3. さまざまなカーネル法

正定値カーネルによるデータ解析
－ カーネル法の基礎と展開 －

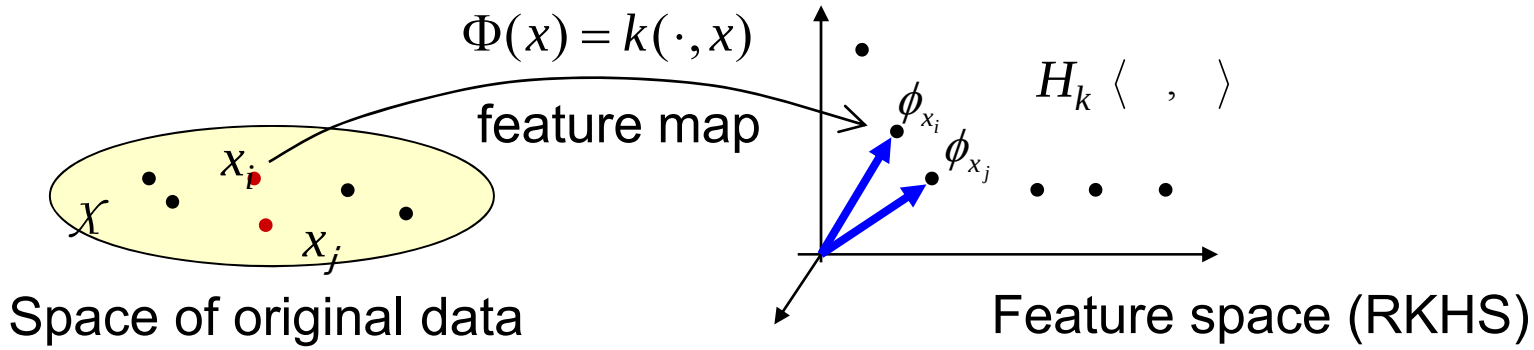
福水健次
統計数理研究所／総合研究大学院大学

統計数理研究所　公開講座
2011年1月13,14日

概要

- Kernel PCAの応用
- Kernel CCA（カーネル正準相関分析）
- サポートベクターマシンの基礎
- カーネル法の方法論
  - 共通する方法
  - Representer定理
- その他の話題
  - カーネルの選択
  - 低ランク近似

カーネル法のアイデア



– データの非線形特徴・高次モーメントを表現するため，特徴写像によってデータを特徴空間（RKHS）に写像し，特徴空間で線形のデータ解析アルゴリズムを適用する.

– 特徴写像: $\Phi : \Omega \to H_k, \quad \Phi(x) = k(\cdot, x)$

– 高次元（無限次元）特徴空間で，内積の簡単な計算が可能.

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

特徴写像

- 正定値カーネルによる高次モーメントの抽出

  Example

  – Polynomial kernel: $k(y, x) = (yx + 1)^d$ on $\mathbf{R}$

  – 特徴写像

$$\Phi(X) = k(u, X) = (uX + 1)^d$$
$$= X^d u^d + a_{d-1} X^{d-1} u^{d-1} + a_{d-2} X^{d-2} u^{d-2} + \cdots + a_1 Xu + 1$$

  – 基底$\{1, u, u^2, ..., u^d\}$を用いると, ベクトル成分表示は

$$\Phi(X) \sim (1, a_1 X, ..., a_{d-1} X^{d-1}, X^d)$$

  高次統計量 $X, X^2, ..., X^d$ を含む.

  – ガウスカーネルなど, 他の非線形カーネルでも事情は同様.

  – ただし, カーネル法では成分表示／基底展開せずに内積計算が可能.

概要

- Kernel PCAの応用
- Kernel CCA（カーネル正準相関分析）
- サポートベクターマシンの基礎
- カーネル法の方法論
  - 共通する方法
  - Representer定理
- その他の話題
  - カーネルの選択
  - 低ランク近似

カーネルPCA（復習）

- 特徴写像　$X_1,\ldots,X_N \quad \rightarrow \quad \Phi(X_1),\ldots,\Phi(X_N)$
- 目的関数：　特徴空間でのPCA

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

$$\text{ただし} \qquad \tilde{\Phi}(X_i) = \Phi(X_i) - \frac{1}{N}\sum_{j=1}^{N}\Phi(X_j)$$

