カーネル法

正定値カーネルを用いたデータ解析

統計数理研究所　福水健次

2004年11月24~26日　公開講座「機械学習の最近の話題」

Final version. Nov.26, 2004

講義の概要 I

1. イントロ ー 非線形データ解析としてのカーネル法

2. 正定値カーネルの基礎
   - 正定値カーネルの定義と代表的な例
   - 正定値カーネルと関数空間

3. 線形アルゴリズムの非線形化としてのカーネル法
   - 正定値カーネルによる非線形化
   - サポートベクターマシン, スプライン平滑化, カーネルPCA, カーネルCCA

4. 正定値カーネルの性質
   - 正定値性の判定
   - Bochnerの定理
   - representer定理

講義の概要 II

5. 構造化データのカーネル
   - 複雑な構造を持つ非ベクトルデータ（ストリング，ツリー，グラフ）の数量化としてのカーネル法

6. 独立性・条件付独立性とカーネル
   - 確率変数の独立性・条件付独立性の特徴づけ
   - カーネルICA，カーネル次元削減法

7. まとめ

用語に関する注意

- 「カーネル」という用語は，統計学では伝統的にノンパラメトリックな確率密度推定

$$p(x) = \frac{1}{N}\sum_{i=1}^{N} g(x - x_i)$$

に用いる密度関数 $g(x)$ の意味に使われることも多い（Parzen windowともいう）

- 最近では、正定値カーネルのことを「カーネル」「カーネル法」と呼ぶことが多いので、注意して区別する必要がある

- 本講義の「カーネル」は後者の意味である

1. イントロダクション

- このセクションの目的
  - カーネル法に関して大まかなイメージを持ってもらう
  - くわしい説明はあとできちんとやる

非線形データ解析としてのカーネル法

- 非線形データ解析の重要性
  - 古典的なデータ解析

    データの行列表現

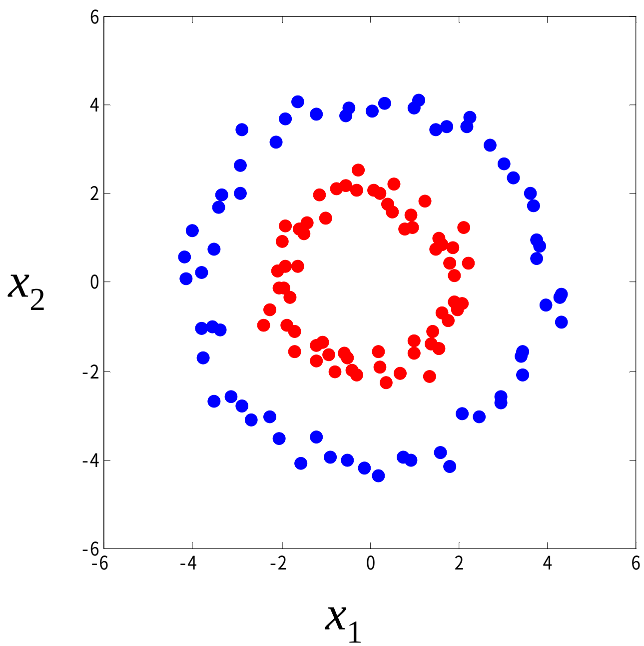
    m 次元　N 点のデータ

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

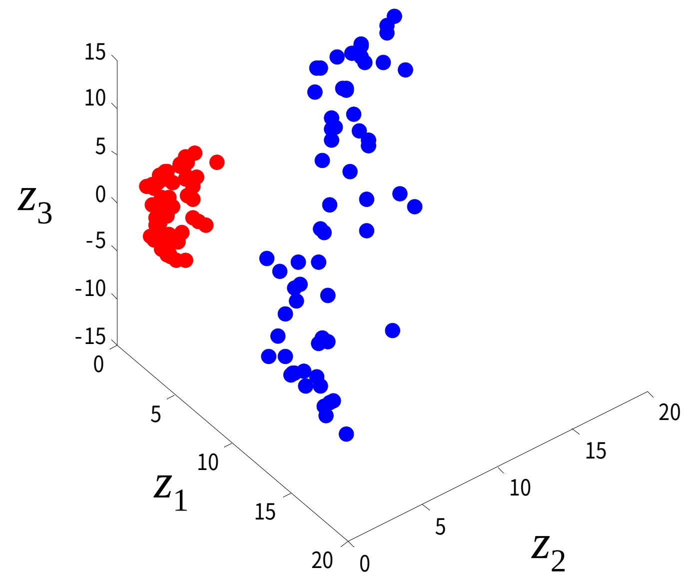
⇒ 線形の処理（主成分分析, 正準相関分析, 線形回帰. . . ）

  - 線形で十分か？

線形識別**不能**



線形識別**可能**

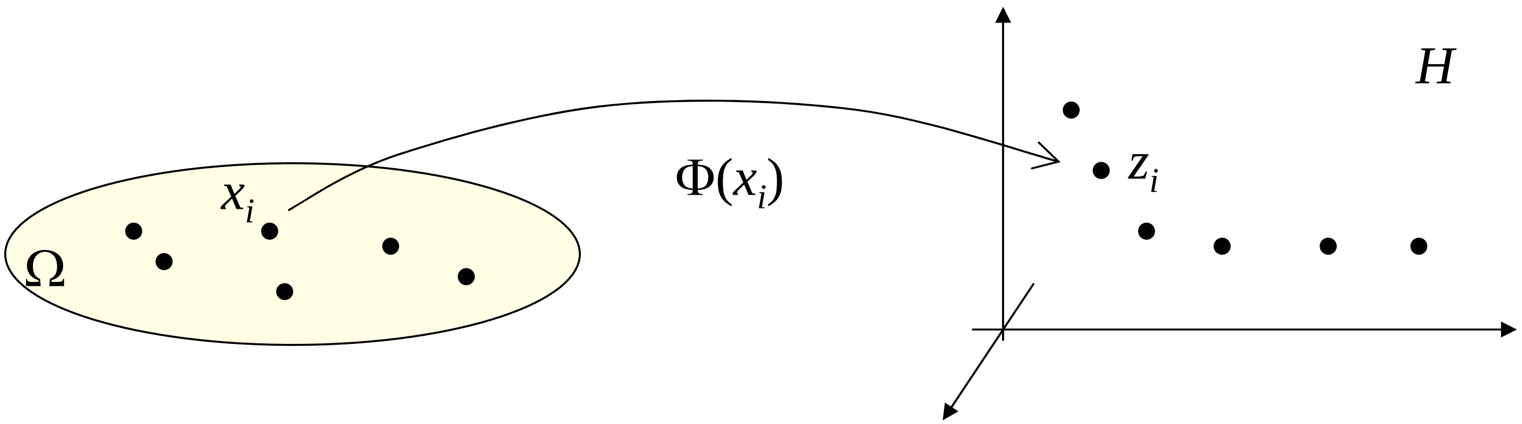


$$(z_1, z_2, z_3) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$

カーネル法の概略

- **高次元空間への非線形写像**

  データを高次元のベクトル空間（一般には無限次元の関数空間）へ写像し、解析しやすいデータに変換する。



Ω ：もとのデータの空間

H ：高次元ベクトル空間（ヒルベルト空間）

$\Phi : \Omega \to H$ 　変換写像

- H = **特徴ベクトルの空間 (特徴空間**, feature space)
  - $\Phi(x)$ **はデータ** x **に対する特徴ベクトルと考えることができる**
  - **もとの空間** Ω **でなく、(高次元、または無限次元) 特徴空間** H **でデータ解析を行う**
  - **もとの空間** Ω **は、ベクトル空間でなくてもよい**

    $\Rightarrow$ **データ** x **はツリー、グラフなどでもよい。**

- 「カーネルトリック」

  高次元 (無限次元) ベクトル空間 H において、内積が容易に計算できる。

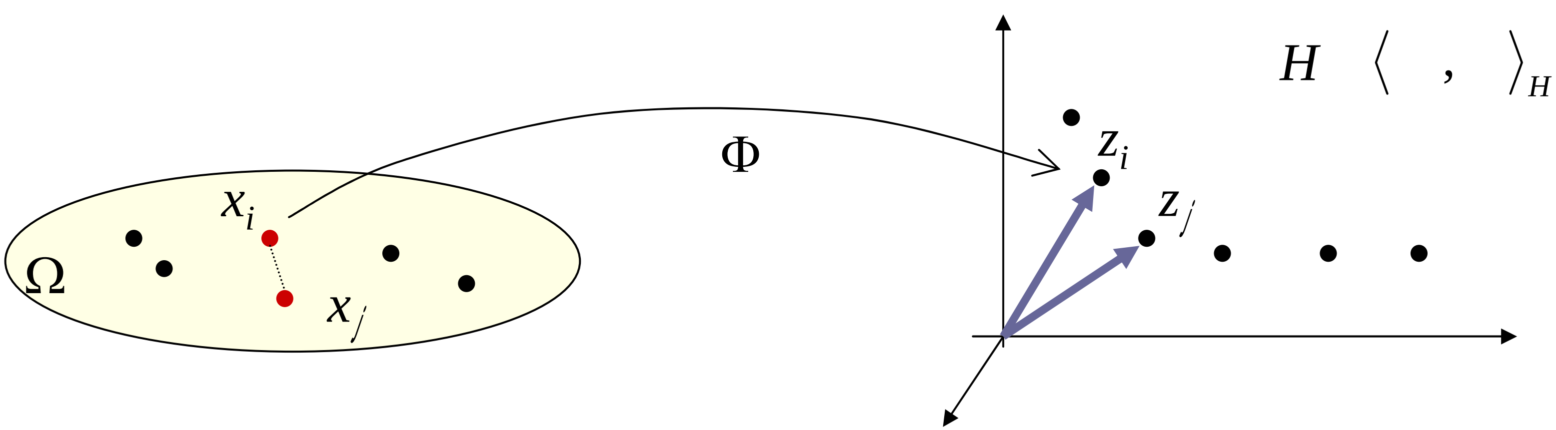
  ⇒ さまざまなデータ解析手法を H 上で適用可能

  Support vector machine, Kernel PCA, Kernel CCA

$$\langle \Phi(x_i), \Phi(x_j) \rangle_H = k(x_i, x_j)$$

・・・ 正定値カーネル

データ x_i と x_j の類似度を定める



2. 正定値カーネルの基礎

- このセクションの目的
  - カーネル法で用いられる基礎的な概念を述べる
  - 正定値カーネルの代表的な例を紹介する
  - 正定値カーネルがヒルベルト空間を定めることを説明する

正定値カーネル

- **正定値カーネル**

  Ω：集合．　$k:\Omega\times\Omega\to\mathbf{R}$

  $k(x,y)$ がΩ上の**正定値カーネル**であるとは，次の2つを満たすことをいう

  1．(対称性)　$k(x,y) = k(y,x)$

  2．(正定値性)　任意の自然数 n と，任意のΩ の点 $x_1, \ldots, x_n$ に対し，

  $n\times n$ 行列
$$\left(k(x_i,x_j)\right)_{i,j=1}^{n} = \begin{pmatrix} k(x_1,x_1) & \cdots & k(x_1,x_n) \\ \vdots & \ddots & \vdots \\ k(x_n,x_1) & \cdots & k(x_n,x_n) \end{pmatrix}$$

  が (半) 正定値．すなわち，任意の実数 $c_1,\ldots, c_n$ に対し，
$$\sum_{i,j=1}^{n} c_i c_j k(x_i,x_j) \geq 0$$

  - 対称行列　$\left(k(x_i,x_j)\right)_{i,j=1}^{n}$　のことを，**グラム行列**と呼ぶ

複素数値の正定値カーネル

$$k : \Omega \times \Omega \to \mathbf{C}$$

の場合にも正定値性は以下のように定義される．

任意の自然数 n と，任意のΩ の点 $x_1, \ldots, x_n$ と，任意の複素数 $c_1,\ldots, c_n$ に対し，

$$\sum_{i,j=1}^{n} c_i \overline{c}_j k(x_i, x_j) \geq 0 \qquad (\overline{c}_j \text{ は複素共役})$$

が成り立つとき，$k(x,y)$ を正定値カーネルという．

カーネル／正定値カーネル

正定値カーネルのことを単に「カーネル」と呼ぶ場合も多い

正定値カーネルの例

- **多項式カーネル**

$\Omega = \mathbf{R}^m$

$$k(x,y) = (x^T y + c)^d \qquad (d\text{:}\textbf{自然数}\text{, } c \geq 0)$$

- **ガウスカーネル** (RBF**カーネル**)

$\Omega = \mathbf{R}^m$

$$k(x,y) = \exp\left(-\frac{1}{\sigma^2} \|y - x\|^2 \right) \qquad (\sigma > 0)$$

- Fourier**カーネル(複素数値)**

$\Omega = \mathbf{R}^m$

$$k(x,y) = e^{\sqrt{-1}\omega^T (x-y)} \qquad (\omega \in \mathbf{R}^m)$$

  - **正定値性であることはあとでチェックする**

ヒルベルト空間の復習

- (実)ヒルベルト空間

  ベクトル空間で, 内積 $\langle\ ,\ \rangle$ が与えられている.

  完備性を満たす. (→ 本講義では使わないので説明は省略)

復習

ベクトル空間 V の内積 $\langle\ ,\ \rangle$ とは, 次の3条件を満たす

V×V→R の写像

(1) 線形性 $\langle \alpha f + \beta g, h \rangle = \alpha \langle f, h \rangle + \beta \langle g, h \rangle$ $(\alpha, \beta \in \mathbf{R},\ f, g \in \mathrm{V})$

(2) 対称性 $\langle g, f \rangle = \langle f, g \rangle$

(3) 強正定値性 $\langle f, f \rangle \geq 0$ かつ $\langle f, f \rangle = 0 \Leftrightarrow f = 0$

- 複素ヒルベルト空間も定義されるが, 以下では実ヒルベルト空間を考える.

- 内積があると，ノルムが $\|f\| = \sqrt{\langle f, f \rangle}$ により定まる．

- ヒルベルト空間は，無限次元でもよい

- 例） $L_2(a,b)$：区間 (a,b) 上の2乗可積分関数全体

  内積 $\langle f, g \rangle = \int_a^b f(x) g(x) dx$

- ユークリッド空間 $\mathbf{R}^m$ もヒルベルト空間のひとつ．

正定値カーネルとヒルベルト空間

- **定理**

$k(x,y)$: 集合 Ω 上の正定値カーネル

⇨

Ω 上の関数からなるヒルベルト空間 H_k が一意に存在して，
次の3つを満たす

(1) $k(\cdot\,, x) \in H_k$ ($x \in \Omega$ は任意に固定)

(2) 有限和 $f = \sum_{i=1}^{n} c_i k(\cdot\,, x_i)$ の形の元は H_k の中で稠密

(3) (再生性) $f(x) = \langle f, k(\cdot\,, x) \rangle$ $\forall f \in H_k,\ x \in \Omega$

注) $k(\cdot, x)$ ・・・ x を固定した1変数関数

- **再生核ヒルベルト空間** (Reproducing Kernel Hilbert Space)
  - **集合 Ω 上の関数を要素に持つヒルベルト空間 H が**
    再生核ヒルベルト空間であるとは,
    任意の $x \in \Omega$ に対して $\phi_x \in H$ があって, 任意の $f \in H$ に対し
    $$\langle f, \phi_x \rangle = f(x)$$
    が成り立つことをいう.
  - ϕ_x **のことを再生核という.**

以下 RKHS と略する場合がある

- **正定値カーネルとRKHS**
  - **正定値カーネル ⇒** RKHS

    正定値カーネル $k(x,y)$ **により定まる** H_k **は再生核を持つ(定理の**(3))

$$\phi_x = k(\cdot\,,x) \qquad \Rightarrow \qquad \langle f, \phi_x \rangle = f(x)$$

  - RKHS **⇒ 正定値カーネル：　再生核** ϕ_x **は正定値カーネルを定める**

$$k(y,x) = \phi_x(y)$$ **と定義**

$$\Rightarrow$$

$$k(y,x) = \phi_x(y) = \langle \phi_x, \phi_y \rangle$$ **(対称性もわかる)**

$$\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) = \sum_{i,j=1}^{n} c_i c_j \langle \phi_{x_i}, \phi_{x_j} \rangle = \left\langle \sum_{i=1}^{n} c_i \phi_{x_i}, \sum_{j=1}^{n} c_j \phi_{x_j}, \right\rangle$$

$$= \left\| \sum_{i=1}^{n} c_i \phi_{x_i} \right\|^2 \geq 0$$ **(正定値性)**

  - **正定値カーネル** ⟺ **再生核ヒルベルト空間**

再生核ヒルベルト空間の性質

- 関数の値が扱える

  L^2 空間などでは関数の値は定まらない (測度0の集合上の値を変更しても同じ元)

- 再生性
  - 関数の値が内積で計算できる
  - 内積が関数の値で計算できる ・・・ カーネルトリック (次のスライド)

- 連続性

  $\Omega = \mathbf{R}^m$, 正定値カーネル $k(x, y)$ が連続だとすると,
  定義されるRKHS H_k の関数はすべて連続関数

$$\because) \quad |f(x) - f(y)|^2 = |\langle f, k(\cdot, x) - k(\cdot, y)\rangle|^2 \le \|f\|^2 \|k(\cdot, x) - k(\cdot, y)\|^2$$
$$= \|f\|^2 (k(x,x) - 2k(x,y) + k(y,y)) \to 0 \quad (x \to y)$$

