Phi}(X^{(i)}) \right\rangle^2 = \sum_i \left\langle \sum_j c_j \tilde{\Phi}(X^{(j)}), \tilde{\Phi}(X^{(i)}) \right\rangle^2 = c^T \tilde{K}^2 c$

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

中心化グラム行列
(centered Gram matrix)

– 目的関数

Kernel PCA：

$$\max c^T \tilde{K}^2 c \qquad \text{subject to*} \quad c^T \tilde{K} c = 1$$

*) “subject to” は，最適化問題で制約条件を書くときの慣用句.

- Kernel PCA も固有値問題に還元される
- Kernel PCA アルゴリズム
  - 行列 $\tilde{K}$ の計算
  - $\tilde{K}$ の固有値分解

$$\tilde{K} = \sum_{i=1}^{N} \lambda_i u_i u_i^T$$

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

$$u_1, u_2, \ldots, u_N$$ unit eigenvectors

  - 第p主軸

$$f^p = \sum_{i=1}^{N} c_p^i \tilde{\Phi}(X^{(i)}), \quad c_p = \frac{1}{\sqrt{\lambda_p}} u_p$$

  - $X^{(i)}$の第p主成分

$$= \left\langle f^p, \tilde{\Phi}(X^{(i)}) \right\rangle = \sqrt{\lambda_p} u_p^T X^{(i)}$$

– 内積計算のチェック:　for　$f = \sum_{i=1}^{N} c_i \tilde{\Phi}(X^{(i)})$

- $\| f \|^2 = c^T \tilde{K} c$

$$\|f\|^2 = \left\langle \sum_i c_i \tilde{\Phi}(X^{(i)}), \sum_j c_j \tilde{\Phi}(X^{(j)}) \right\rangle = \sum_{i.j} c_i c_j \left\langle \tilde{\Phi}(X^{(i)}), \tilde{\Phi}(X^{(j)}) \right\rangle$$

$\tilde{K}_{ij} := \left\langle \tilde{\Phi}(X^{(i)}), \tilde{\Phi}(X^{(j)}) \right\rangle$　を展開せよ.

$$\begin{aligned}
\tilde{K}_{ij} &= \left\langle \Phi(X_i) - \tfrac{1}{N}\sum_a \Phi(X_a), \Phi(X_j) - \tfrac{1}{N}\sum_b \Phi(X_b) \right\rangle \\
&= \left\langle \Phi(X_i), \Phi(X_j) \right\rangle - \tfrac{1}{N}\sum_a \left\langle \Phi(X_a), \Phi(X_j) \right\rangle \\
&\qquad - \tfrac{1}{N}\sum_b \left\langle \Phi(X_i), \Phi(X_b) \right\rangle + \tfrac{1}{N^2}\sum_{a,b} \left\langle \Phi(X_a), \Phi(X_b) \right\rangle \\
&= k(X_i, X_j) - \tfrac{1}{N}\sum_a k(X_a, X_j) - \tfrac{1}{N}\sum_b k(X_i, X_b) + \tfrac{1}{N^2}\sum_{a,b} k(X_a, X_b)
\end{aligned}$$

- $\sum_i \left\langle f, \tilde{\Phi}(X^{(i)}) \right\rangle^2 = c^T \tilde{K}^2 c$

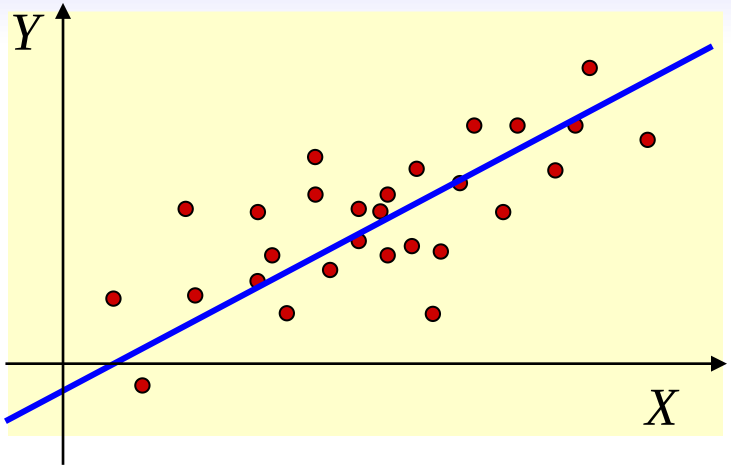
上と同様に計算できる([Exercise])