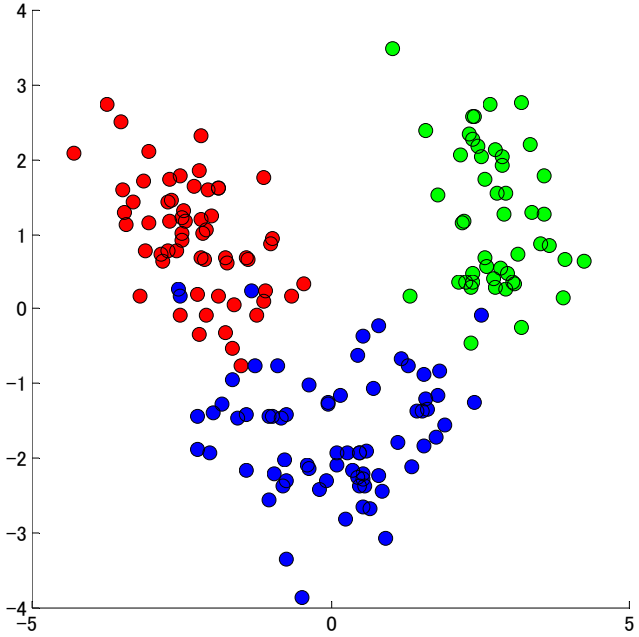
- （中心化された）データの線形和で解が求まる　$f = \sum_{i=1}^{N} c_i \tilde{\Phi}(X_i)$
- 固有値問題：　中心化グラム行列の固有値分解

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

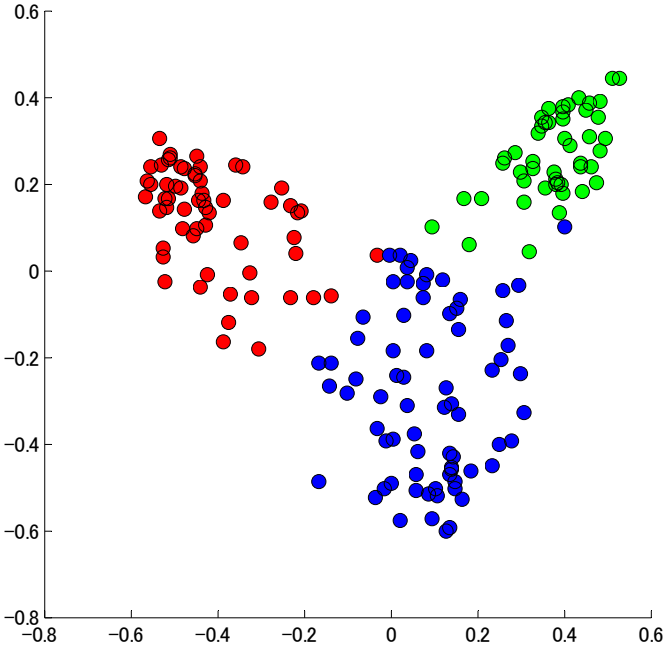
- 解
  - 第p主軸　$f^p = \sum_{i=1}^{N} c_p^i \tilde{\Phi}(X_i), \quad c_p = \frac{1}{\sqrt{\lambda_p}} u_p$
  - データ X_i の第p主成分　$= \left\langle f^p, \tilde{\Phi}(X_i)\right\rangle = \sqrt{\lambda_p} u_p^T X_i$

カーネルPCAの応用

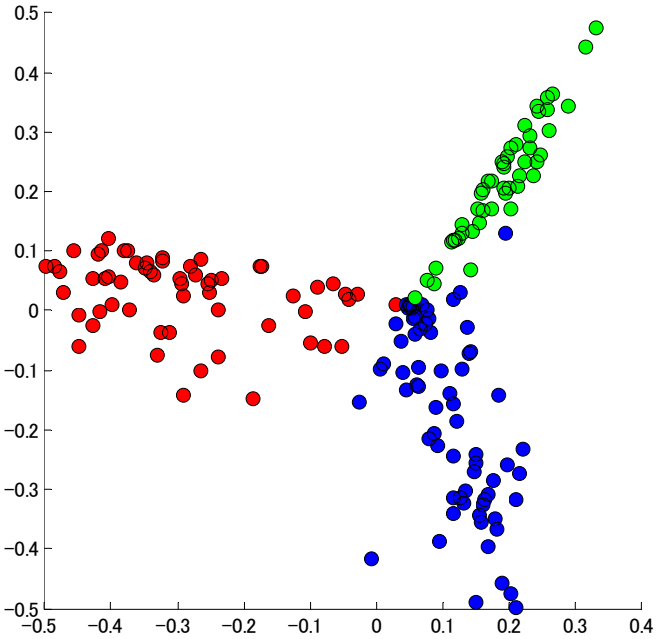
- ‘Wine’ データ（UCI Machine Learning Repository）
  - 13次元（連続値），178データ，3種類（産地）のワインの化学成分に関する属性データ
  - 2つの主成分を取った（3クラスの色は参考に付けたもの，Kernel PCAには用いていない）