実は, $k(x,y)$ が微分可能だと, すべての関数が微分可能

再生核とカーネルトリック

Ω：データが含まれる空間

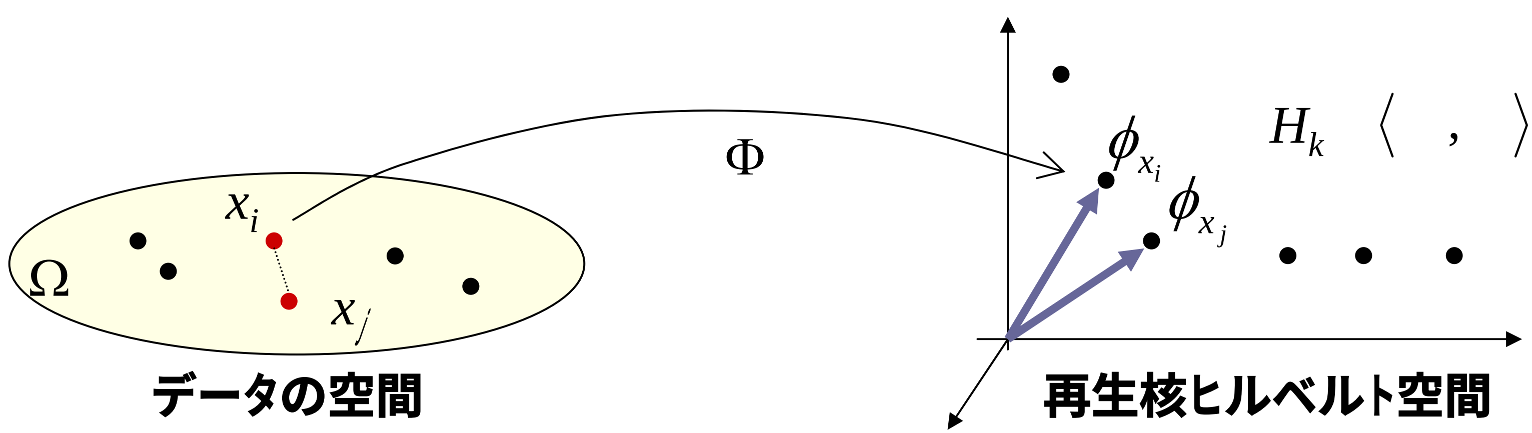
$k(x,y)$：集合 Ω 上の正定値カーネル

H_k ：k により定まる再生核ヒルベルト空間

$$\Phi : \Omega \to H_k, \qquad \Phi(x) = k(\cdot\,, x) = \phi_x$$ により定める

内積計算は関数 k の計算でOK

$$\langle \Phi(x), \Phi(y) \rangle = \langle k(\cdot\,, x), k(\cdot\,, y) \rangle = k(x, y)$$ ・・・ **カーネルトリック**



- **例：多項式カーネル**

$\mathbf{R}^2$ 上の正定値カーネル $k(x,y) = (x^T y)^2 = (x_1 y_1 + x_2 y_2)^2$

k **が定める再生核ヒルベルト空間** H_k **と、写像** $\Phi: \mathbf{R}^2 \to H_k$ **を求めよう。**

$H : \mathbf{R}^2$ **上の2次関数全体**

$$f(z) = \alpha_{11} z_1^2 + \alpha_{12}(\sqrt{2} z_1 z_2) + \alpha_{22} z_2^2$$

$z_1^2, \sqrt{2} z_1 z_2, z_2^2$ **を正規直交基底として、内積を以下で定義**

$$f(z) = \alpha_{11} z_1^2 + \alpha_{12}(\sqrt{2} z_1 z_2) + \alpha_{22} z_2^2, \quad g(z) = \beta_{11} z_1^2 + \beta_{12}(\sqrt{2} z_1 z_2) + \beta_{22} z_2^2$$

$$\langle f, g \rangle_H = \alpha_{11}\beta_{11} + \alpha_{12}\beta_{12} + \alpha_{22}\beta_{22}$$

H **は** $\mathbf{R}^3$ **と同型になる。**

- $H \cong H_k$

1． $k(\cdot, x) \in H$ $\because)$ $k(z,x) = \boxed{x_1^2} \cdot z_1^2 + \boxed{\sqrt{2}x_1x_2} \cdot \sqrt{2}z_1z_2 + \boxed{x_2^2} \cdot z_2^2$

係数

2．再生性： 任意の H の元 $f(z) = \alpha_{11}z_1^2 + \alpha_{12}(\sqrt{2}z_1z_2) + \alpha_{22}z_2^2$ に対し

$$\langle f, k(\cdot, x)\rangle_H = \alpha_{11}x_1^2 + \alpha_{12}\sqrt{2}x_1x_2 + \alpha_{22}x_2^2 = f(x)$$

- カーネルトリック

$$\Phi(x) = k(\cdot, x) \leftrightarrow \begin{pmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{pmatrix}$$

$z_1^2, \sqrt{2}z_1z_2, z_2^2$ と基底とした表現

$$\langle \Phi(x), \Phi(y)\rangle_H = k(x,y) = (x_1y_1 + x_2y_2)^2$$

H (3次元) の内積　　　Ω (2次元) の計算で済んでいる

多項式の次数が高ければ圧倒的に $k(x,y)$ の計算が有利

セクション2のまとめ

- 正定値カーネルの定義
  - グラム行列の(半)正定値性

- 正定値カーネルの代表的な例
  - 多項式カーネル　$k(x,y)=(x^T y + c)^d$
  - ガウスカーネル　$k(x,y)=\exp\left(-\frac{1}{\sigma^2}\|y-x\|^2\right)$
  - Fourierカーネル (複素数値カーネル)　$k(x,y)=e^{\sqrt{-1}\omega^T(x-y)}$

- 再生核ヒルベルト空間
  - 正定値カーネルは, 特別な内積を持つ関数空間を定める
  - 再生核ヒルベルト空間は都合のよい性質を持つ
    - 再生性, 関数の値が定まる, (場合によっては) 連続性, 微分可能性

3．線形アルゴリズムの非線形化としてのカーネル法

- このセクションの目的
  - 線形なデータ解析法をカーネルによって非線形化する方法を紹介する
  - 具体的なカーネル化アルゴリズム

    スプライン平滑化
    サポートベクターマシン
    カーネルPCA
    カーネルCCA

特徴空間での線形アルゴリズム

- **線形のアルゴリズム**

  データが $\mathbf{R}^m$ のベクトル

  → 線形アルゴリズムの利用

  線形回帰、主成分分析、正準相関分析 etc

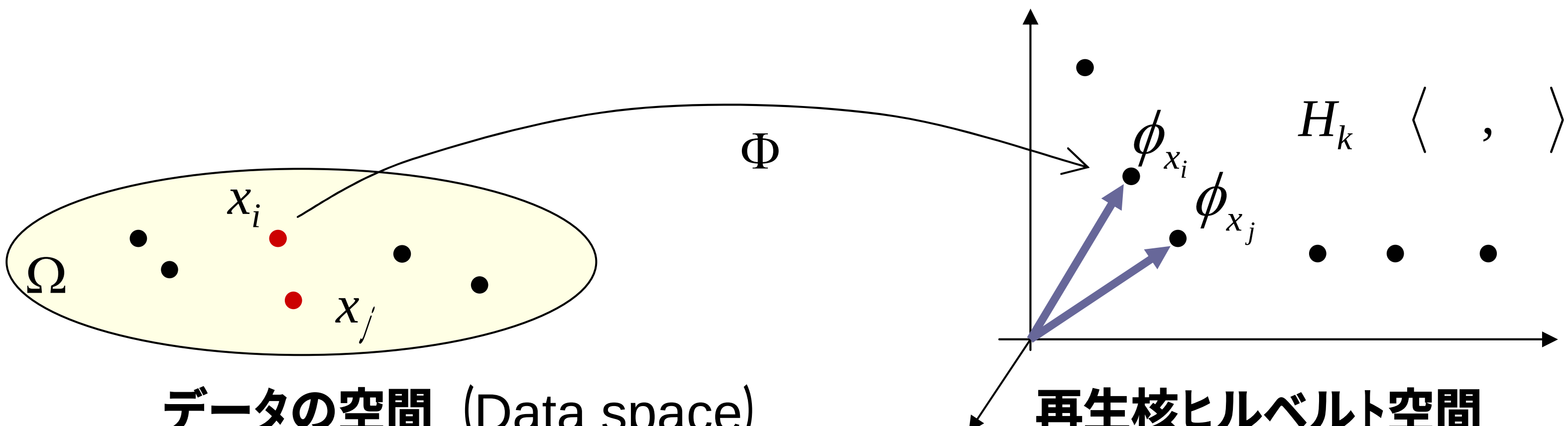
  相関、分散共分散行列の計算が本質的

内積計算ができれば、ヒルベルト空間内のデータにも適用可能

カーネルによる非線形化

正定値カーネルにより定まるヒルベルト空間は**特徴空間**とも呼ばれる

$x \mapsto \Phi(x) = k(\cdot, x)$　　x の**特徴ベクトル**とみなせる



データの空間 (Data space)

再生核ヒルベルト空間
= **特徴空間** (feature space)

特徴空間における線形アルゴリズム
→ データの空間での非線形アルゴリズム

カーネルPCA, カーネルCCA
SVM, スプライン平滑化　etc

PCAとカーネルPCA

- **主成分分析** (PCA, **復習**)

  m **次元データ** $X_1, \ldots, X_N$

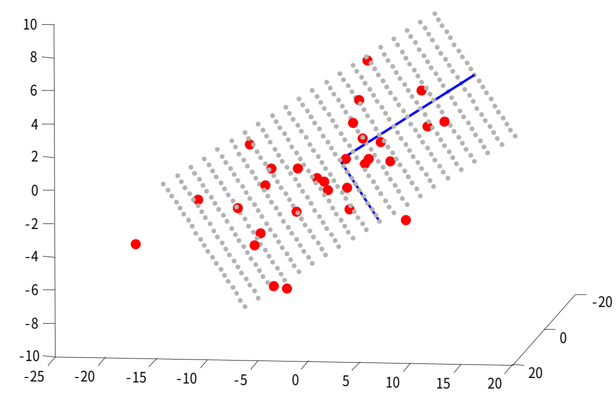
  主成分分析　･･･　分散が最大になる方向 (部分空間) にデータを射影

  単位ベクトル a **方向の分散**：$\mathrm{Var}[a^T X] = \frac{1}{N}\sum_{i=1}^{N}(a^T \tilde{X}_i)^2 = a^T V a$

  $V = \frac{1}{N}\sum_{i=1}^{N}\tilde{X}_i \tilde{X}_i^T$ **分散共分散行列**　　$\tilde{X}_i = X_i - \frac{1}{N}\sum_{j=1}^{N} X_j$ (**中心化**)

  V **の固有ベクトル** $u_1, u_2, \ldots, u_m$ (**ノルム1**)

  $(\lambda_1 \geqq \lambda_2 \geqq \ \ldots \geqq \lambda_m)$



  ⇨

  第 p **主成分の軸　＝**　u_p

  データ X_j **の第** p **主成分　＝**　$u_p^T X_j$

- **カーネルPCA** (Schölkopf et al 98)

データ $X_1, \ldots, X_N$ ⇨ **特徴ベクトル** $\phi_1, \ldots, \phi_N$

カーネル k を設定

$\phi_j = k(\cdot, X_j) \in H_k$

特徴空間での単位ベクトル h 方向の分散 $= \frac{1}{N}\sum_{i=1}^{N}\langle h, \tilde{\phi}_i \rangle^2$

ただし $\tilde{\phi}_i = \phi_i - \frac{1}{N}\sum_{j=1}^{N}\phi_j$ **(中心化)**

$h = \sum_{i=1}^{N}\alpha_i \tilde{\phi}_i$ **としてよい (∵ 直交する方向は分散に寄与しない)**

⇨ **分散** $= \frac{1}{N}\sum_{a=1}^{N}\left\langle \sum_{j=1}^{N}\alpha_j \tilde{\phi}_j, \tilde{\phi}_a \right\rangle^2 = \frac{1}{N}\alpha^T \tilde{K}^2 \alpha$ **ただし** $\tilde{K}_{ij} = \langle \tilde{\phi}_i, \tilde{\phi}_j \rangle$

主成分は

$$\begin{cases} \max_{\alpha} \alpha^T \tilde{K}^2 \alpha \\ \textbf{制約条件} \quad \alpha^T \tilde{K} \alpha = 1 \end{cases}$$

$\Leftarrow \| h \|_{H_k} = \left\langle \sum_i \alpha_i \tilde{\phi}_i, \sum_i \alpha_i \tilde{\phi}_i \right\rangle = \alpha^T \tilde{K} \alpha$

- **カーネルPCAのアルゴリズム**

分散最大の方向 ＝ $\tilde{K}$ の最大固有値の固有ベクトル方向

$$\tilde{K}_{ij} = K(X_i, X_j) - \tfrac{1}{N}\textstyle\sum_{a=1}^{N} K(X_i, X_a) - \tfrac{1}{N}\sum_{a=1}^{N} K(X_a, X_j) + \tfrac{1}{N^2}\sum_{a,b=1}^{N} K(X_a, X_b)$$
$$= \left(Q_N K Q_N\right)_{ij}$$

ただし $Q_N = I_N - \tfrac{1}{N}\mathbf{1}_N\mathbf{1}_N^T, \quad \mathbf{1}_N = (1,\ldots,1)^T$

$\tilde{K}$ の固有値分解

$$\tilde{K} = \textstyle\sum_{a=1}^{N} \lambda_a u^a u^{aT}$$

第 p 主成分を与える α： $\alpha^{(p)} \propto u^p$

$\alpha^T \tilde{K} \alpha = 1$ **ゆえ** $\alpha^{(p)} = \dfrac{1}{\sqrt{\lambda_p}} u^p$

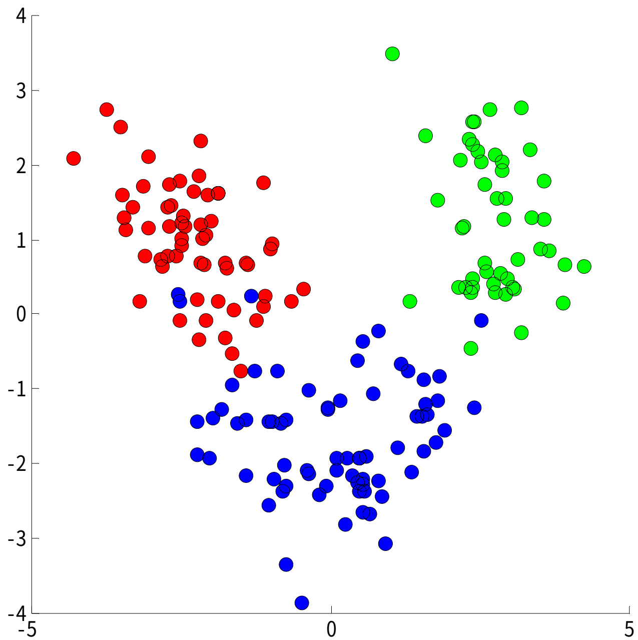
⇨ $$\boxed{\text{データ } X_j \text{ の第 } p \text{ 主成分} = \left\langle \textstyle\sum_{i=1}^{N} \alpha_i^{(p)} \tilde{\phi}_i, \tilde{\phi}_j \right\rangle = \sqrt{\lambda_p}\, u_j^p}$$

- カーネルPCAの実験例

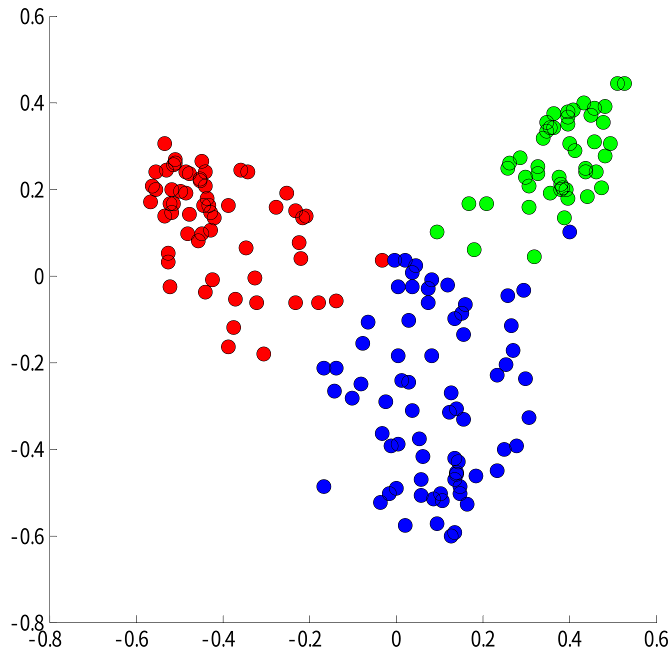
'Wine' データ (UCI Machine Learning Repository)

13次元，178データ，3種類のワインの属性データ

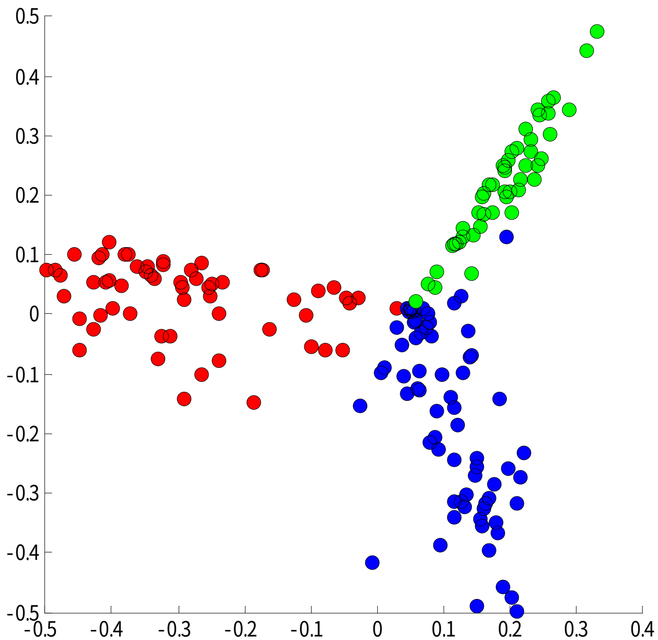
2つの主成分を取った (3クラスの色は参考に付けたもの)