リッジ回帰

- **線形回帰(復習)**

$$(X^{(1)},Y^{(1)}),\ldots,(X^{(N)},Y^{(N)})$$

$$X^{(i)} \in \mathbf{R}^m,\ \ Y^{(i)} \in \mathbf{R}$$



問題 $\min \sum_{i=1}^{N} (Y^{(i)} - f_w(X^{(i)}))^2$

を達成する線形関数 $f_w(x) = w^T x$

データ行列 $X = \begin{pmatrix} X_1^1 & \cdots & X_m^1 \\ X_1^2 & \cdots & X_m^2 \\ \vdots & & \vdots \\ X_1^N & \cdots & X_m^N \end{pmatrix}, \quad Y = \begin{pmatrix} Y^1 \\ Y^2 \\ \vdots \\ Y^N \end{pmatrix}$

最適解 $\hat{w} = (X^T X)^{-1} X^T Y$

最適な関数 $f_{\hat{w}}(x) = Y^T X (X^T X)^{-1} x$

- リッジ回帰（Ridge regression）
  - 問題 $\min \sum_{i=1}^{N} (Y^{(i)} - f_w(X^{(i)}))^2 + \lambda \| w \|^2$

    を達成する線形関数 $f_w(x) = w^T x$

    最適解は $\hat{w} = (X^T X + \lambda I_N)^{-1} X^T Y$

    最適な関数は $f_{\hat{w}}(x) = Y^T X (X^T X + \lambda I_N)^{-1} x$

  - リッジ回帰は、$X^T X$ が特異また特異に近いときによく使われる.
  - Bayes的な解釈もできる（正則化：後述)

リッジ回帰のカーネル化

– Data: $(X^{(1)}, Y^{(1)}), \ldots, (X^{(N)}, Y^{(N)})$

$X^{(i)}$: 任意の集合 Ω に値を持つ.　　$Y^{(i)} \in \mathbf{R}$

– 特徴写像

$$X^{(1)}, \ldots, X^{(N)} \quad \to \quad \Phi(X^{(1)}), \ldots, \Phi(X^{(N)})$$

- 仮定 :
  - 特徴空間 H は内積 $\langle\ ,\ \rangle$ を持ち，特徴ベクトルの内積が

$$\left\langle \Phi(X^{(i)}), \Phi(X^{(j)}) \right\rangle = k(X^{(i)}, X^{(j)}) \qquad \text{(kernel trick)}$$

により計算可能

– 特徴空間 H におけるリッジ回帰

$$\min_{f \in H} \sum_{i=1}^{N} \left(Y^{(i)} - \left\langle f, \Phi(X^{(i)}) \right\rangle \right)^2 + \lambda \| f \|_H^2$$

– 最適解：は以下の形を持つ

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

∵）データ $\{\Phi(X_i)\}$ の張る H の部分空間をH_0，直交補空間を$H_\perp$とするとき，$f=h_0+h_1$ の分解で，目的関数の第１項は h_0 のみに依存．第２項は $\|f\|^2=\|h_0\|^2+\|h_1\|^2$ により $h_1=0$のときが最適．

– 内積計算

- ２乗ノルム：　$\|f\|^2=c^T Kc$　　　$K_{ij}=k(X^i, X^j)$ グラム行列
- 線形関数：　$\left\langle f, \Phi(X^{(i)})\right\rangle = \left\langle \sum_j c_j \Phi(X^{(j)}), \Phi(X^{(i)})\right\rangle = (Kc)_i$

– カーネルリッジ回帰

$$\min_{c\in\mathbf{R}^N} \quad \|Y-Kc\|^2+\lambda c^T Kc$$

最適解

$$f(x)=Y^T(K+\lambda I_N)^{-1}\mathbf{k}(x), \qquad \mathbf{k}(x)=\begin{pmatrix} k(x,X^{(1)}) \\ \vdots \\ k(x,X^{(N)}) \end{pmatrix}$$

[Exercise] 導出を確認せよ.

カーネル法の原理

– 特徴写像によって，内積 $\langle\ ,\ \rangle$ を持つ特徴空間Hにデータを写像する．

$$X^{(1)},\ldots,X^{(N)} \quad \rightarrow \quad \Phi(X^{(1)}),\ldots,\Phi(X^{(N)})$$

– 特徴ベクトル $\{\Phi(X^{(i)})\}_{i=1}^{N}$ に，H 上で線形の解析手法を適用する．

– 多くの場合，最適解（H の元）は次の形を持つことがわかる．

$$f = \sum_i c_i \Phi(X^{(i)})$$

– 問題は，内積 $\left\langle \Phi(X^{(i)}), \Phi(X^{(j)}) \right\rangle$ を用いて表現される．

– カーネル法では，上の内積が

$$\left\langle \Phi(X^{(i)}), \Phi(X^{(j)}) \right\rangle = k(X^{(i)}, X^{(j)}) \qquad \text{(kernel trick)}$$

によって効率的に計算される．基底による展開や表示は必要ない．

Question:

カーネルトリック

$$\langle \Phi(x), \Phi(y) \rangle = k(x, y)$$

を満たすような，特徴写像 Φ と k はどのようなものか？

➡ 正定値カーネル