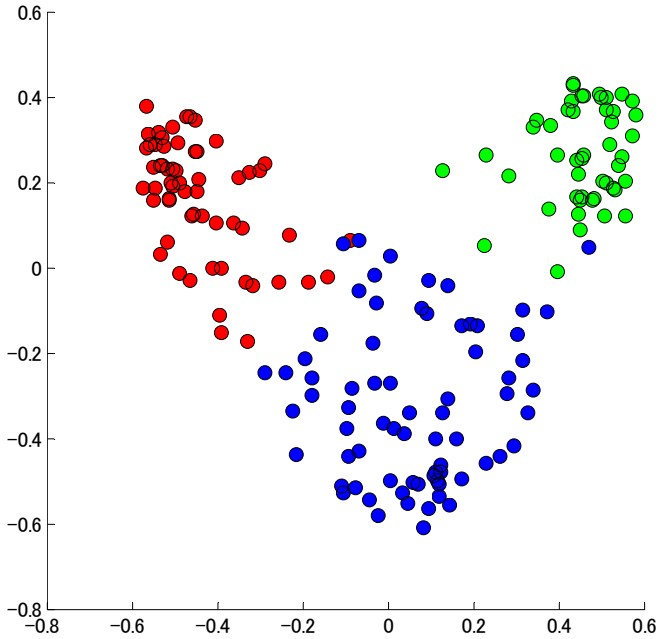
PCA（線形）



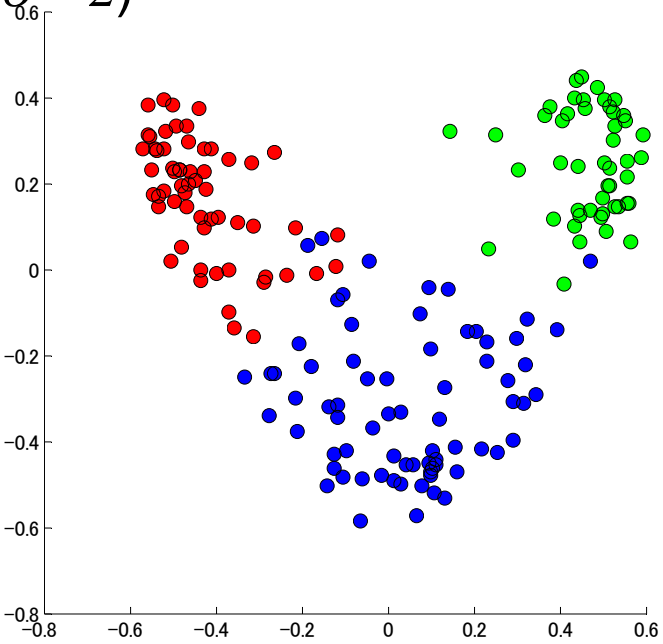
KPCA（Gauss，$\sigma = 3$）



KPCA（Gauss，$\sigma = 2$）



KPCA（Gauss，$\sigma = 4$）



KPCA（Gauss，$\sigma = 5$）

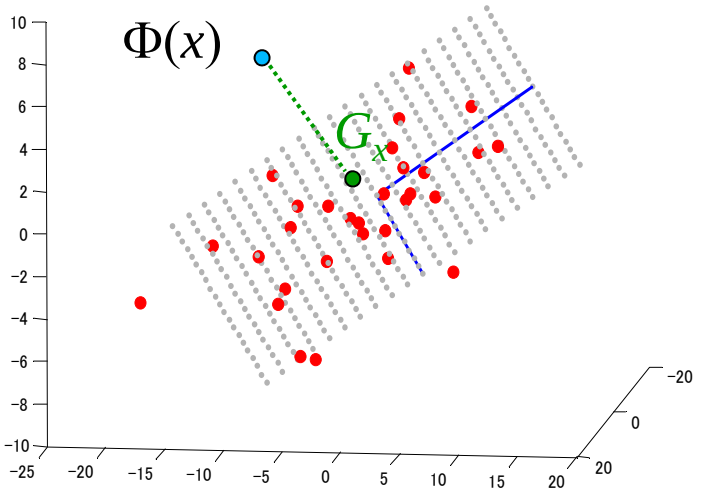
カーネルPCAの雑音除去への応用

- （カーネル）PCAによる雑音除去

  高次元の空間のなかで、d 個の主成分の軸 $F_1, F_2, ..., F_d$
  が張る d 次元部分空間へ、データ $\Phi(x)$ を射影した点を G_x とおく。

  雑音除去された特徴ベクトル

  G_x に最も近い埋め込み点を探す
  （pre-imageの問題）



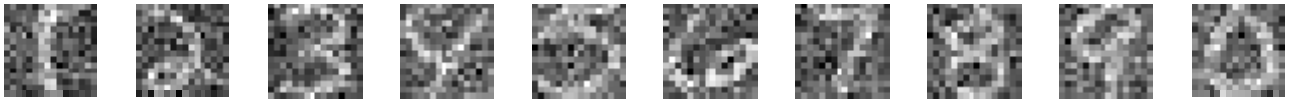
$$\hat{x} = \arg\min_{x'} \left\| \Phi(x') - G_x \right\|^2$$