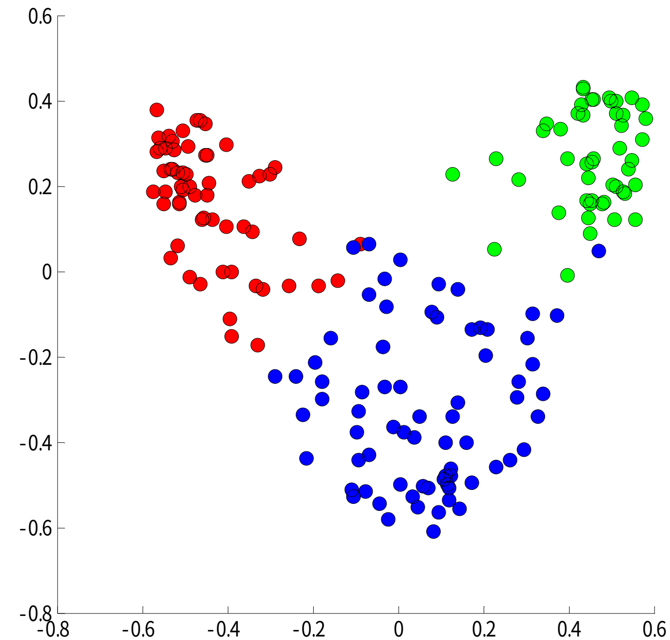
PCA (線形)



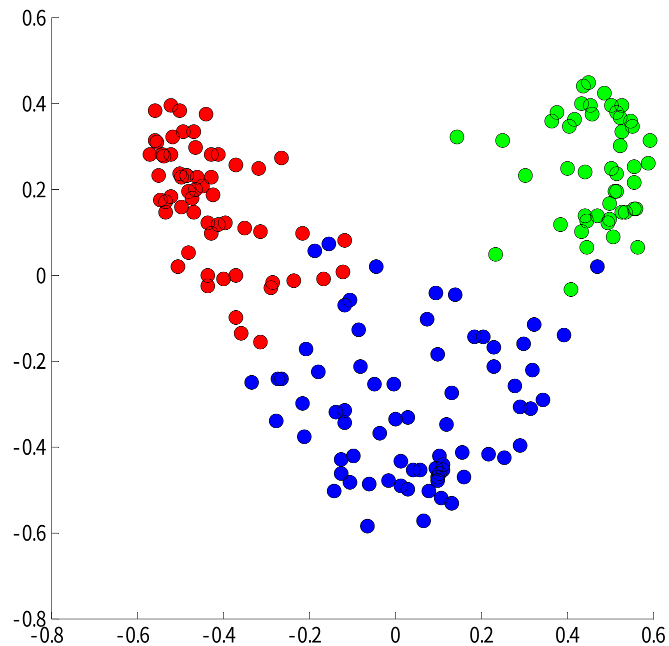
KPCA (RBF, $\sigma = 3$)



KPCA (RBF, $\sigma = 2$)



KPCA (RBF, $\sigma = 4$)



KPCA (RBF, $\sigma = 5$)

- **カーネルPCAの特徴**
  - **非線形な方向でのデータのばらつきが扱える.**
  - **結果はカーネルの選び方に依存するので，解釈には注意が必要**
    ガウスカーネルの分散パラメータなど

    どうやって選ぶか？　→ 必ずしも明確でない
  - **前処理として使える**
    後の処理の結果を改良するための非線形特徴抽出

カーネルCCA

- **正準相関分析** (CCA, **復習**)

  CCA ・・・ **2種類の多次元データの相関を探る**

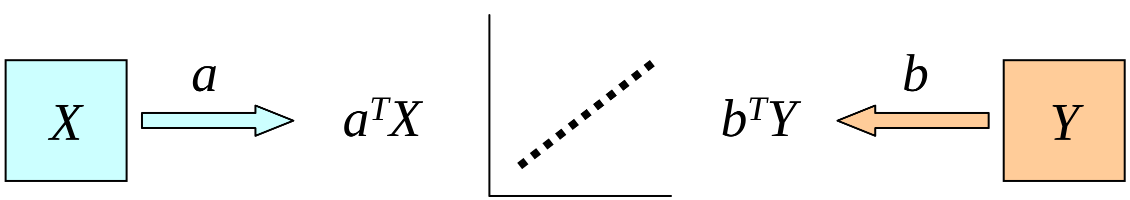
  m **次元データ** $X_1, \ldots, X_N$

  n **次元データ** $Y_1, \ldots, Y_N$

  X **を** a **方向,** Y **を** b **方向に射影したときに相関が大きくなる** (a,b) **を求める**

正準相関
$$\rho = \max_{\substack{a \in \mathbf{R}^m \\ b \in \mathbf{R}^n}} \frac{\frac{1}{N}\sum_i \left(a^T \tilde{X}_i\right)\left(b^T \tilde{Y}_i\right)}{\sqrt{\frac{1}{N}\sum_i \left(a^T \tilde{X}_i\right)^2}\sqrt{\frac{1}{N}\sum_i \left(b^T \tilde{Y}_i\right)^2}} = \max_{\substack{a \in \mathbf{R}^m \\ b \in \mathbf{R}^n}} \frac{a^T V_{XY} b}{\sqrt{a^T V_{XX} a}\sqrt{b^T V_{YY} b}}$$

ただし $V_{XY} = \frac{1}{N}\sum_i \tilde{X}_i \tilde{Y}_i^T$ **など**



$$\rho = \max_{\substack{a \in \mathbf{R}^m \\ b \in \mathbf{R}^n}} \frac{\left(a^T V_{XX}^{1/2}\right)\left(V_{XX}^{-1/2} V_{XY} V_{YY}^{-1/2}\right)\left(V_{YY}^{1/2} b\right)}{\left\| V_{XX}^{1/2} a \right\| \left\| V_{YY}^{1/2} b \right\|} = \max_{\substack{u \in \mathbf{R}^m \\ v \in \mathbf{R}^n}} \frac{u^T \left(V_{XX}^{-1/2} V_{XY} V_{YY}^{-1/2}\right) v}{\|u\| \, \|v\|}$$

特異値分解

$$V_{XX}^{-1/2} V_{XY} V_{YY}^{-1/2} = U \Lambda V^T$$

$$U = (u_1, \ldots, u_m) \qquad V = (v_1, \ldots, v_n)$$

$$\Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \ddots & & O \\ & & \lambda_m & \end{pmatrix} \qquad \begin{array}{c} \lambda_1 \geq \cdots \geq \lambda_\ell \geq 0 \\ \ell = \min\{m, n\} \end{array}$$

$$\Rightarrow \quad \begin{cases} a = V_{XX}^{-1/2} u_1 \\ b = V_{YY}^{-1/2} v_1 \end{cases}$$

$$\rho = \lambda_1$$

カーネルCCA (Akaho 2001, Bach and Jordan 2002)

データ $X_1, \ldots, X_N$ → 特徴ベクトル $\phi^X{}_1, \ldots, \phi^X{}_N$

$Y_1, \ldots, Y_N$ 　　　　　　　　　$\phi^Y{}_1, \ldots, \phi^Y{}_N$

カーネル k_X, k_Y を設定

$$\phi_j^X = k_X(\cdot, X_j) \in H_{k_X}$$

$$\phi_j^Y = k_Y(\cdot, Y_j) \in H_{k_Y}$$

- カーネルCCA：特徴空間での相関を最大化する射影方向 f, g を求める

$$\rho = \max_{\substack{f \in H_{k_X} \\ g \in H_{k_Y}}} \frac{\frac{1}{N}\sum_i \langle f, \tilde{\phi}_i^X \rangle_{H_{k_X}} \langle g, \tilde{\phi}_i^Y \rangle_{H_{k_Y}}}{\sqrt{\frac{1}{N}\sum_i \langle f, \tilde{\phi}_i^X \rangle_{H_{k_X}}^2} \sqrt{\frac{1}{N}\sum_i \langle g, \tilde{\phi}_i^Y \rangle_{H_{k_Y}}^2}}$$

カーネルPCA同様 $f = \sum_{\ell=1}^N \alpha_\ell \tilde{\phi}_\ell^X, \quad g = \sum_{\ell=1}^N \beta_\ell \tilde{\phi}_\ell^Y$ としてよい.

$$\Rightarrow \quad \rho = \max_{\substack{\alpha \in \mathbf{R}^N \\ \beta \in \mathbf{R}^N}} \frac{\alpha^T \tilde{K}_X \tilde{K}_Y \beta}{\sqrt{\alpha^T \tilde{K}_X^2 \alpha} \sqrt{\beta^T \tilde{K}_Y^2 \beta}}$$

- 正則化

ところが，，，　$\tilde{K}_X, \tilde{K}_Y$ はゼロ固有値を持つので，正則化を施す

結局：

$$\tilde{\rho} = \max_{\substack{\alpha\in\mathbf{R}^N \\ \beta\in\mathbf{R}^N}} \frac{\alpha^T \tilde{K}_X \tilde{K}_Y \beta}{\sqrt{\alpha^T \left(\tilde{K}_X + \varepsilon I_N\right)^2 \alpha} \sqrt{\beta^T \left(\tilde{K}_Y + \varepsilon I_N\right)^2 \beta}}$$

⇒

$$\left(\tilde{K}_X + \varepsilon I_N\right)^{-1} \tilde{K}_X \tilde{K}_Y \left(\tilde{K}_Y + \varepsilon I_N\right)^{-1} = U \Lambda V^T$$ 特異値分解

$$U = (u_1, \ldots, u_N) \quad V = (v_1, \ldots, v_N)$$

$$\Lambda = \begin{pmatrix} \lambda_1 & & O \\ & \ddots & \\ O & & \lambda_N \end{pmatrix} \quad \lambda_1 \geq \cdots \geq \lambda_N \geq 0$$

⇒

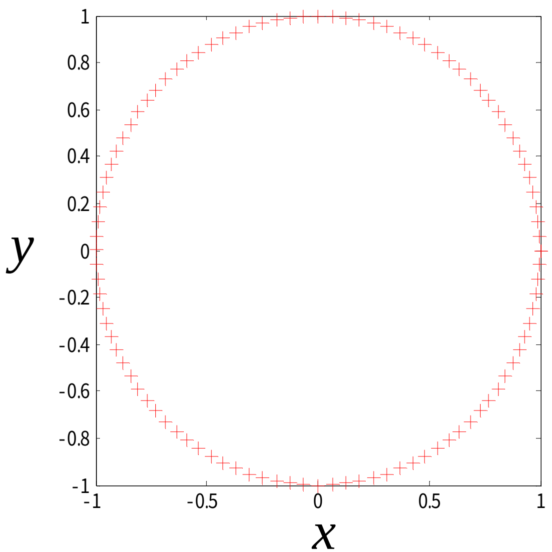
$$\alpha = \left(\tilde{K}_X + \varepsilon I_N\right)^{-1} u_1$$

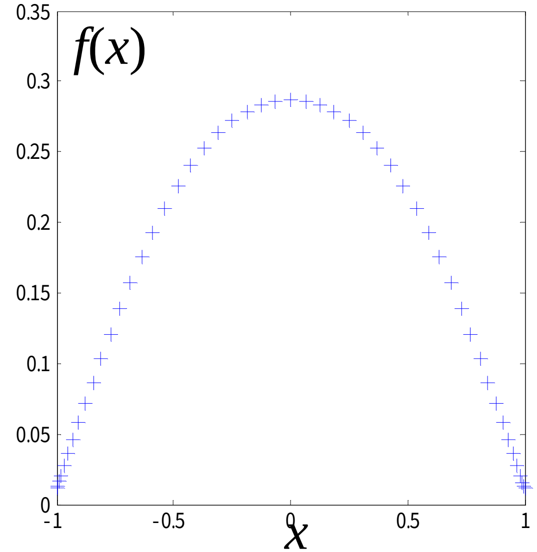
$$\beta = \left(\tilde{K}_Y + \varepsilon I_N\right)^{-1} v_1$$

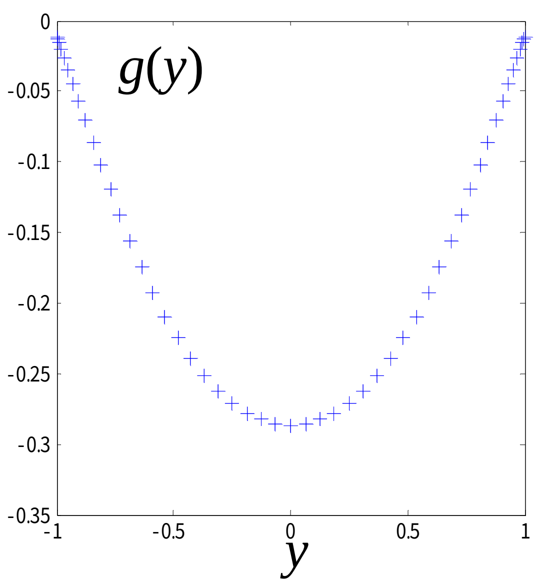
$$\boxed{f = \sum_{i=1}^N \alpha_i \tilde{k}_X(\cdot, X_i), \qquad g = \sum_{i=1}^N \beta_i \tilde{k}_Y(\cdot, Y_i)}$$

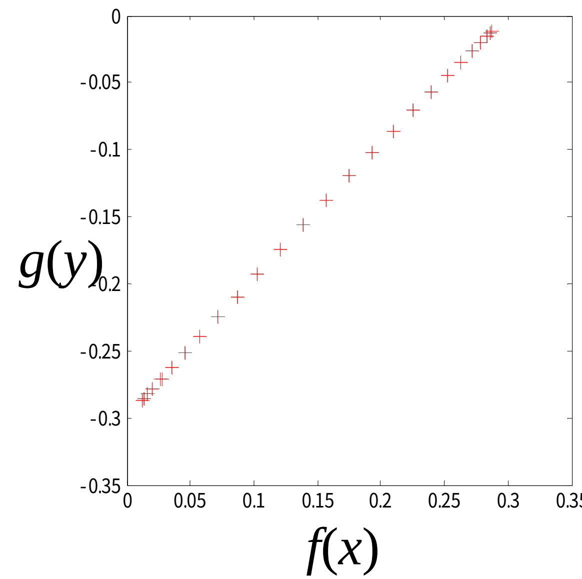
カーネルCCAの実験例

ガウスカーネル









SVM: マージン最大化による2値識別

- **2クラス識別問題**

データ　$(X^1, Y^1), \ldots, (X^N, Y^N)$　　$X_i \in \mathbf{R}^m$

$Y_i \in \{1, -1\}$　・・・ **2クラスのクラスラベル**

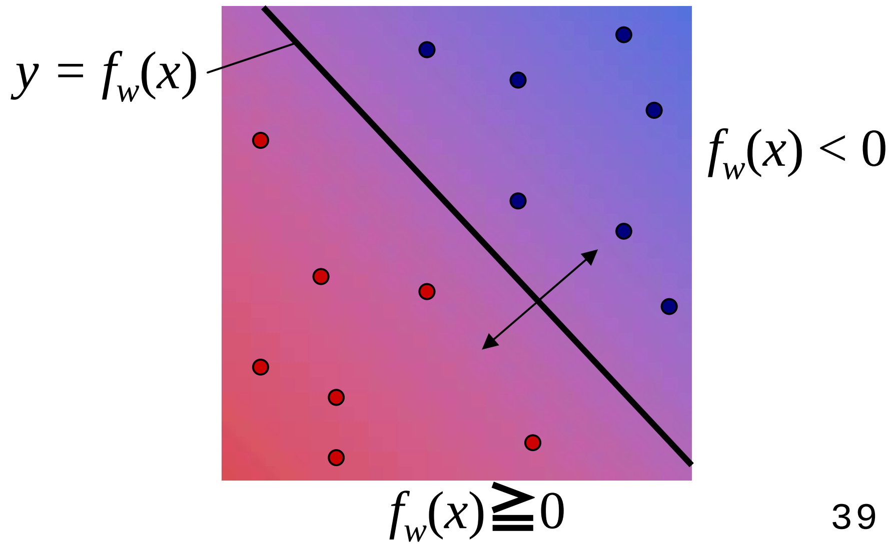
線形識別関数

$$f_w(x) = a^T x + b$$

$$w = (a, b)$$

$$\begin{cases} f_w(x) \geq 0 & \Rightarrow \quad y = 1 \text{ (クラス1)と判定} \\ f_w(x) < 0 & \Rightarrow \quad y = -1 \text{ (クラス2)と判定} \end{cases}$$

問題：　**未知の** x **に対しても**
正しく答えられるように
$f_w(x)$　**を構成せよ**

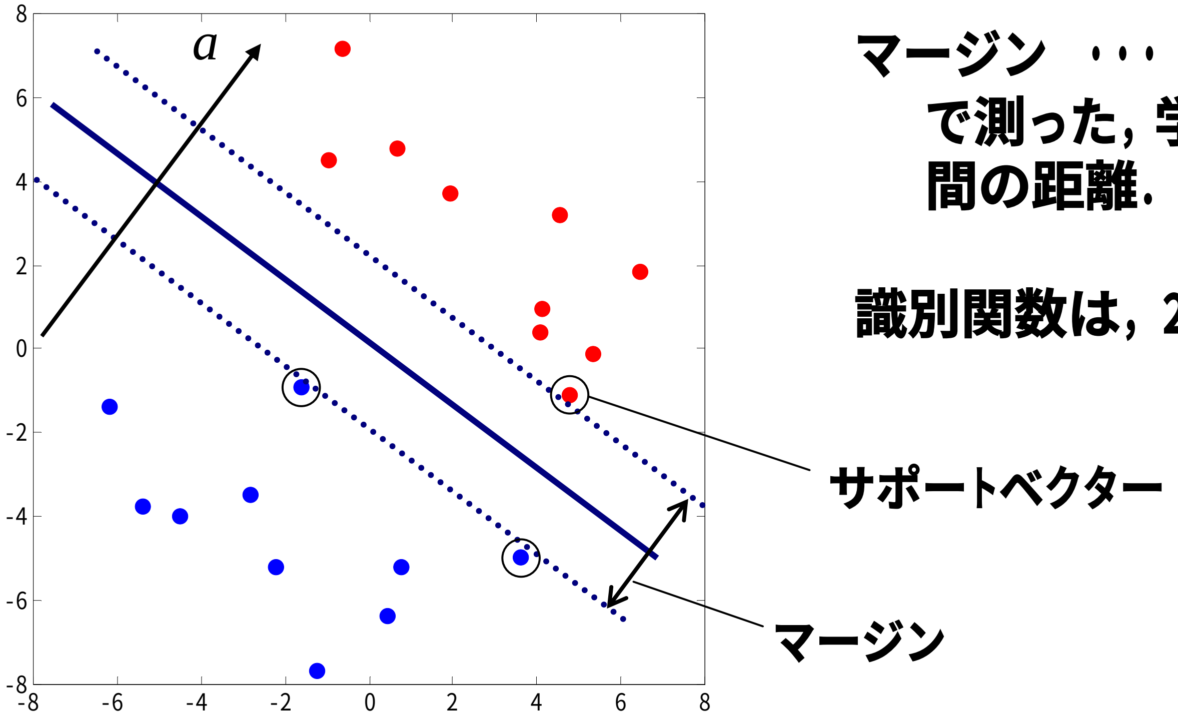


■ マージン最大化

学習データは線形識別が可能と仮定

⇒　学習データを分類する線形識別関数は無数にある.