カーネル $k(x_1, x_2)$ を使って表せる

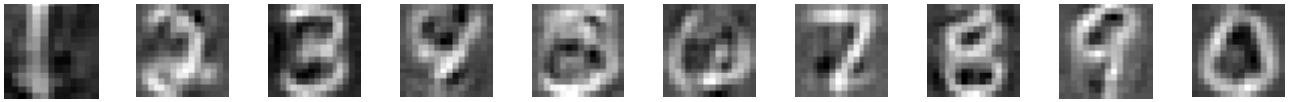
- USPS　手書き数字データベース

16x16画素(256次元) 7291データ

元の画像 (データとしては使用していない)



ノイズつきの画像



ノイズ除去された画像（linear PCA）

ノイズ除去された画像（kernel PCA, Gaussカーネル）

Matlab stprtool (by V. Franc) により作成

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． $\max_f \mathrm{Var}[\, f(X) \,]$
- 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， 目的に依存
- 前処理として使われることが多い
  後の処理の結果を改良するための非線形特徴抽出

  例えば、

  カーネルPCA ＋ 識別機（SVM） によるクラス識別問題
  → 最終的な識別の正答率を向上させるカーネルがよい
  （Cross-Validationの適用可能)

概要

- Kernel PCAの応用
- Kernel CCA（カーネル正準相関分析）
- サポートベクターマシンの基礎
- カーネル法の方法論
  - 共通する方法
  - Representer定理
- その他の話題
  - カーネルの選択
  - 低ランク近似

正準相関分析

- 正準相関分析（Canonical Correlation Analysis, CCA）

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 , Y を b 方向に射影したときに相関を最大 にする(a,b) を求める

正準相関

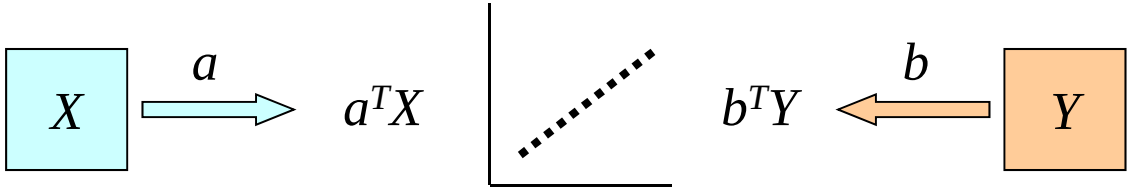
$$\rho = \max_{\substack{a\in\mathbf{R}^m \\ b\in\mathbf{R}^n}} \mathrm{Corr}[a^T X, b^T Y] = \max_{\substack{a\in\mathbf{R}^m \\ b\in\mathbf{R}^n}} \frac{\frac{1}{N}\sum_i \left(a^T \tilde{X}_i\right)\left(b^T \tilde{Y}_i\right)}{\sqrt{\frac{1}{N}\sum_i \left(a^T \tilde{X}_i\right)^2}\sqrt{\frac{1}{N}\sum_i \left(b^T \tilde{Y}_i\right)^2}}$$

$$= \max_{\substack{a\in\mathbf{R}^m \\ b\in\mathbf{R}^n}} \frac{a^T V_{XY} b}{\sqrt{a^T V_{XX} a}\sqrt{b^T V_{YY} b}}$$

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

$\tilde{X}_i = X_i - \frac{1}{N}\sum_j X_j$

など



$$\rho = \max\ a^T V_{XY} b \qquad \text{subj. to} \quad a^T V_{XX} a = b^T V_{YY} b = 1$$

⇨

Lagrange乗数法

$$\max \quad a^T V_{XY} b + \frac{\mu}{2}(a^T V_{XX} a - 1) + \frac{\nu}{2}(b^T V_{YY} b - 1)$$

⇨

一般化固有値問題

$$\begin{pmatrix} O & V_{XY} \\ V_{YX} & O \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \rho \begin{pmatrix} V_{XX} & O \\ O & V_{YY} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$

$[\ \mu = \nu = -\rho\ ]$

最大固有値，固有ベクトルを求めればよい

カーネル正準相関分析

- カーネルCCA（赤穂, 2000, Melzer et al. 2001）
  - Data: $(X_1, Y_1), ..., (X_N, Y_N)$.

    X, Y: arbitrary variables taking values in Ω_X and Ω_Y (resp.).
  - 特徴写像

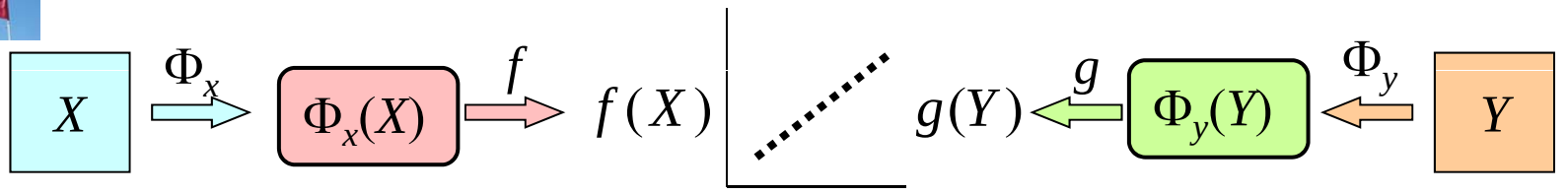
$$\Phi_X : \Omega_X \to H_X, \quad \Phi_X(X_i) = k_X(\cdot, X_i)$$
$$\Phi_Y : \Omega_Y \to H_Y, \quad \Phi_Y(Y_i) = k_Y(\cdot, Y_i)$$

  - Apply CCA on $\Phi_X(X_1),...,\Phi_X(X_N)$ and $\Phi_Y(Y_1),...,\Phi_Y(Y_N)$.

$$\max_{f,g} \mathrm{Corr}\left[\langle f, \Phi_X(X_i)\rangle, \langle \Phi_Y(Y_i), g\rangle\right] = \max_{f,g} \mathrm{Corr}\left[f(X), g(Y)\right]$$

$$= \max \frac{\sum_{i=1}^{N} \left\langle f, \tilde{\Phi}_X(X_i)\right\rangle \left\langle \tilde{\Phi}_Y(Y_i), g\right\rangle}{\sqrt{\sum_{i=1}^{N} \left\langle f, \tilde{\Phi}_X(X_i)\right\rangle^2} \sqrt{\sum_{i=1}^{N} \left\langle \tilde{\Phi}_Y(Y_i), g\right\rangle^2}}$$

‘*flag*’
‘*sky*’,



カーネルPCA同様 $f = \sum_{i=1}^{N} \alpha_i \tilde{\Phi}_X(X_i),\ g = \sum_{i=1}^{N} \beta_i \tilde{\Phi}_Y(Y_i)$ としてよい.

⇒ (カーネルトリック)

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

$\tilde{K}_X, \tilde{K}_Y$: 中心化グラム行列

実は ill-posed. $R\left(\tilde{K}_X\right) \cap R\left(\tilde{K}_Y\right) \neq \{0\}$ であれば*, 常に正準相関 = 1.

⇒ 正則化

$$\max_{f,g} \frac{\frac{1}{N}\sum_{i=1}^{N} \left\langle f, \tilde{\Phi}_X(X_i) \right\rangle \left\langle \tilde{\Phi}_Y(Y_i), g \right\rangle}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} \left\langle f, \tilde{\Phi}_X(X_i) \right\rangle^2 + \varepsilon_N \| f \|_{H_X}^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N} \left\langle \tilde{\Phi}_Y(Y_i), g \right\rangle^2 + \varepsilon_N \| g \|_{H_Y}^2}}$$

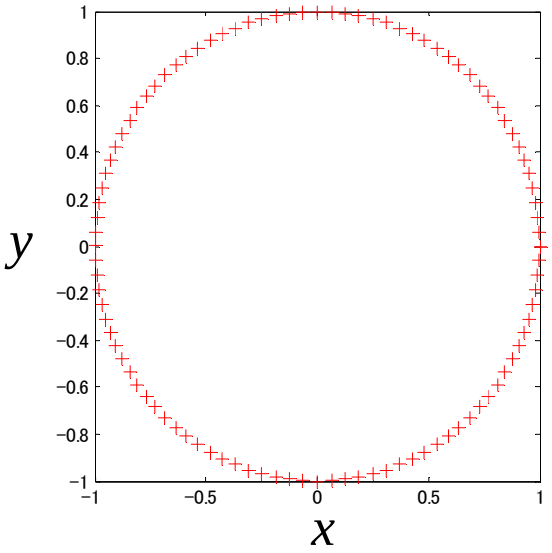
一般化固有値問題として解ける

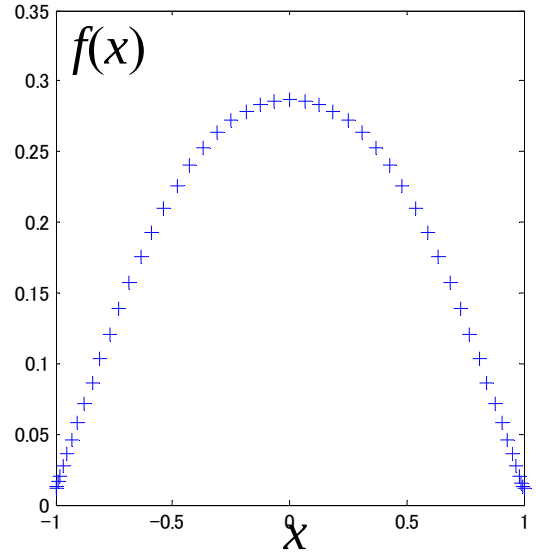
$$\begin{pmatrix} O & \tilde{K}_X \tilde{K}_Y \\ \tilde{K}_Y \tilde{K}_X & O \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \rho \begin{pmatrix} (\tilde{K}_X + N\varepsilon_N I_N)^2 & O \\ O & (\tilde{K}_Y + N\varepsilon_N I_N)^2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