マージンを最大化する方向を選ぶ



マージン　・・・　ベクトル a の方向で測った, 学習データのクラス間の距離.

識別関数は, 2つの境界の真中

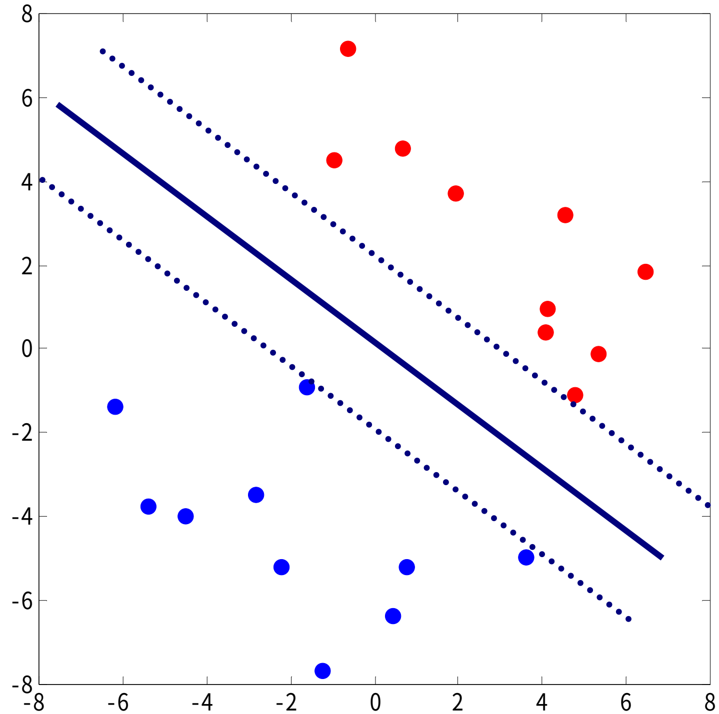
マージンの計算

(a,b) を定数倍しても識別境界は不変なので，スケールを一つ決める

$$\begin{cases} \min(a^T X^i + b) = 1 & Y^i = 1 \text{ のとき} \\ \max(a^T X^i + b) = -1 & Y^i = -1 \text{ のとき} \end{cases}$$

⇒

$$\text{マージン} = \frac{2}{\|a\|}$$



■ マージン最大化識別関数

$$\max_{a,b} \frac{1}{\|a\|}$$

制約条件

$$\begin{cases} \min\limits_{i:\,Y_i=1}(a^T X^i + b) = 1 \\ \min\limits_{i:\,Y_i=-1}\left(-(a^T X^i + b)\right) = 1 \end{cases}$$

$$\min_{a,b} \|a\|^2$$

制約条件

$$Y_i(a^T X^i + b) \geq 1 \quad (\forall i)$$

2次最適化：有効な最適化アルゴリズムの利用が可能
数値的に，必ず解を得ることができる

- ソフトマージン

線形識別可能の仮定は強すぎるので，少し弱める

ハードな制約条件

$$Y_i(a^T X^i + b) \geq 1$$

→

ソフトな制約条件

$$Y_i(a^T X^i + b) \geq 1 - \xi_i \qquad (\ \xi_i \geqq 0\ )$$

ソフトマージンの識別関数

$$\min_{a,b,\xi_i} \| a \|^2 + C\sum_{i=1}^N \xi_i$$

制約条件

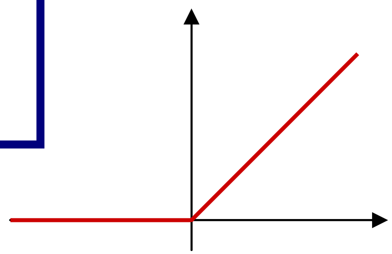
$$Y_i(a^T X^i + b) \geq 1 - \xi_i$$

$$\xi_i \geqq 0$$

- 正則化問題としての表現

$$\boxed{\min_{a,b} \quad \sum_{i=1}^N \left(1 - Y^i(a^T X^i + b)\right)_+ + \frac{\lambda}{2}\| a \|^2}$$

ただし $(z)_+ = \max(z, 0)$



サポートベクターマシン

- **特徴空間でのソフトマージン最大化**

線形の場合 $\min_{a,b} \quad \sum_{i=1}^{N}\left(1-Y^i(\underline{a^T X^i}+b)\right)_+ + \frac{\lambda}{2}\underline{\|a\|^2}$

線形の関数

カーネル k **を用意**

⇩

非線形化 $\min_{f\in H,b} \quad \sum_{i=1}^{N}\left(1-Y^i(\underline{f}(X^i)+b)\right)_+ + \frac{\lambda}{2}\underline{\|f\|_H^2}$

RKHS**の元**

H : **正定値カーネル** k **により定まる再生核ヒルベルト空間**

- Representer 定理による解の構成

最適化問題
$$\min_{f,b} \quad \sum_{i=1}^{N}\left(1-Y^{i}(f(X^{i})+b)\right)_{+} + \frac{\lambda}{2}\| f \|_{H}^{2}$$

の解は
$$f(x) = \sum_{i=1}^{N}\alpha_i k(x, X^{i})$$

の形で与えられる (Representer theorem, **セクション4で詳しく述べる**)

$$\min_{\alpha,b} \quad \sum_{i=1}^{N}\left(1-Y^{i}\sum_{j=1}^{N}\alpha_j k(X^{i},X^{j})+b\right)_{+} + \frac{\lambda}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j k(X^{i},X^{j})$$

あるいは, ソフトマージンに戻って

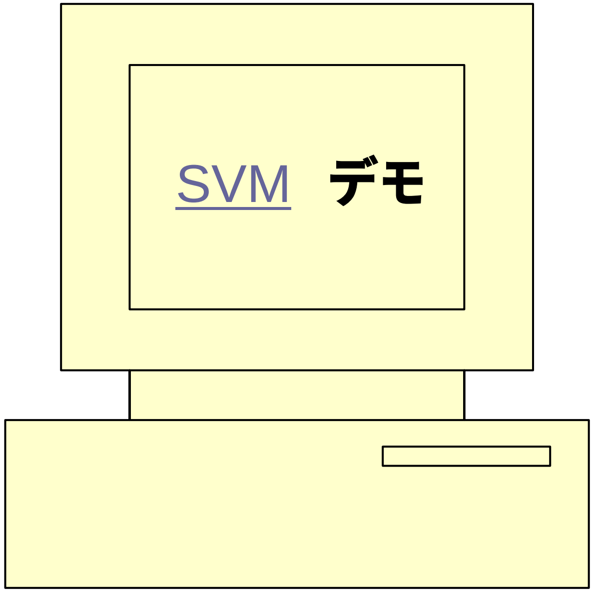
$$\min_{\alpha,b,\xi_i} \alpha^{T}K\alpha + C\sum_{i=1}^{N}\xi_i$$

制約条件
$$Y_i((K\alpha)_i + b) \geq 1-\xi_i$$
$$\xi_i \geqq 0$$

2次最適化問題として解ける

$K_{ij} = k(X_i, X_j)$ **グラム行列**

- SVMの例
  - ガウスカーネル
  - 複雑な識別境界が実現可能



http://svm.dcs.rhbnc.ac.uk/

リッジ回帰とスプライン平滑化

- **線形回帰(復習)**

$$(X^1,Y^1),\dots,(X^N,Y^N) \qquad X^i \in \mathbf{R}^m,\ Y^i \in \mathbf{R}$$

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 $X = \begin{pmatrix} X_1^1 & \cdots & X_m^1 \\ X_1^2 & \cdots & X_m^2 \\ \vdots & & \vdots \\ X_1^N & \cdots & X_m^N \end{pmatrix}, \qquad Y = \begin{pmatrix} Y^1 \\ Y^2 \\ \vdots \\ Y^N \end{pmatrix}$ **を使うと,**

最適解は $\hat{w} = (X^T X)^{-1} X^T Y$

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

- **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、** $X^T X$ **が特異になるときに特に有効**
  - Bayes**的な解釈もできる** (**縮小推定**)

スプライン平滑化

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

カーネル k
を用意

↓

非線形化　$\min_{f \in H} \sum_{i=1}^{N} \left(Y^i - f(X^i)\right)^2 + \lambda \| f \|_H^2$　・・・ スプライン平滑化

H : 正定値カーネル k により定まる再生核ヒルベルト空間

解の構成

Representer theorem より　$f(x) = \sum_{i=1}^{N} \alpha_i k(x, X^i)$　の形なので、

$$\min_{\alpha} \|Y - K\alpha\|^2 + \lambda \alpha^T K \alpha$$ を解けばよい

解　$f(x) = Y^T (K + \lambda I_N)^{-1} g(x)$

$K_{ij} = k(X^i, X^j)$　グラム行列

$g_i(x) = k(x, X^i)$

正則化としてのスプライン平滑化

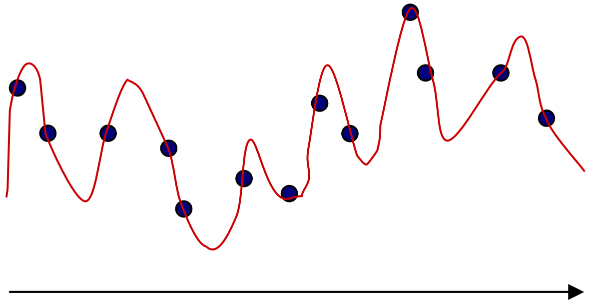
2乗誤差によるカーブフィッティング

$$\min_f \quad \sum_{i=1}^N \left(Y^i - f(X^i)\right)^2$$

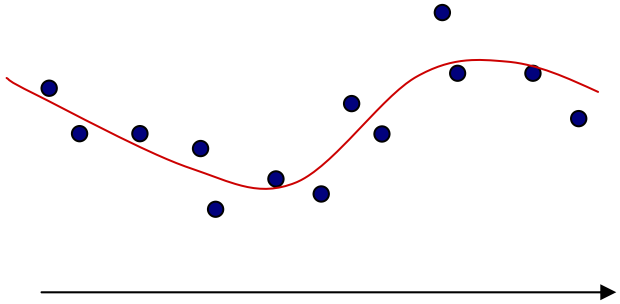
正則化

$$\min_f \quad \sum_{i=1}^N \left(Y^i - f(X^i)\right)^2 + \boxed{\Psi(f)}$$

誤差 0 となる f は無数にある



解が一意になるように正則化項を付加



- 滑らかさによる正則化 (平滑化)

$$\min_f \quad \sum_{i=1}^N \left(Y^i - f(X^i)\right)^2 + \lambda \int \left\{ c_1 \left| \frac{df(x)}{dx} \right|^2 + \cdots + c_m \left| \frac{d^m f(x)}{dx^m} \right|^2 \right\} dx$$

$(c_i \geqq 0)$

実は、微分が2乗可積分な関数全体

＝　ある正定値カーネル k に対する再生核ヒルベルト空間

かつ

$$\int \left\{ c_1 \left| \frac{df(x)}{dx} \right|^2 + \cdots + c_m \left| \frac{d^m f(x)}{dx^m} \right|^2 \right\} dx = \| f \|_H^2$$

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

- ガウスカーネルも、ある種の微分作用素に対応する正定値カーネル
  RBFスプライン
- SVMも同様の正則化
  ただし2乗誤差ではない

$$\min_{f,b} \quad \sum_{i=1}^N \left(1 - Y^i (f(X^i) + b)\right)_+ + \lambda \| f \|_H^2$$

セクション3のまとめ

- カーネルによる非線形化
  - 線形データ解析アルゴリズムを特徴空間で行うことによって
    非線形アルゴリズムが得られる　→　カーネル化 (kernelization)
  - 「内積」を使って表される線形手法なら拡張が可能
    射影, 相関, 分散共分散, etc
  - 例:　サポートベクターマシン, スプライン平滑化, カーネルPCA,
    カーネルCCA, など

- 非線形アルゴリズムの特徴
  - 線形ではとらえられない性質が調べられる.
  - とらえることのできる非線形性はカーネルの選び方に影響を受ける

4. 正定値カーネルの性質

- このセクションの目的
  - 正定値性を保つ演算と正定値性のチェック
  - R^m上の正定値カーネルの特徴づけ： Bochnerの定理
  - 無限次元の問題を有限次元へ還元： Representer定理

正定値性を保つ演算

- **和と積**

$k_1(x,y)$, $k_2(x,y)$ ：Ω 上の正定値カーネル．次も正定値カーネル

(1) 非負実数 a_1 と a_2 に対する線形和　$a_1 k_1(x,y) + a_2 k_2(x,y)$

(2) 積　$k_1(x,y) k_2(x,y)$

- **正定値カーネルの収束列**

$k_1(x,y)$, $k_2(x,y)$, …, $k_n(x,y)$, … を Ω 上の正定値カーネル列とするとき

$$k(x,y) = \lim_{n\to\infty} k_n(x,y) \qquad (\text{任意の } x, y \in \Omega)$$

ならば，$k(x,y)$ も正定値カーネル

- Ω 上の正定値カーネル全体は，積について閉じた（各点位相での）閉凸錐

- **正規化**
  - $k(x,y)$ は Ω 上の正定値カーネル, $f:\Omega\to\mathbf{R}$ は任意の関数とすると,

$$\tilde{k}(x,y) = f(x)k(x,y)f(y)$$

  は正定値カーネル

  - 特に正定値カーネルが $k(x,x) > 0$ （$x\in\Omega$ は任意）を満たすとき,

$$\tilde{k}(x,y) = \frac{k(x,y)}{\sqrt{k(x,x)k(y,y)}}$$

  は正定値カーネル ・・・ normalized カーネル

例)

$$k(x,y) = (x^T y + c)^d \quad (c>0) \quad \Longrightarrow \quad \tilde{k}(x,y) = \frac{(x^T y + c)^d}{(x^T x + c)^{d/2}(y^T y + c)^{d/2}}$$

- **証明**
  - **和・収束列　⇒ グラムル行列の半正定値性は明らか.**
  - **正規化:**

$x_1, \ldots, x_n \in \Omega$, $c_1, \ldots, c_n \in \mathbf{R}$

$$\sum_{i,j=1}^n c_i c_j \tilde{k}(x_i, x_j) = \sum_{i,j=1} c_i c_j f(x_i) k(x_i, x_j) f(x_j)$$

$$= \sum_{i,j=1} \underbrace{\boxed{c_i f(x_i)}}_{d_i} \underbrace{\boxed{c_j f(x_j)}}_{d_j} k(x_i, x_j) \quad \geq 0$$

  - **積:** $\left(k_1(x_i, x_j)\right)_{i,j=1}^n$ **は半正定値なので, 対角化により**

$$k_1(x_i, x_j) = \sum_{p=1}^n \lambda_p U_p^i U_p^j \qquad (\lambda_p \geqq 0)$$

$$\Rightarrow \sum_{i,j=1}^n c_i c_j k_1(x_i, x_j) k_2(x_i, x_j)$$

$$= \sum_{p=1}^n \sum_{i,j=1}^n c_i c_j \lambda_p U_p^i U_p^j k_2(x_i, x_j)$$

$$= \lambda_1 \left(\sum_{i,j=1}^n c_i U_1^i c_j U_1^j k_2(x_i, x_j) \right) + \cdots + \lambda_n \left(\sum_{i,j=1}^n c_i U_n^i c_j U_n^j k_2(x_i, x_j) \right) \geq 0$$