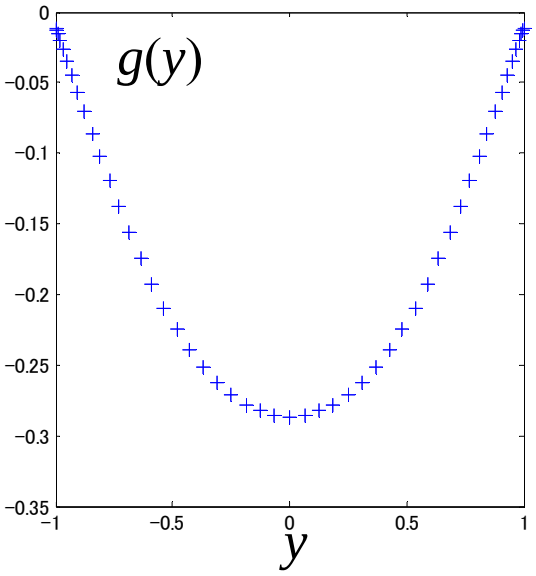
* $R(T)$: 行列Tの像空間

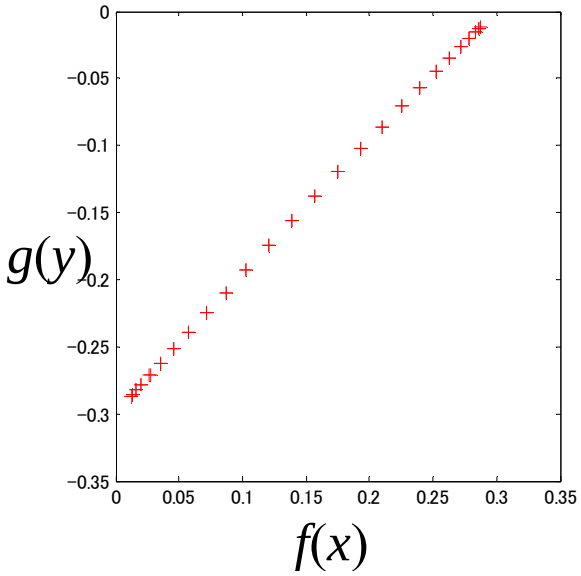
- カーネルCCAの実験例

ガウスカーネル









Kernel CCAの性質

- 複数の固有ベクトルも考えられる(第2，第3，．．．固有ベクトル）も考えられる.（もとのデータが1次元でも！）

- 結果はカーネルや正則化係数ε_Nに依存する.

- 正準相関の値によって相関の大小を解釈するのは難しい（正則化のために正規化されない）

- カーネル／正則化係数の選択:
  - Cross-validationが用いられる場合もある.
  - 他に提案されている方法もある(Hardoon et al 2004. See later.).

- ε_Nが十分ゆっくり0に近づく場合，一致性があることが知られている. (Fukumizu et al 2007).

Kernel CCAの画像検索への応用

Idea: d 個の固有ベクトル $f_1, ..., f_d$ と $g_1, ..., g_d$ を X と Y の依存性が最も強く表れている特徴空間とみなす.

– X: image, Y: text (extracted from the same webpage).

X_i 

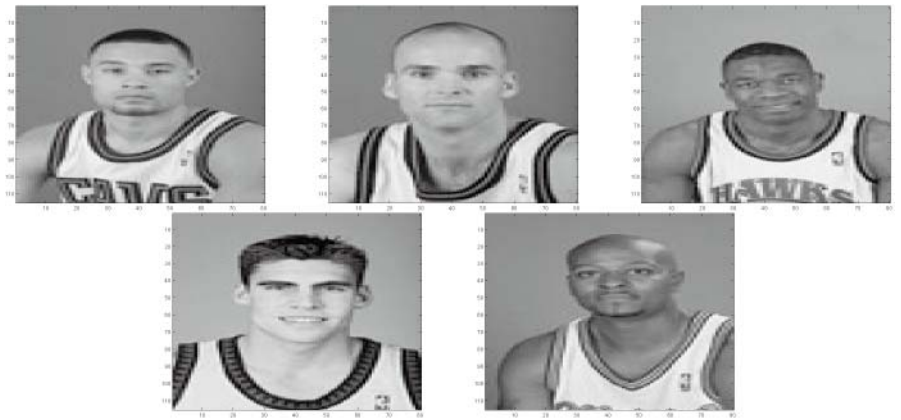
Y_i: 'sky', 'Phoenix', 'harbor', ...