カーネルの設計：Marginalized kernel

隠れた構造を利用

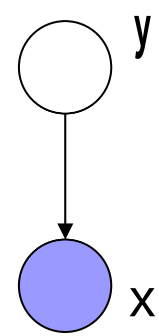
$z = (x, y)$

x : **観測される変数**

y : **観測されない隠れ変数**

$p(x,y)$ ： (x, y) **に対する確率モデル**

$k_z(z_1, z_2)$ ： z **に対する正定値カーネル**



e.g.) HMM

系列

$$k(x_1, x_2) = \sum_{y_1}\sum_{y_2} p(y_1 \mid x_1)\, p(y_2 \mid x_2)\, k_z((x_1, y_1), (x_2, y_2))$$

＊ $k_z(z_1, z_2) = k(y_1, y_2)$ （y **のみに依存**） **でも**OK

正定値性の判定

- **正定値カーネルの例： 正定値性の証明**
  - **多項式カーネル**

    $x^T y$ **は正定値**　∵）　$\sum_{i,j=1}^{n} c_i c_j x_i^T x_j = \left(\sum_{i=1}^{n} c_i x_i\right)^T \left(\sum_{j=1}^{n} c_j x_j\right) \geq 0$

    ⇒　$x^T y + c$ **は正定値**（$c \geqq 0$）

    ⇒　$(x^T y + c)^d$ **は正定値**（d **個の積ゆえ**）

  - Fourier **カーネル**

$$\exp\left(\sqrt{-1}\omega^T (x-y)\right) = \exp\left(\sqrt{-1}\omega^T x\right)\exp\left(-\sqrt{-1}\omega^T y\right)$$

$$= f(x)\overline{f(y)}$$

⇒　**正定値**

ガウスカーネル

- **復習**

$$\exp\left(-\frac{\|x\|^2}{2}\right) = \frac{1}{(2\pi)^m}\int_{R^m} \exp\left(-\frac{\|\omega\|^2}{2}\right)\exp\left(\sqrt{-1}\omega^T x\right)d\omega$$

ガウス関数の Fourier**変換は，またガウス関数**

- **正定値性の証明**

$$\sum_t \underbrace{\exp\left(-\frac{\|\omega_t\|^2}{2}\right)}_{\text{正の数}} \underbrace{\exp\left(\sqrt{-1}\omega_t{}^T (x-y)\right)}_{\text{正定値カーネル}} \rightarrow \exp\left(-\frac{\|x-y\|^2}{2}\right)$$

⇓

正定値

$\mathbf{R}^m$上の正定値カーネル

- Bochner の定理

  $k(x,y) = \phi(x-y)$ の形により与えられる $\mathbf{R}^m$ 上のカーネルが正定値であるための必要十分条件は，関数 $\phi(z)$ が, ある非負可積分関数 $f(z)$ によって

$$\phi(z) = \int_{\mathbf{R}^m} \underbrace{f(z)}_{\text{非負}} \underbrace{\exp\left(\sqrt{-1}\omega^T z\right)}_{\text{Fourierカーネル(正定値)}} d\omega$$

  と表されることである.

  すなわち，$\phi(z)$ のFourier変換が非負実数関数となることである.

  - 上の積分表示を持つ $\phi(z)$ が正定値カーネルを与えることは, ガウスの場合と同様. Bochnerの定理は, その逆も成り立つことを主張している.

  - Fourierカーネル $\exp\left(\sqrt{-1}\omega^T(x-y)\right)$ が, 正定値カーネル全体の成す閉凸錐を張っている.

Representer Theorem

- **正則化の問題 (復習)**
  - **スプライン平滑化** $\min_{f\in H} \sum_{i=1}^{N}\left(Y^i - f(X^i)\right)^2 + \lambda \| f \|_H^2$
  - SVM $\min_{f,b} \sum_{i=1}^{N}\left(1 - Y^i (f(X^i) + b)\right)_+ + \lambda \| f \|_H^2$

- **一般化された問題**

k: **正定値カーネル,** H : k **により定まる再生核ヒルベルト空間**

$x_1, \ldots, x_N, \; y_1, \ldots, y_N$ ： **データ (固定)**

$h_1(x), \ldots, h_d(x)$ ： **固定された関数**

(∗) $$\min_{\substack{f \in H \\ (b_\ell) \in \mathbf{R}^d}} L\left(\{x_i\}_{i=1}^N, \{y_i\}_{i=1}^N, \left\{f(x_i) + \sum_{\ell=1}^{d} b_\ell h_\ell(x_i)\right\}_{i=1}^N\right) + \Psi\left(\| f \|_H^2\right)$$

- Representer Theorem

正則化項の関数 Ψ は, $[0, \infty)$ 上の単調増加関数とする.

$\tilde{H}_N = \mathrm{span}\{k(\cdot, x_1), \ldots, k(\cdot, x_N)\}$　　$\{k(\cdot, x_i)\}$ の張る N 次元部分空間

⇒

(∗) の解 f は $\tilde{H}_N$ の中にある.　すなわち

$$f(x) = \sum_{i=1}^{N} \alpha_i k(x, x_i)$$

の形で探してよい.

$$\min_{f \in H \;\; (b_\ell) \in \mathbf{R}^d} L\left(\{x_i\}_{i=1}^N, \{y_i\}_{i=1}^N, \left\{f(x_i) + \sum_{\ell=1}^{d} b_\ell h_\ell(x_i)\right\}_{i=1}^N\right) + \Psi\left(\|f\|_H^2\right)$$

$$= \min_{(\alpha_i) \in \mathbf{R}^N \;\; (b_\ell) \in \mathbf{R}^d} L\left(\{x_i\}_{i=1}^N, \{y_i\}_{i=1}^N, \left\{\sum_{j=1}^{N} K_{ij}\alpha_j + \sum_{\ell=1}^{d} b_\ell h_\ell(x_i)\right\}_{i=1}^N\right) + \Psi\left(\alpha^T K \alpha\right)$$

$K = (k(x_i, x_j))$：グラム行列

H (無限次元) 上の最適化が有限次元の最適化に変換できる

- Representer theorem **の証明**

$$\min \quad L\left(\{x_i\}_{i=1}^N, \{y_i\}_{i=1}^N, \left\{f(x_i) + \sum_{\ell=1}^d b_\ell h_\ell(x_i)\right\}_{i=1}^N\right) + \Psi\left(\| f \|_H^2\right)$$

$$H = \widetilde{H}_N \oplus H_\perp$$ **直交分解**

$$f = \widetilde{f}_N + f_\perp$$

$$\langle f_\perp, k(\cdot, x_i)\rangle = 0 \quad (\forall i)$$

⇨

- $$f(x_i) = \langle f, k(\cdot, x_i)\rangle = \langle \widetilde{f}_N + f_\perp, k(\cdot, x_i)\rangle = \langle \widetilde{f}_N, k(\cdot, x_i)\rangle = \widetilde{f}_N(x_i)$$

  → L の値は $\widetilde{f}_N$ だけで決まる

- $$\| f \|_H^2 = \| \widetilde{f}_N \|_H^2 + \| f_\perp \|_H^2$$

  → Ψ の値は $f_\perp = 0$ のほうがよい

⇨

$f \in \widetilde{H}_N$ **に最適解がある**

(証明終)

セクション4のまとめ

- **正定値性を保つ演算**
  - **非負の和, 積**
  - Normalization
  - **正定値カーネル列の各点収束**

  ⇒ **あたらしいカーネルの作成／カーネルの正定値性のチェック**

- Bochner **の定理**

  $\phi(x-y)$ **型の**$\mathbf{R}^m$ **上のカーネルが正定値** ⇔ ϕ **の**Fourier**変換が非負**

- Representer thoerem

  無限次元空間上の正則化問題を有限次元の問題へ

5. 構造化データのカーネル

- このセクションの目的
  - 複雑な構造を持つデータ（ストリング，ツリー，グラフ）に対して定義されるカーネルとその計算法を紹介する
  - 構造化データが使われる応用を紹介する

構造化データの処理

- カーネルの利用

正定値カーネル $k(x, y)$： x, y はベクトルデータでなくてよい

- どんなデータでもOK
  - 長さの違うシンボル列　＝　ストリング
  - ツリー構造
  - グラフ表現されたデータ
- カーネル法　→ 非ベクトルデータのベクトル化
  カーネルが定義されると, SVM，カーネルPCA, スプライン平滑化などの利用が可能
- 計算すべきもの ＝ データに対するグラム行列　$k(x_i, x_j)$

ストリング

- ストリング
  - アルファベット Σ： 有限集合
  - ストリング： Σ の要素の有限長の列
    - 例） $\Sigma = \{ a, b, c, d, \ldots, z \}$
      ストリング `cat, head, computer, xyydyaa,…`
  - Σ^p： 長さ p のストリング全体
  - Σ^*：任意の長さのストリング全体　$\Sigma^* = \bigcup_{p=0}^{\infty} \Sigma^p$

注） $\Sigma^0 = \{\varepsilon\}$ ：空ストリング

  - 記号法

s： $s_1\ s_2 \ldots s_n$　ストリングに対し

| s |　・・・　ストリング s の長さ = n

s[i : j]　・・・　$s_i \ldots s_j$　という s の部分列

s, t に対し結合 s t = $s_1\ s_2 \ldots s_n\ t_1\ t_2 \ldots t_m$

ストリングカーネル

- ストリングカーネル
  - Σ^*上の定義された正定値カーネル・・・ 2つのストリング s, t の類似度
  - 一致する部分列を数え上げるタイプが多い
  - 効率的な計算の工夫が重要 ・・・ 再帰式 (漸化式)など
    Dynamical Programming (DP)

- 典型的な応用先
  - 自然言語処理
    - 文字列: $\Sigma = \{a, b, c, \ldots, z\}$
    - 単語列: Σ = {単語全体}
  - ゲノム解析
    - ゲノム: Σ = {A, T, G, C}
    - タンパク質: Σ = {アミノ酸} (20種類)

ストリングカーネルの応用

- ゲノム配列のアラインメント
- タンパク質の構造予測
  - アミノ酸配列： Σ = 20種のアミノ酸

7LES_DROME	LKLLRFLGSGAFGEVYEGQLKTE....DSEEPQRVAIKSLRK.......
ABL1_CAEEL	IIMHNKLGGGQYGDVYEGYWK........RHDCTIAVKALK........
BFR2_HUMAN	LTLGKPLGEGCFGQVVMAEAVGIDK.DKPKEAVTVAVKMLKDD.....A
TRKA_HUMAN	IVLKWELGEGAFGKVFLAECHNLL...PEQDKMLVAVKALK........

配列 → 立体構造のクラスを予測

  - データベース

SCOP (Structural Classification of Proteins) など

p-スペクトラムカーネル

- 長さ p の部分列の出現回数を特徴ベクトルとする

$$|\Sigma| = m, \quad u \in \Sigma^p$$

$$\phi_u^p(s) = \left|\{(w_1, w_2) \in \Sigma^* \times \Sigma^* \mid s = w_1 u w_2\}\right|$$ ・・・ s の中の u の出現回数

$$\Phi : \Sigma^* \to H \cong \mathbf{R}^{m^p}, \qquad \Phi^p(s) = \left(\phi_u^p(s)\right)_{u \in \Sigma^p}$$

特徴空間： 長さ p の列全体 ・・・ m^p 次元

$$K_p(s,t) = \sum_{u \in \Sigma^p} \phi_u^p(s)\phi_u^p(t) = \left\langle \Phi^p(s), \Phi^p(t) \right\rangle_H$$

s = “statistics”　　t = “pastapistan”

3-スペクトラム

s: sta, tat, ati, tis, ist, sti, tic, ics

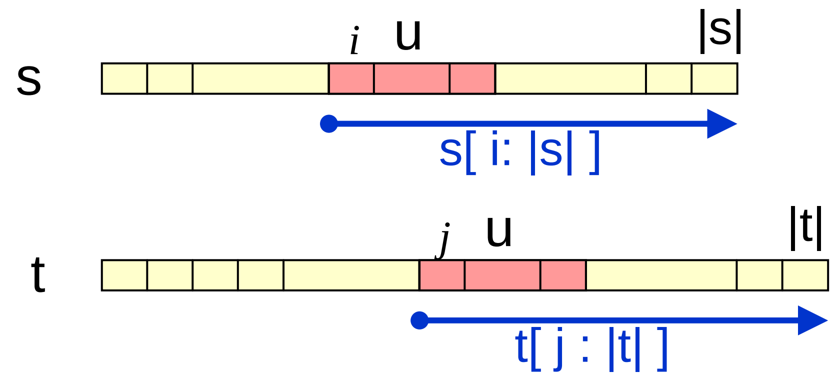
t : pas, ast, sta, tap, api, pis, ist, sta, tan

	sta	tat	ati	tis	ist	sti	tic	ics	pas	ast	tap	api	pis	tan
Φ(s)	1	1	1	1	1	1	1	1	0	0	0	0	0	0
Φ(t)	2	0	0	0	1	0	0	0	1	1	1	1	1	1

$K_3(s, t) = 1 \cdot 2 + 1 \cdot 1 = 3$

p-スペクトラムカーネルの計算法

$$K_p(s,t)=\sum_{u\in\Sigma^p}\phi_u^p(s)\phi_u^p(t) \quad =\left\langle \Phi^p(s),\Phi^p(t)\right\rangle_H$$



直接的な計算

部分列 u **を, 途中から始まる部分列** suffix **(接尾辞) の先頭** (prefix)**と思う**

$$h_u^p(s,t)=\begin{cases}1 & \text{s の } p\text{-prefix} = \text{t の } p\text{-prefix} \\ 0 & \text{s の } p\text{-prefix} \neq \text{t の } p\text{-prefix}\end{cases}$$

$$\Rightarrow \quad K_p(s,t)=\sum_{i=1}^{|s|-p+1}\sum_{j=1}^{|t|-p+1} h_p(s[i:i+p-1],t[j:j+p-1])$$

計算量 = O(p| s || t |)

- p-スペクトラムカーネルの計算法(II)

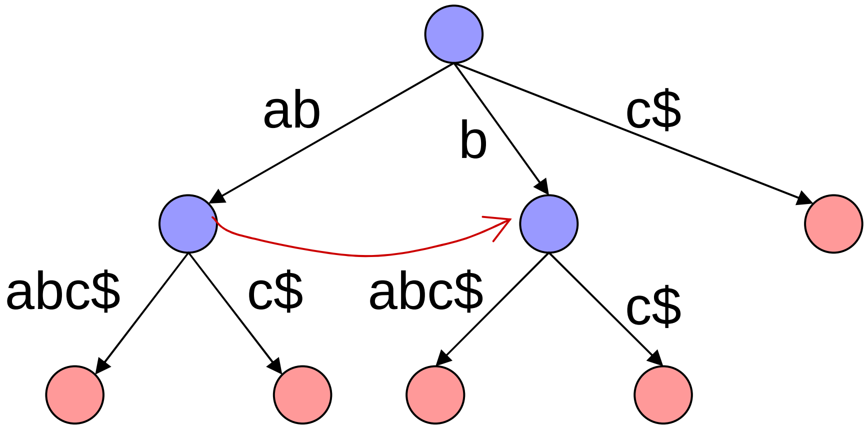
  実は, |s|+|t| に対して線形時間 O(p(| s | + | t |)) で計算する方法がある

  - Suffix Tree

    ストリングのすべての suffix を木構造で効率的に表すアルゴリズム

    例）ababc

ababc
babc
abc
bc
c



  - 詳しくは Vishwanathan & Smola 03, Gusfield 97.

All-subsequences Kernel

- すべての長さの部分列 (gapも許す) の出現回数を特徴ベクトルとする

$|\Sigma| = m, \quad u \in \Sigma^*$

$$\phi_u(s) = \left|\{\vec{i} \mid s[\vec{i}\,] = u\}\right|$$

ただし $\vec{i} = [i_1, i_2, i_3, \ldots, i_\ell] \quad (1 \le i_1 < i_2 < \cdots < i_\ell \le |s|)$ に対し

$$s[\vec{i}\,] = s_{i_1} s_{i_2} s_{i_3} \cdots s_{i_\ell}$$

s: `statsitics` $\quad \vec{i} = [2,3,9] \quad \Rightarrow \quad s[\vec{i}\,] =$ `tac`

特徴空間 ＝任意長のストリングを添字に持つベクトル空間 (無限次元)

$$\Phi : \Sigma^* \to H \cong \mathbf{R}^\infty, \qquad \Phi(s) = (\phi_u(s))_{u \in \Sigma^*}$$

$$\boxed{k(s,t) = \sum_{u \in \Sigma^*} \phi_u(s) \phi_u(t) \quad = \quad \langle \Phi(s), \Phi(t) \rangle}$$

- gapを許す比較

ATGACTAC → ATGACTAC

CATGCGATT → CATG CGATT

u = ATGCA

- 例

s: ATG

t : AGC

$K(s,t) = 4$

ε: 空ストリング

	ε	A	T	G	C	AT	AG	AC	TG	GC	ATG	AGC
Φ(s)	1	1	1	1	0	1	1	0	1	0	1	0
Φ(t)	1	1	0	1	1	0	1	1	0	1	0	1

- All-subsequences kernel **の計算**

再帰的な式: 過去に計算した結果を利用

初期条件

$k(s, \varepsilon) = k(t, \varepsilon) = 1$ （**任意の** s, t） ε **: 空ストリング**

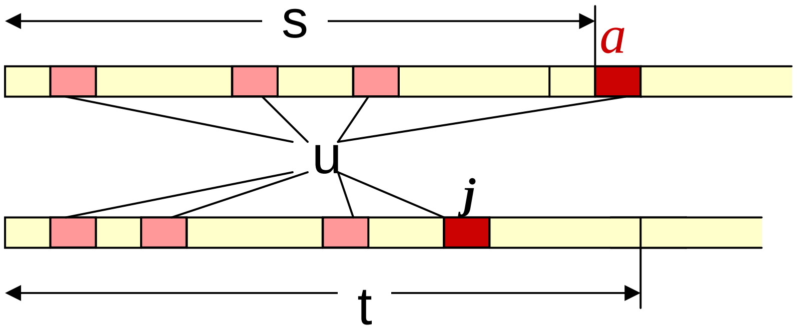
$k(s, t)$ **まで求まっているとして** $k(sa, t) = ?$

$k(sa, t)$

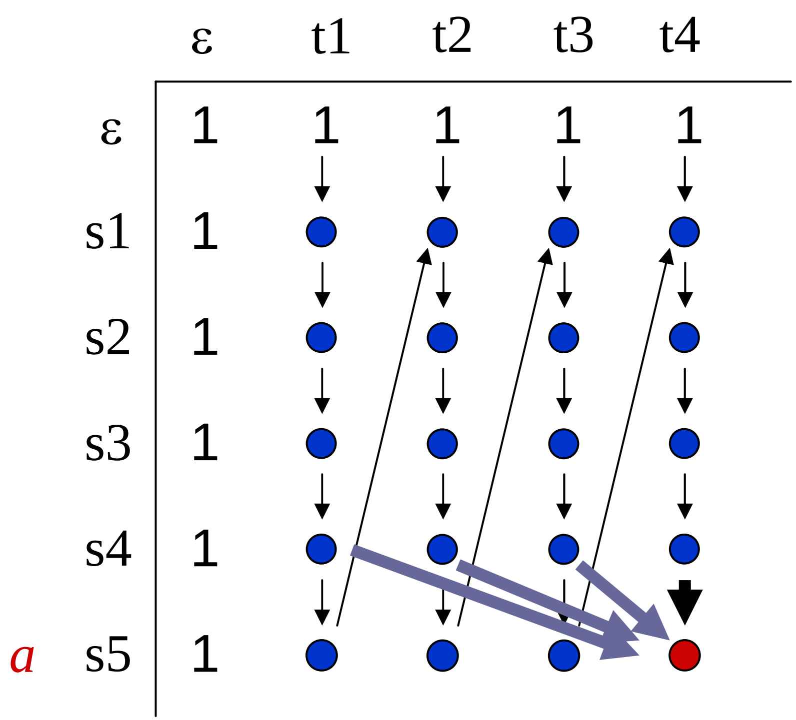
＝（s **内と** t **内での一致によるもの**）

＋ （a **を含む一致**）

$$= k(s,t) + \sum_{\substack{1 \leqq j \leqq |t| \\ j:\, t[j] = a}} k(s, t[1:j-1])$$



$$k(sa,t) = k(s,t) + \sum_{\substack{1 \leqq j \leqq |t| \\ j:\ t[j] = a}} k(s, t[1: j-1])$$



計算量 = O(| s | | t |^2)

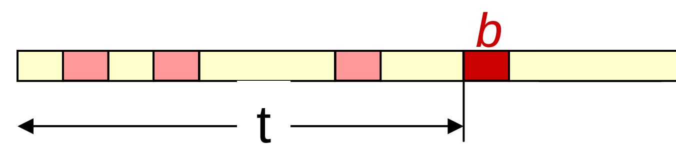
All-subsequences kernel の計算(II)

計算の効率化

$$k(sa,t) = k(s,t) + \boxed{\sum_{\substack{1 \leqq j \leqq |t| \\ j:\, t[j]=a}} k(s,t[1:j-1])}$$

t 回の計算をやりたくない

$$\| \quad \tilde{k}(sa,t)$$ とおくと



t に関する再帰式

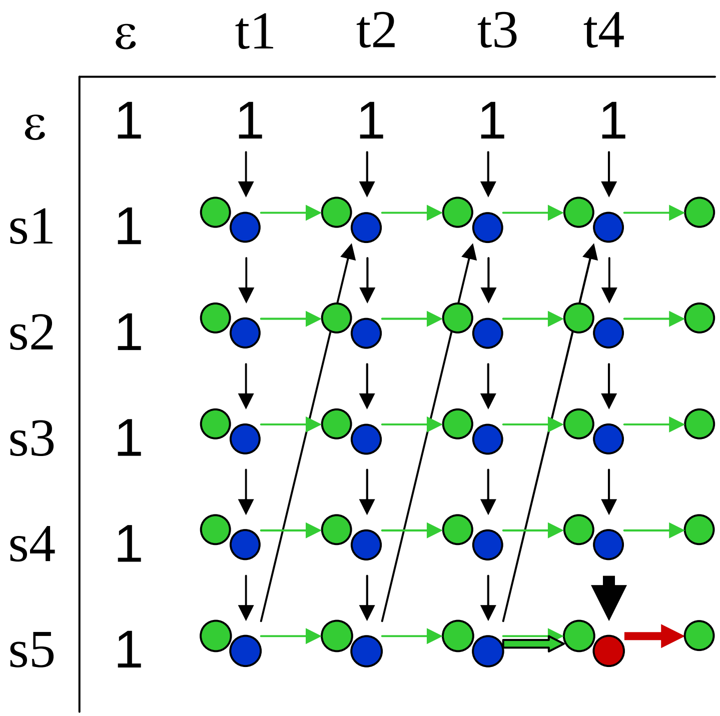
$$\tilde{k}(sa,tb) = \underbrace{\sum k(s,t[1:j-1])}_{1 \leqq j \leqq |t| \text{ の場合}} + \underbrace{\delta_{ab} k(s,t)}_{j=|tb| \text{ の場合}}$$

$$= \tilde{k}(sa,t) + \delta_{ab} k(s,t)$$

$$\begin{cases} k(sa,t) = k(s,t) + \tilde{k}(sa,t) & (s \text{ についての再帰式}) \\ \tilde{k}(sa,tb) = \tilde{k}(sa,t) + \delta_{ab} k(s,t) & (t \text{ についての再帰式}) \end{cases}$$

$$k(sa,t) = k(s,t) + \tilde{k}(sa,t)$$

$$\tilde{k}(sa,tb) = \tilde{k}(sa,t) + \delta_{ab} k(s,t)$$



計算量 = O(| s | | t |)

Gap-weighted subsequence kernel

- Gap に対してペナルティをつけた特徴ベクトル

  $|\Sigma| = m, \quad u \in \Sigma^p, \quad 0 < \lambda \leqq 1$

$$\phi_u^p(s) = \sum_{\vec{i}\,:\,u = s[\vec{i}\,]} \lambda^{\ell(\vec{i}\,)}$$

ただし $\vec{i} = [i_1, \ldots, i_r]$ に対し

$\ell(\vec{i}\,) = | s[i_1 : i_r] |$

C T G A C T G
u = CAT

$\Rightarrow \vec{i} = [1,4,6] \Rightarrow \ell(\vec{i}\,) = 6$

gap が多い一致は割り引く

$$\Phi : \Sigma^* \to H \simeq \mathbf{R}^{m^p}, \qquad \Phi^p(s) = \left(\phi_u^p(s)\right)_{u \in \Sigma^p}$$

特徴空間: 長さ p の列全体 ・・・ m^p 次元

$$K_p(s,t) = \sum_{u \in \Sigma^p} \phi_u^p(s)\phi_u^p(t) \quad = \left\langle \Phi^p(s), \Phi^p(t) \right\rangle_H$$

- Gap-weighted subsequences kernel **の例**

s: ATGC
t : AGCT
p = 2

	AT	AG	AC	TG	TC	GC	GT	CT
Φ(s)	λ^2	λ^3	λ^4	λ^2	λ^3	λ^2	0	0
Φ(t)	λ^4	λ^2	λ^3	λ^4	0	λ^2	λ^3	λ^2

$$k(s,t) = \lambda^4 + \lambda^5 + 2\lambda^6 + \lambda^7$$

- **効率的なアルゴリズムとして再帰式によるものがある**
  計算量 = O(p | s | | t |)

- **正規化が行われることが多い**

$$\tilde{k}_p(s,t) = \frac{k_p(s,t)}{\sqrt{k_p(s,s)k_p(t,t)}}$$

- **詳しくは** Lohdi et al. (JMLR, 2002)
  Rousu & Shawe-Taylor (2004)

他のストリングカーネル

- Fisher kernel: Jaakkola & Haussler (1999)
- Mismatch kernel: Leslie et al. (2003)

Marginalized kernel

- **確率モデルにもとづくカーネル設計**

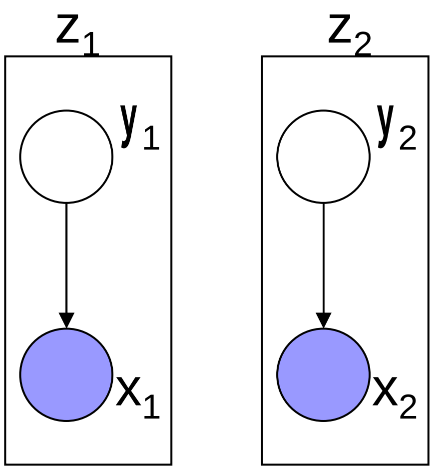
$z = (x, y)$

x : **観測される変数**

y : **観測されない隠れ変数**
(データを生成する構造)

$p(x,y)$ ： (x, y) **に対する確率モデル**

$k_z(z_1, z_2)$ ： z **に対する正定値カーネル**



y_1, y_2 **の状態全体**

$$k(x_1, x_2) = \boxed{\sum_{y_1}\sum_{y_2}} p(y_1 \mid x_1)\, p(y_2 \mid x_2)\, k_z((x_1, y_1), (x_2, y_2))$$

- 例

unknown

y_1 1 2 2 1 2 2 1 2 2

x_1 **A C G G T T C A A**

known

unknown

y_2 1 2 2 1 2 2 1

x_2 **A C C G T A C**

known

exon / intron

DNA

- $p(x, y)$ は隠れマルコフモデル (HMM) によって記述済み (y: 隠れ状態)
- $$k_z(z_1, z_2) = \frac{1}{|z_1||z_2|} \sum_{i=1,2} \sum_{a \in \{A,T,G,C\}} C_{ai}(z_1) C_{ai}(z_2)$$

  $C_{ai}(z)$: (a, i) のカウント

1 A	1 C	1 G	1 T	2 A	2 C	2 G	2 T
1	1	1	0	2	1	1	2

- Marginalized kernel

$$k(x_1, x_2) = \sum_{y_1} \sum_{y_2} \underline{p(y_1 | x_1) p(y_2 | x_2)} k_z(z_1, z_2)$$

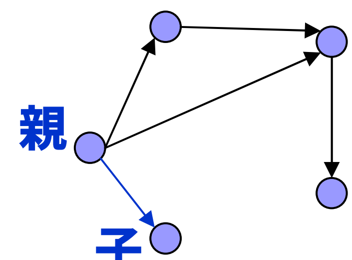
HMMから計算

1 A	1 C	1 G	1 T	2 A	2 C	2 G	2 T
1	1	1	0	1	2	0	1

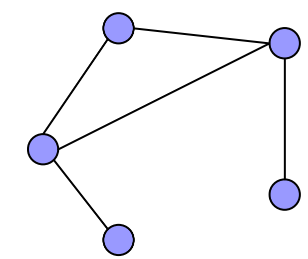
グラフとツリー

- グラフ
  - V： **ノード** (node, vertex) ・・・ **有限集合**
  - E： **エッジ** (edge) ・・・ V x V **の部分集合**
  - **有向グラフ**： E **の向きを考えたもの**

    (a, b) ∈ E **のとき**, a**から**b**へ矢印を描く**
    - **ノード** a **の親**： (b,a) ∈E **なる** b
    - **ノード** a **の子**： (a,b) ∈E **なる** b
  - **無向グラフ**： E **の向きを忘れたもの**

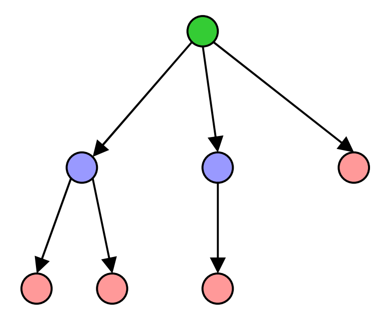


有向グラフ



無向グラフ

- **ツリー** (directed rooted tree)
  - **連結した有向グラフで, 親の無いルートノードが存在し, 他の各ノードは親を1個だけ持つもの**
  - **リーフ：ツリーの中で子の無いノード**



ツリー

ツリーカーネル

- ツリー全体の集合上に定義された正定値カーネル

$$\Phi: \text{ ツリー } T \mapsto \Phi(T) \in H$$ 特徴空間 (ベクトル空間)

- 代表的な例
  サブツリーの一致によりカーネルを定義する
  - All-subtrees kernel

$$k(T_1,T_2)=\sum_{S:\text{ ツリー}}\phi_S(T_1)\phi_S(T_2)$$

$$\phi_S(T)=\begin{cases}1 & T \text{ が } S \text{ をサブツリーとして含む} \\ 0 & T \text{ が } S \text{ をサブツリーとして含まない}\end{cases}$$

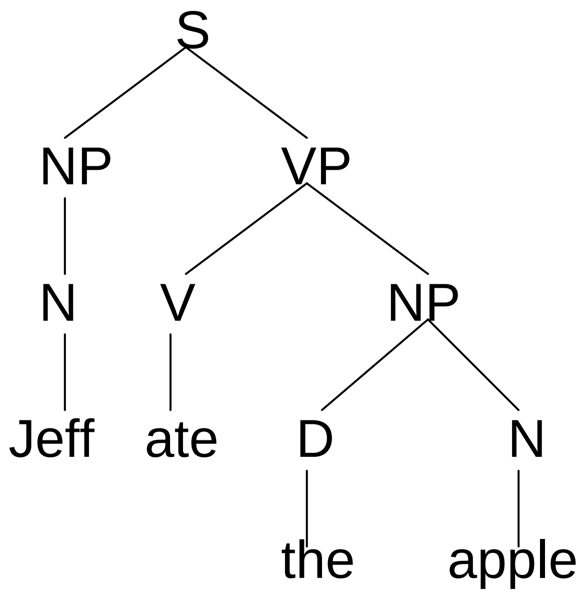
  - 再帰式で計算可能．　計算量 ＝ $O(|T_1|\,|T_2|)$

- 詳細は Collins & Duffy (2002, NIPS) などを参照

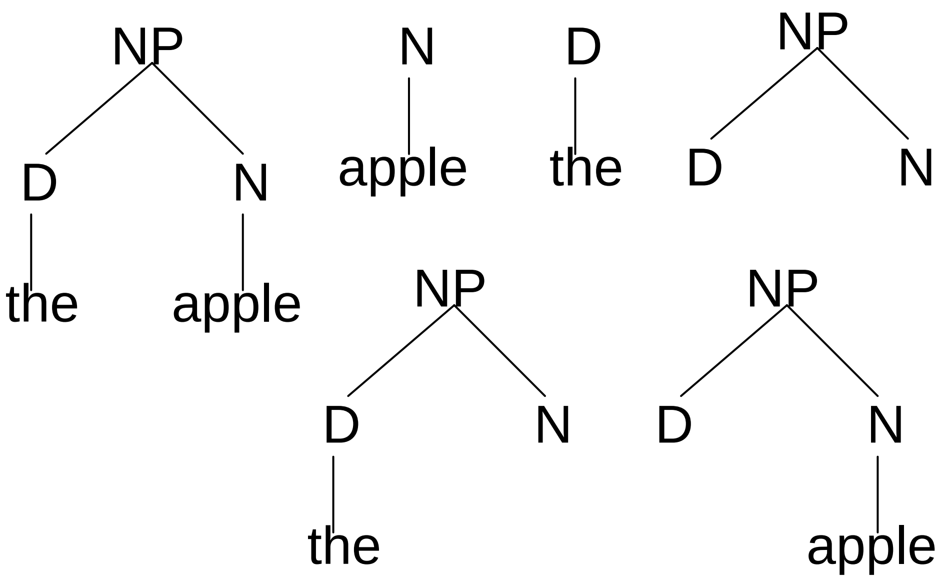
- 自然言語処理への応用

構文解析

Jeff ate the apple.



サブツリーの例



グラフカーネル

- グラフ上に定義された正定値カーネル
  - グラフとグラフの類似度を測る.
  - ラベル付グラフ

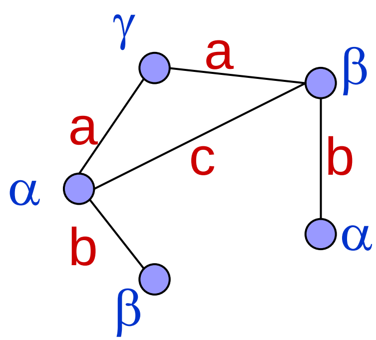
    ノードとエッジにラベルがついている.