– 方法

$$X_i \xrightarrow{\Phi_x} \Phi_x(X_i) \rightarrow \begin{pmatrix} \langle f_1, \Phi_x(X_i) \rangle \\ \vdots \\ \langle f_d, \Phi_x(X_i) \rangle \end{pmatrix} \bullet \begin{pmatrix} \langle g_1, \Phi_y(Y) \rangle \\ \vdots \\ \langle g_d, \Phi_y(Y) \rangle \end{pmatrix} \xleftarrow{g} \Phi_y(Y) \xleftarrow{\Phi_y} Y$$

'*flag*'
'*sky*',

– 例



テキストからの画像検索例: ”height: 6-11 weight: 235 lbs position: forward
born: september 18, 1968, split, croatia college: none”

Hardoon et al. *Neural Computation* (2004).

– テキスト → “bag-of-words” kernel (単語の頻度分布) を用いる.

– 正則化係数 ε_N の決め方:

$$\varepsilon = \arg\max_{\varepsilon} \| \lambda(\varepsilon) - \lambda_R(\varepsilon) \|$$

$\lambda(\varepsilon)$：KCCAの固有値分布,
$\lambda_R(\varepsilon)$：XとYをランダム化したときのKCCAの固有値分布

概要

- Kernel PCAの応用
- Kernel CCA（カーネル正準相関分析）
- サポートベクターマシンの基礎
- カーネル法の方法論
  - 共通する方法
  - Representer定理
- その他の話題
  - カーネルの選択
  - 低ランク近似

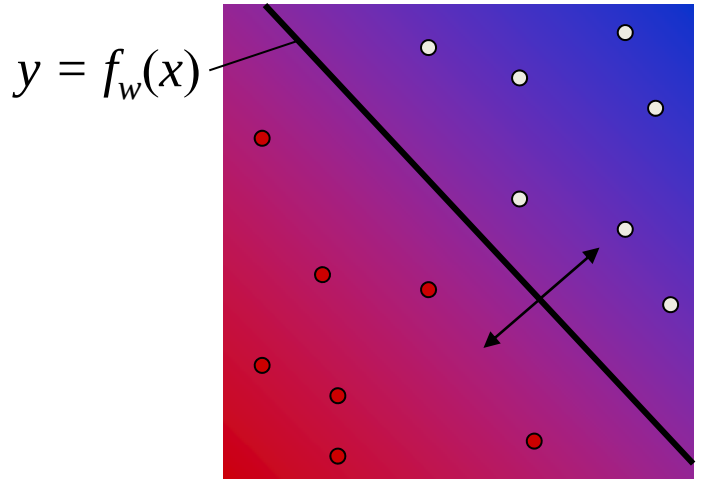
線形識別

- 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

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

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

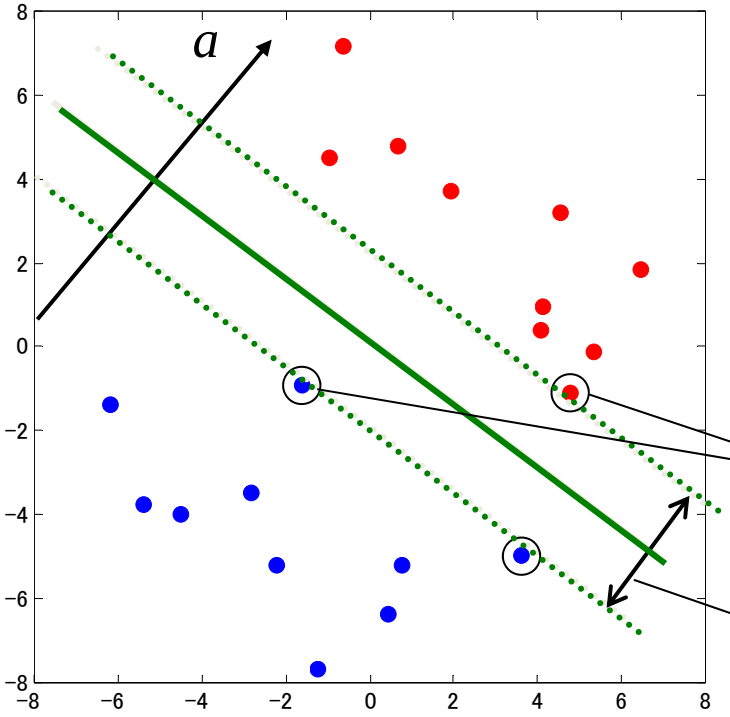
  - 問題：　未知の x に対しても
    正しく答えられるように
    $f_w(x)$　を構成せよ



マージン最大化

- 線形のサポートベクターマシン
  - 仮定： データは線形分離可能

    ⇒ 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，データの2クラス間の距離.