    L: ラベルの集合 (有限集合)

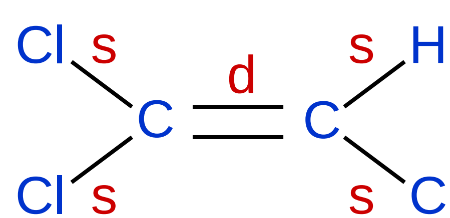
    ラベル付グラフ　G = (V, E, h)

    V: ノード, E：エッジ,

    h : V∪E → L　ラベル付けの写像



  - 応用
    - 化合物の毒性予測
    - 自然言語処理



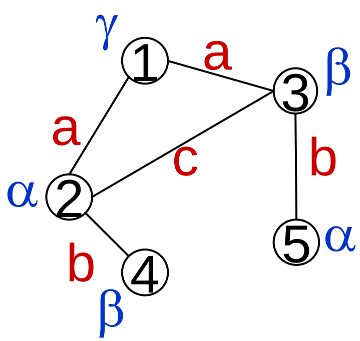
- Marginalized graph kernel

  - 系列のラベル

    s: $v_1\ v_2\ v_3\ v_5\ v_3 \cdots$

    → $H(s) = h(v_1)h(e_{12})h(v_2)h(e_{23})h(v_3)h(v_{35})h(v_5)\cdots$

    $= \gamma\ a\ \alpha\ c\ \beta\ b\ \alpha\ b\ \beta \cdots$



  - 系列の確率 – ランダムウォーク
    - ノード間の遷移確率

$$p(v_j \mid v_i) = \begin{cases} 1/(i\ \text{の隣接ノードの数}) & (i, j) \in E \\ 0 & (i, j) \notin E \end{cases}$$

    - 系列の確率

$$p(s) = p(v_1)\,p(v_2 \mid v_1)\,p(v_3 \mid v_2)\,p(v_5 \mid v_3)\,p(v_5 \mid v_3)\cdots$$

グラフ上のランダムウォークにより生じる系列の確率

- ラベル系列に対するカーネル

$$K_L : L^* \times L^*, \qquad K_L(H_1, H_2) = \begin{cases} 1 & (H_1 = H_2) \\ 0 & (H_1 \neq H_2) \end{cases}$$

- Marginalized graph kernel

$G_1 = (V_1, E_1, h_1), \quad G_2 = (V_2, E_2, h_2)$

$$K(G_1, G_2) = \sum_{\substack{s \in V_1^* \\ t \in V_2^*}} p_1(s) p_2(t) K_L(H_1(s), H_2(t))$$

H_1, H_2 : それぞれ h_1, h_2 から決まるラベル関数
V_1^*, V_2^* : それぞれ V_1, V_2 をアルファベットとする系列全体

  - ランダムウォークにおいて, 同じパスが生じる確率
  - Marginalized kernel のひとつとみなせる

- 詳しくは, Kashima et al. (2003), Mahé, et al. (2004)

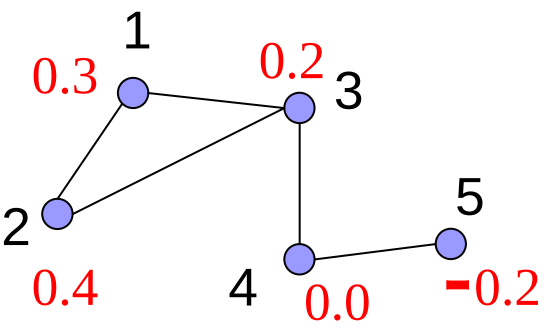
構造化データ上のカーネルの問題点

- 計算量
  - k(x,y) の計算にかかる時間
    O(|s| |t|)　でも，サイズが大きくなると困難
  - データ数
    グラム行列の計算は　(データ数)^2 のオーダー
  - SCOPデータベース：　配列の長さ～数百，　配列データの数～数千

ちょっと話を変えて. . .

グラフLaplacian と Diffusion Kernel

- 「ノード」間の近さを表す正定値カーネル



無向グラフ G = (V, E)

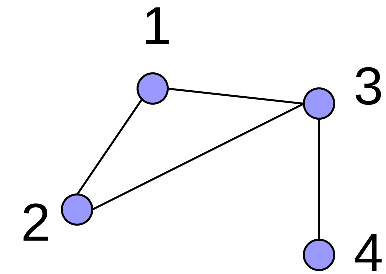
各ノードが値をもつ: $f_1, \ldots, f_n$

グラフ＝「隣接した2ノードは近い値をとりやすい」という情報の表現
⇒ 相関構造の導入

ノード集合 $\{1,\ldots, n\}$ 上のカーネル ⇒ n 次元ベクトルの再生核

注意: グラフとグラフの近さではない！！

グラフのLaplacian



無向グラフ $G = (V, E)$ **ノード数** n

隣接行列 A

$$A_{ij} = \begin{cases} 1 & (i, j) \in E \\ 0 & (i, j) \notin E \end{cases}$$

$$A = \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \end{array} \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array}$$

Laplacian L

$$L = D - A$$

ただし D **は対角行列で** $D_{ii} = d_i = \sum_{j=1}^{n} A_{ij}$

$$L = \begin{pmatrix} 2 & -1 & -1 & 0 \\ -1 & 2 & -1 & 0 \\ -1 & -1 & 3 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix}$$

Normalized Laplacian $\tilde{L}$

$$\tilde{L} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} A D^{-1/2}$$

- Laplacian の意味

V 上の関数　$f : V \to \mathbf{R}$　　（要はベクトル　($f(1), f(2), \ldots, f(n)$)　）

$$\boxed{(f, Lf) = \frac{1}{2} \sum_{i \sim j} \big(f(i) - f(j)\big)^2}$$

(f, Lf) 小　⇔　隣接ノードで近い値

特に L は(半)正定値行列　　・・・　正定値カーネルに使える

c.f)　$\mathbf{R}^n$ 上の Laplacian

$$\Delta f = \left(\frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \cdots + \frac{\partial^2}{\partial x_n^2} \right) f$$

これを離散点上で差分化したもの

Diffusion Kernel

- Diffusion Kernel **の定義**

$$\beta > 0$$

$$K = e^{\beta L} \quad = I + \beta L + \frac{1}{2!}\beta^2 L^2 + \frac{1}{3!}\beta^3 L^3 + \cdots$$

n x n **正定値行列**

隣接以外の近傍の効果が L **より強調されている**

  - **再生核ヒルベルト空間 ＝** n **次元ベクトル空間**

    内積 $\langle f, g \rangle_H = f^T e^{-\beta L} g = f^T K^{-1} g$

  - **拡散方程式との関連**

$$\frac{d}{d\beta} e^{\beta L} = L e^{\beta L}$$

c.f.) $\frac{\partial}{\partial t} H(x,t) = \Delta H(x,t)$

補足：有限集合上の再生核ヒルベルト空間

- $V = \{1,\ldots,n\}$ 上のRKHS ⇒ n 次元ベクトル空間
- 正定値カーネル

  $$K(i,j) \qquad (i,j=1,\ldots,n) \qquad \cdots \quad n \times n \text{ 行列}$$

- RKHSの内積

  $$f(x) = \sum_{i=1}^{n} u_i K(x,i) \qquad g(x) = \sum_{i=1}^{n} w_i K(x,i)$$

  ベクトル表示すると

  $$f = Ku \qquad g = Kw$$

  $$\langle f, g \rangle_H = \left\langle \sum_{i=1}^{n} u_i K(\cdot, i), \sum_{j=1}^{n} w_i K(\cdot, j), \right\rangle = u^T K w$$

  $$\Rightarrow \qquad \langle f, g \rangle_H = f K^{-1} g$$

- 相関構造

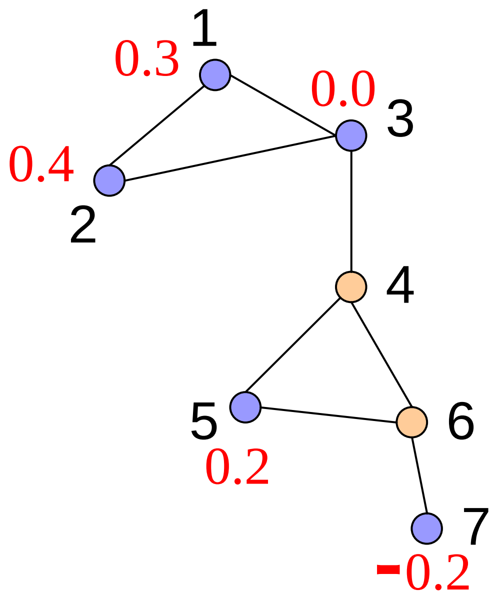
  K で「相関」を定める場合，　$\langle f, g \rangle_H = f K^{-1} g$ ＝ Mahalanobis距離

Laplacian, Diffusion Kernel の応用

- **グラフ上の関数の補間** (semi-supervised learning)
  - **グラフ構造は既知**
  - **ノードの一部の値が未知**

正則化問題

$$\min_{f \in H} \quad \sum_{i\,:\,\text{既知}} (y_i - f(i))^2 \quad + \lambda \| f \|_H^2$$



- **グラフ上の関数の平滑化 (スプライン)**
  - **すべての値が既知の場合**

$$\min_{f \in H} \quad \sum_{i=1}^{n} (y_i - f(i))^2 + \lambda \| f \|_H^2$$

セクション5のまとめ

- **構造化データのカーネル**
  - **非ベクトルデータのベクトル化**
  - **ストリング**
    - p-spectral kernel, all-subsequences kernel, gap-weighted kernel, …
  - **ツリー, グラフ**
  - **計算の効率化が重要**

- **グラフ**Laplacian **と** Diffusion Kernel
  - n **次元ベクトルデータに用いる**
  - **グラフによる離散点の相関構造の定義**

6. 独立性･条件付独立性とカーネル

- このセクションの目的
  - 独立性や条件付独立性が正定値カーネルで特徴付けられることを説明する
  - この特徴づけをもとにした，独立成分分析，次元削減の方法を紹介する
  - 「特徴空間での線形アルゴリズム」という今までの手法とは異なる

確率変数の独立性

- **独立性の定義**

X, Y： **確率変数**

P_{XY}： **同時確率,** P_X, P_Y： **周辺確率,**

X **と** Y **が独立** ⇔ $P_{XY}(A\times B) = P_X(A)P_Y(B)$

- **特性関数による特徴づけ**

X **と** Y **が独立** $\Leftrightarrow E_{XY}\left[e^{\sqrt{-1}\omega^T X} e^{\sqrt{-1}\eta^T Y}\right] = E_X\left[e^{\sqrt{-1}\omega^T X}\right] E_Y\left[e^{\sqrt{-1}\eta^T Y}\right]$

⇔ $e^{\sqrt{-1}\omega^T X}$ **と** $e^{\sqrt{-1}\eta^T Y}$ **の相関が0** (∀ω, η)

独立性 ⇔ 十分豊かな非線形相関が0

Fourier**カーネル** $e^{\sqrt{-1}\omega^T x}$ **と** $e^{\sqrt{-1}\eta^T y}$ **は非線形相関をはかるテスト関数**

再生核ヒルベルト空間と独立性

- **再生核ヒルベルト空間** (RKHS)**による独立性の特徴づけ**

X, Y： **それぞれ** Ω_X **と** Ω_Y **に値をとる確率変数**

H_X： Ω_X **上の**RKHS

H_Y： Ω_Y **上の**RKHS

X **と** Y **が独立**

$$\Leftrightarrow \quad E_{XY}\left[f(X)g(Y)\right] = E_X\left[f(X)\right]E_Y\left[g(Y)\right] \quad \text{for all } f \in H_X,\ g \in H_Y$$

$\Leftarrow$ **がいつ成り立つか？**　　(⇒ **は常に成立**)

H_X **と** H_Y **が ガウスカーネルの**RKHS**なら成立**　(Bach and Jordan 02)

∵) **十分豊かな非線形相関が表現できる**

カーネルICA

- 独立成分分析 (ICA)

  Z：m 次元確率変数 (ベクトル)

$$\begin{pmatrix} U^1 \\ \vdots \\ U^m \end{pmatrix} = A \begin{pmatrix} Z^1 \\ \vdots \\ Z^m \end{pmatrix}$$

  U の成分 $U_1, \ldots, U_m$ が独立になるように $m \times m$ 行列 A を見つける

  - さまざまなアルゴリズムが知られている
    KLダイバージェンスにもとづく方法, 高次モーメントにもとづく方法など
    (村田04参照)

- カーネルCCAによるICA

簡単のため2変数で説明

$U = (X,Y)^T = AZ$

H_X , H_Y ： ガウスカーネルのRKHS

X と Y が独立 $\Leftrightarrow \mathrm{Cov}_{XY}[f(X), g(Y)] = 0 \qquad (\forall f \in H_X, g \in H_Y)$

$$\Leftrightarrow \max_{\substack{f \in H_X \\ g \in H_Y}} \frac{\mathrm{Cov}_{XY}[f(X), g(Y)]}{\mathrm{Var}[f(X)]^{1/2}\mathrm{Var}[g(Y)]^{1/2}} = 0$$

データ $X_1, \ldots, X_N, \; Y_1, \ldots, Y_N$ を使うと

$$\max_{\substack{f \in H_X \\ g \in H_Y}} \frac{\frac{1}{N}\sum_i \langle f, k_X(\cdot, X_i)\rangle_{H_X} \langle g, k_Y(\cdot, Y_i)\rangle_{H_Y}}{\sqrt{\frac{1}{N}\sum_i \langle f, k_X(\cdot, X_i)\rangle_{H_X}^2}\sqrt{\frac{1}{N}\sum_i \langle g, k_Y(\cdot, Y_i)\rangle_{H_Y}^2}} = 0$$

カーネルCCAの正準相関

カーネルCCAによるICAアルゴリズム （2変数）

$$\tilde{\rho} = \max_{\substack{\alpha \in \mathbf{R}^N \\ \beta \in \mathbf{R}^N}} \frac{\alpha^T \tilde{K}_X \tilde{K}_Y \beta}{\sqrt{\alpha^T \left(\tilde{K}_X + \varepsilon I_N\right)^2 \alpha} \sqrt{\beta^T \left(\tilde{K}_Y + \varepsilon I_N\right)^2 \beta}}$$

$\Longrightarrow$ Aについて最小化

ただし　$A = (a_1, a_2)$

$$\tilde{K}_X = Q_N \left(k_X(a_1^T Z_i, a_1^T Z_j)\right) Q_N$$
$$\tilde{K}_Y = Q_N \left(k_Y(a_2^T Z_i, a_2^T Z_j)\right) Q_N$$

$\tilde{\rho}$：$\left(\tilde{K}_X + \varepsilon I_N\right)^{-1} \tilde{K}_X \tilde{K}_Y \left(\tilde{K}_Y + \varepsilon I_N\right)^{-1}$ の最大特異値

- Kernel generalized variance によるICA

$$I_N - \left(\tilde{K}_X + \varepsilon I_N\right)^{-1} \tilde{K}_X \tilde{K}_Y \left(\tilde{K}_Y + \varepsilon I_N\right)^{-1}$$ の特異値を大きくすればよい．

$$\hat{\Sigma}_{XY} = \tilde{K}_X \tilde{K}_Y \qquad \hat{\Sigma}_{XX} = (\tilde{K}_X + \varepsilon I_N)^2 \qquad \hat{\Sigma}_{YY} = (\tilde{K}_Y + \varepsilon I_N)^2$$ とおく

$$\begin{pmatrix} I_N & \hat{\Sigma}_{XX}^{-1/2}\hat{\Sigma}_{XY}\hat{\Sigma}_{YY}^{-1/2} \\ \hat{\Sigma}_{YY}^{-1/2}\hat{\Sigma}_{YX}\hat{\Sigma}_{XX}^{-1/2} & I_N \end{pmatrix} = \begin{pmatrix} \hat{\Sigma}_{XX}^{-1/2} & O \\ O & \hat{\Sigma}_{YY}^{-1/2} \end{pmatrix} \begin{pmatrix} \hat{\Sigma}_{XX} & \hat{\Sigma}_{XY} \\ \hat{\Sigma}_{YX} & \hat{\Sigma}_{YY} \end{pmatrix} \begin{pmatrix} \hat{\Sigma}_{XX}^{-1/2} & O \\ O & \hat{\Sigma}_{YY}^{-1/2} \end{pmatrix}$$

の固有値を大きくすればよい．

⇨

$$\boxed{\max_A \frac{\det\begin{pmatrix} \hat{\Sigma}_{XX} & \hat{\Sigma}_{XY} \\ \hat{\Sigma}_{YX} & \hat{\Sigma}_{YY} \end{pmatrix}}{\det \hat{\Sigma}_{XX} \det \hat{\Sigma}_{YY}}}$$

・・・ Kernel generalized variance

- 非線形最適化による A の最適化
- ガウス分布の相互情報量の一般化

回帰問題における次元削減

- 回帰問題における有効な部分空間
  - 回帰問題 ・・・ Y を X で説明する

    $p(Y \mid X)$ の推定

  - 次元削減

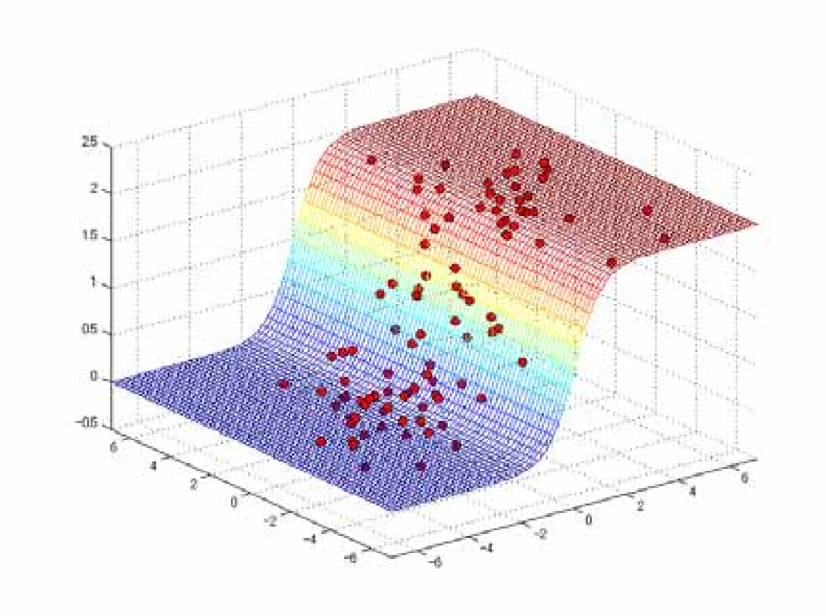
    $X : m$ 次元ベクトル
    $B = (b_1, \ldots, b_d)$ $m \times d$ 行列

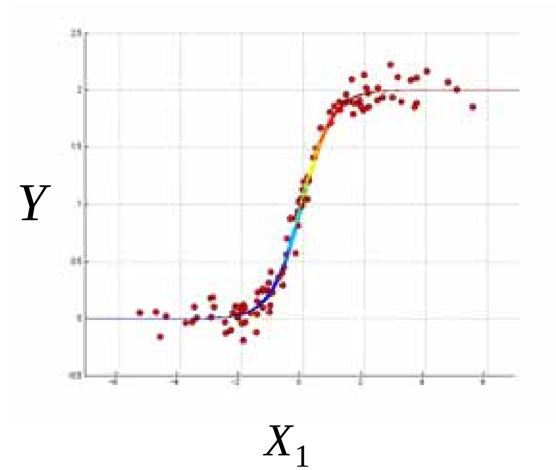
    $$p(Y \mid X) = p(Y \mid B^T X)$$

    となる B を探す.

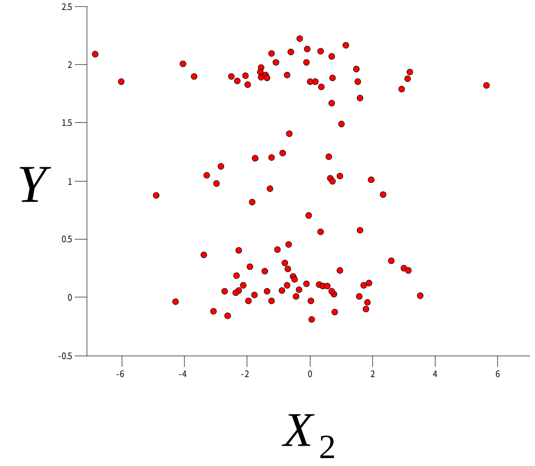
$B^T X = \underline{(b_1^T X, .., b_d^T X)}$ は，Y を説明する目的では, X と同じ情報を持つ

有効部分空間 (特徴ベクトル)





$$Y = \frac{2}{1+\exp(-2X_1)} + N(0; 0.1^2)$$



次元削減と条件付独立性

X の分解　$(U, V) = (B^T X, C^T X)$　　$(B, C) \in O(m)$　直交行列

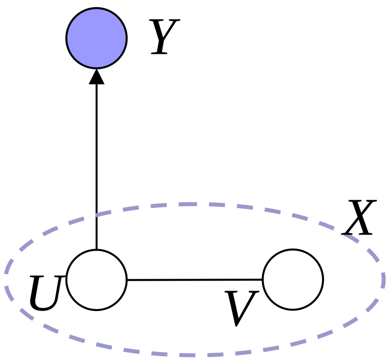
U：有効なベクトルの候補,　V：それに直交する方向

B が有効な部分空間を与える

$\Leftrightarrow \quad p_{Y|X}(y \mid x) = p_{Y|U}(y \mid B^T x)$

$\Leftrightarrow \quad p_{Y|U,V}(y \mid u, v) = p_{Y|U}(y \mid u)$　for all y, u, v

$\Leftrightarrow$　Y と V は U のもと**条件付独立**　$Y \perp V \mid U$



条件付独立性と条件付分散

- **条件付分散**
  - X, Y: **ガウスの場合**

    $$\mathrm{Var}[Y \mid X] = V_{YY} - V_{YX} V_{XX}^{-1} V_{XY} = (Y\text{ を }X\text{ で線形回帰した残差})$$

    $\mathrm{Var}[Y \mid X]$ **が小さいほど** X **は** Y **の情報を多く含んでいる**

  - X, Y: **一般の場合**
    - $X = (U, V)$ **と分解すると**

      $$E_X[\mathrm{Var}[f(Y) \mid X]] \le E_U[\mathrm{Var}[f(Y) \mid U]] \qquad (\forall f)$$

      Y **に関する情報が増えることはない**

    - $Y \perp V \mid U$ **ならば**

      $$E_X[\mathrm{Var}[f(Y) \mid X]] = E_U[\mathrm{Var}[f(Y) \mid U]] \qquad (\forall f)$$

      Y **に関する情報は落ちない**

RKHSと条件付独立性

Q: 逆に

$$E_X[\mathrm{Var}[f(Y)\,|\,X]] \;=\; E_U[\mathrm{Var}[f(Y)\,|\,U]]$$

ならば $Y \perp V \,|\, U$ か？

A: $f(Y)$ が Y に関する情報のバリエーションを十分表現できれば「Yes」

H_Y： ガウスカーネルのRKHS

$$Y \perp V \,|\, U$$

$$\Leftrightarrow \quad E_X[\mathrm{Var}[f(Y)\,|\,X]] = E_U[\mathrm{Var}[f(Y)\,|\,U]] \qquad \forall f \in H_Y$$

証明は Fukumizu et al. 04 参照

条件付分散の推定

有限個のデータ $X_1, \ldots, X_N,\ Y_1, \ldots, Y_N$ から条件付分散を推定

$f = \sum_{i=1}^{N} \alpha_i k_Y(\cdot, Y_i)$ の形に限ると

$$E_X[\mathrm{Var}[f(Y) \mid X]] \approx \alpha^T \underline{\left(\hat{\Sigma}_{YY} - \hat{\Sigma}_{YX} \hat{\Sigma}_{XX}^{-1} \hat{\Sigma}_{XY}\right)} \alpha$$

ここで $\hat{\Sigma}_{XY} = \tilde{K}_X \tilde{K}_Y$ $\hat{\Sigma}_{XX} = (\tilde{K}_X + \varepsilon I_N)^2$ $\hat{\Sigma}_{YY} = (\tilde{K}_Y + \varepsilon I_N)^2$

c.f.) ガウス分布の条件付分散： $V_{YY} - V_{YX} V_{XX}^{-1} V_{XY}$

カーネル次元削減法

- **条件付分散の最小化**
  - $E_U[\mathrm{Var}[f(Y)|U]]$ **を小さくする** $U = B^T X$ **を探したい**

⇨ $\hat{\Sigma}_{YY} - \hat{\Sigma}_{YU}\hat{\Sigma}_{UU}^{-1}\hat{\Sigma}_{UY}$ **最小化**

⇨ $\hat{\Sigma}_{YU}\hat{\Sigma}_{UU}^{-1}\hat{\Sigma}_{UY}$ **最大化** （$\because$ $\hat{\Sigma}_{YY}$ **は一定**）

➡ Kernel ICA **の時と同様に**

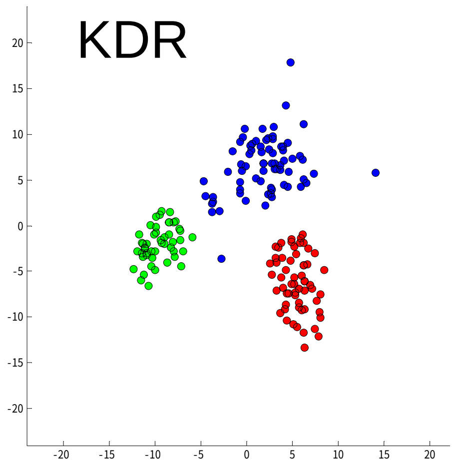
$$\boxed{\min_B \frac{\det\begin{pmatrix} \hat{\Sigma}_{UU} & \hat{\Sigma}_{UY} \\ \hat{\Sigma}_{YU} & \hat{\Sigma}_{YY} \end{pmatrix}}{\det\hat{\Sigma}_{UU}\,\det\hat{\Sigma}_{YY}}}$$

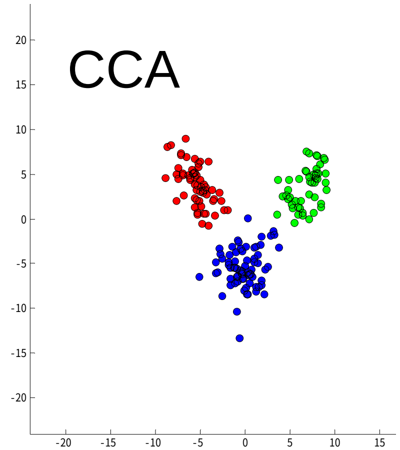
再び Kernel generalized variance

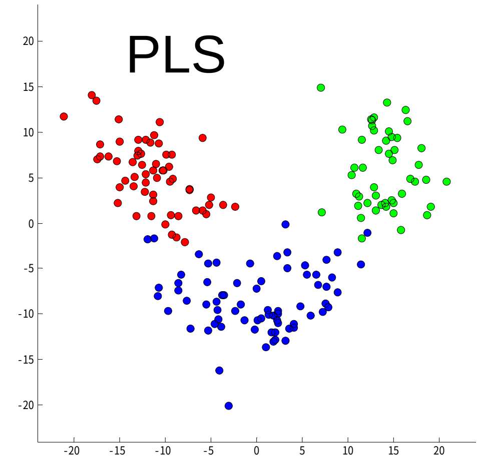
Kernel Dimensionality Reduction (KDR)

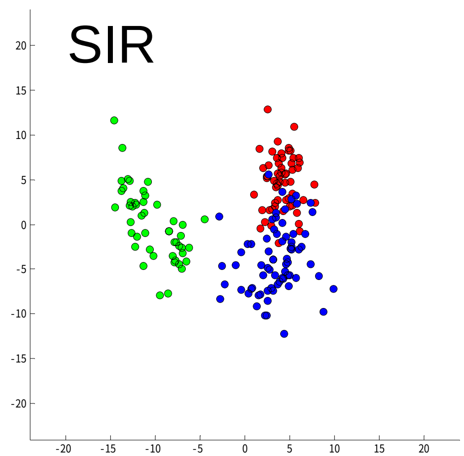
Wine data　(他の次元削減法との比較)

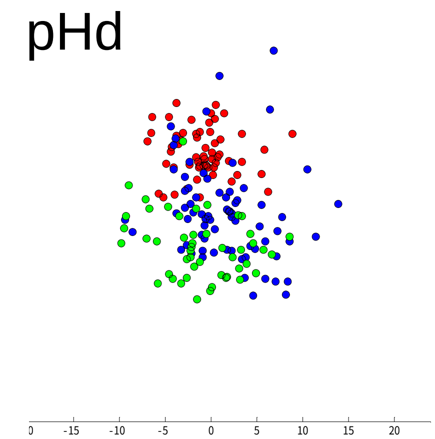
- Data
  13 dim. 178 data.
  3 classes
  2 dim. projection











カーネル次元削減法の特徴

- どんなデータでも扱える
  - 回帰問題の次元削減としては最も一般的
    $p(Y|X)$ のモデル (線形など)を使わない.
  - X, Y の分布に条件がいらない. Y が離散値や高次元でもOK.
    従来法 (SIR, pHd, CCA, PLS, etc) ではさまざまな制約
- 計算量の問題 (カーネルICA, KDRに共通)
  - N x N 行列を用いた演算.
    → Incomplete Cholesky decomposition の利用
  - 非線形最適化に伴う局所解／計算時間の問題

セクション6のまとめ

- RKHSによる独立性・条件付独立性の特徴づけ
  - RKHSは非線形性な関係をはかるテスト関数として有効
    - L^2などより「狭い」空間．関数の連続性，微分可能性
    - 各点での値が意味を持つ

- カーネルICA
  - 従来のICAと異なる理論にもとづくアルゴリズム

- カーネル次元削減法 (KDR)
  - 回帰問題に対する，もっとも一般的な次元削減法

7. まとめ

- **カーネル法**
  - **正定値カーネルによる特徴写像で, 線形アルゴリズムを非線形化**
  - **非ベクトルデータ／構造化データの数量化**
  - **再生核ヒルベルト空間＝非線形性をはかるテスト関数の空間**
    独立性・条件付独立性の特徴づけ
  - **現在も発展途上**

- **講義で扱わなかったこと**
  - SVM**の詳細**
    - **最適化に関わる点**, Vapnik**流の誤差の上界評価**
  - **再生核ヒルベルト空間と確率過程の関係** (Parzen 61)

補足：RKHSと確率過程

Ω ：集合

Ω 上の正定値カーネル $k(\cdot, t)$	1対1 $\Longleftrightarrow$	相関 $k(s,t)$ を持つ (2乗可積分な) 確率過程 X_t
⋮		⋮
RKHS	同型 $\cong$	X_t の張る部分空間 ⊆ L^2空間
$k(\cdot, t)$	$\leftrightarrow$	X_t

$$k(t, s) = E[X_t X_s]$$

カーネルには，背後に確率過程がある

参考となる資料

- **ホームページとソフトウエア**
  - **カーネル関連ポータルサイト** http://www.kernel-machines.org
  - http://www.euro-kermit.org Kernel Methods for Image and Text（**欧州のプロジェクト**）
  - SVM Java applet: http://svm.dcs.rhbnc.ac.uk/
  - SVM C++**プログラム**（Web**上で**Job**も受け付ける**）GIST
    http://svm.sdsc.edu/cgi-bin/nph-SVMsubmit.cgi
  - UCI Machine Learning Repository
    http://www.ics.uci.edu/~mlearn/MLRepository.html

- **論文誌, 国際会議**
  - Journal of Machine Learning Research
    http://jmlr.csail.mit.edu/ （**論文は全て無料でダウンロード可**）
  - Neural Computation
  - Neural Information Processing Systems (NIPS, Conference)
  - International Conference on Machine Learning (ICML)

- **全般的な文献**
  - Schölkopf, B. and A. Smola. *Learning with Kernels*. MIT Press. 2002.
  - **津田宏治.「カーネル法の理論と実際」** in **統計科学のフロンティア6：パターン認識と学習の統計学（岩波書店）** 2003.
  - Müller K.-R., S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. (2001) An introduction to kernel-based learning algorithms. *IEEE Trans. Neural Networks, 12(2),* pp.181-201.
  - John Shawe-Taylor & Nelo Cristianini. *Kernel Methods for Pattern Analysis.* Cambridge Univ. Press. 2004.

- **個別の話題に関する文献（特にセクション5，6）**
  - SVM **関連**
    - Schölkopf, B., C. Burges, and A. Smola (eds). *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1999.
    - Smola, A., P. Bartlett, B. Schölkopf, and D. Schuurmans (eds). *Advances in Large Margin Classifiers*. MIT Press, 2000.
    - Vladimir Vapnik. *Statistical Learning Theory*. Wiley, NY, 1998.
  - Kernel PCA
    - Schölkopf, B., A. Smola, K.-R. Müller. (1998) Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation 10*, 1299–1319.

- Kernel CCA
  - Akaho, S. (2001) A kernel method for canonical correlation analysis. *International Meeting on Psychometric Society (IMPS2001).*
  - See also Bach and Jordan (JMLR, 2002) in Kernel ICA.
- String kernel
  - Lodhi, H., C. Saunders, J. Shawe-Taylor, N. Cristianini, C. Watkins. (2002) Text Classification using String Kernels. *J. Machine Learning Research*, 2 (Feb): 419-444.
  - Leslie, C., E. Eskin, A. Cohen, J. Weston and W. S. Noble. (2003) Mismatch string kernels for SVM protein classification. *Advances in Neural Information Processing Systems 15*, pp. 1441-1448.
  - Rousu, J., and J. Shawe-Taylor. (2004) Efficient computation of gap-weighted string kernels on large alphabets. *Proc. PASCAL Workshop Learning Methods for Text Understanding and Mining*.
- Suffix Tree
  - Dan Gusfield. *Algorithms on Strings, Trees, and Sequences*. Cambridge Univ. Press. 1997.
- Fisher Kernel
  - Jaakkola, T.S. and D. Haussler. (1999) Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems 11*. pp.487-493.
- Tree kernel
  - Collins, M. & N. Duffy. (2002) Convolution Kernels for Natural Language. *Advances in Neural Information Processing Systems 14.*
- Marginalized kernel
  - Tsuda, K., T. Kin, and K. Asai. (2002) Marginalized kernels for biological sequences. *Bioinformatics, 18.* S268-S275.

- Graph kernel
  - Kashima, H., K. Tsuda and A. Inokuchi. (2003) Marginalized Kernels Between Labeled Graphs. *Proc. 20th Intern. Conf. Machine Learning (ICML2003).*
  - Mahé, P., N. Ueda, T. Akutsu, J.-L. Perret and J.-P. Vert. (2004) Extensions of marginalized graph kernels. *Proc. 21th Intern. Conf. Machine Learning (ICML 2004),* p.552-559.
- Diffusion kernel
  - Kondor, R.I., and J.D. Lafferty (2002) Diffusion Kernels on Graphs and Other Discrete Input Spaces. *Proc. 19th Intern. Conf. Machine Learning (ICML2002)*: 315-322.
  - Lafferty, J.D. and G. Lebanon. (2003) Information Diffusion Kernels. *Advances in Neural Information Processing Systems 15,* 375-382.
- Kernel ICA / Kernel Dimensionality Reduction
  - Bach, F.R. and M.I. Jordan. Kernel independent component analysis. *J. Machine Learning Research, 3*, 1-48, 2002.
  - Fukumizu, K., F.R. Bach, and M.I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces, *J. Machine Learning Research*, 5, 73-99, 2004.
- ICA
  - **村田昇.「入門独立成分分析」東京電機大学出版局.** 2004
- RKHS**と確率過程**
  - Parzen, E. (1961) An Approach to Times Series Analysis. *The Annals of Statistics, 32.* pp.951-989.

- **その他**
  - Computational biology **へのカーネル法の応用**
    - Schlkopf，B., K. Tsuda, J-P. Vert (Editor) *Kernel Methods in Computational Biology*. Bradford Books. 2004.
  - **数学理論に関する本**
    - Berlinet, A. and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2003.
    - Berg, C., J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups. Theory of Positive Definite and Related Functions* (Graduate Texts in Mathematics Vol. 100). Springer 1984.