識別関数は，2つの境界の真中

サポートベクター

マージン

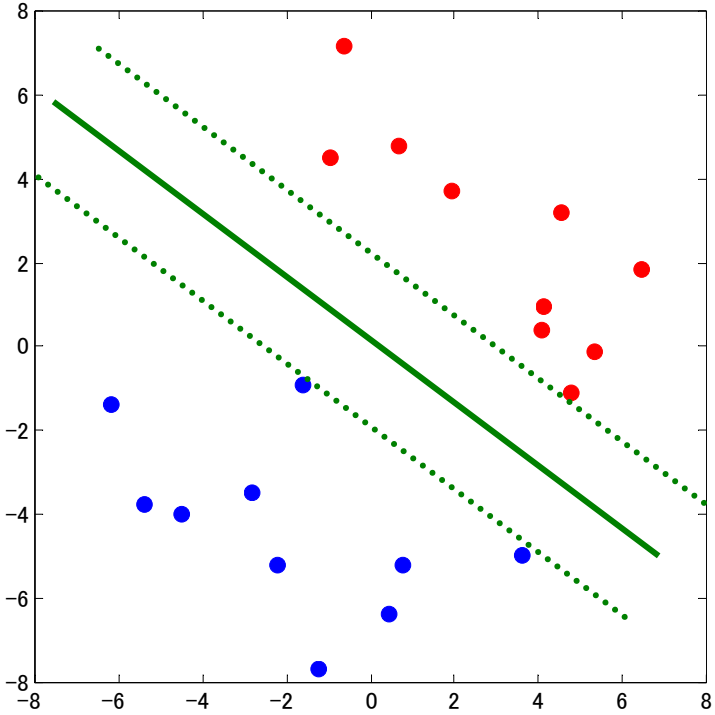
– マージンの計算

(a,b) を正の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

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

➡

$$\text{マージン} = \frac{2}{\|a\|}$$



サポートベクターマシン：ハードマージン

- マージン最大化基準による識別関数

$$\max_{a,b} \frac{1}{\|a\|} \quad \text{subject to} \quad \begin{cases} a^T X^i + b \geq 1, & \text{if } Y_i = 1 \\ a^T X^i + b \leq -1, & \text{if } Y_i = -1 \end{cases}$$

サポートベクターマシン（ハードマージン）

$$\min_{a,b} \|a\|^2 \quad \text{subject to} \quad Y_i(a^T X^i + b) \geq 1 \quad (\forall i)$$

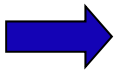
  - 2次最適化（Quadratic Program, QP）
    - 線形制約のもとでの2次関数の最小化.
    - 凸最適化： 局所最適解の問題がない.
    - 有効なソフトウエアの利用が可能

サポートベクターマシン：ソフトマージン

- 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

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

サポートベクターマシン（ソフトマージン）

$$\min_{a,b,\xi_i} \| a \|^2 + C\sum_{i=1}^{N} \xi_i \qquad \text{subject to} \qquad Y_i(a^T X^i + b) \geq 1 - \xi_i$$

$$\xi_i \geqq 0$$

  - 上の問題もQP.
  - C はユーザが決める必要がある．Cross-validation がよく使われる．

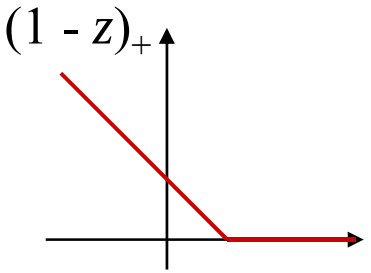
SVMと正則化

– ソフトマージンSVMは以下の正則化問題と同値 （$C = 1/\lambda$ ）

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

損失関数　　正則化項

ただし $(z)_+ = \max(z, 0)$



– 損失関数： ヒンジ損失

$$\ell(f(x), y) = \left(1 - yf(x)\right)_+$$

– [Exercise] ソフトマージンSVMと上の正則化の同値性を確認せよ.

カーネル化されたSVM

- SVMのカーネル化（kernelization)
  - Data: $(X_1, Y_1), \ldots, (X_N, Y_N)$
    - X_i: 任意の集合 Ω に値を持つ. $Y_i \in \{+1, -1\}$
  - 正定値カーネル k による特徴ベクトル

$$X_1, \ldots, X_N \quad \rightarrow \quad \Phi(X_1), \ldots, \Phi(X_N)$$

  - RKHS H における線形識別関数

$$f(x) = \mathrm{sgn}\left(\langle h, \Phi(x) \rangle + b\right) = \mathrm{sgn}\left(\underline{h(x) + b}\right) \qquad (h \in H)$$

非線形関数

- マージン最大化

$$\min_{h,b,\xi_i} \| h \|_H^2 + C\sum_{i=1}^{N} \xi_i \qquad \text{subject to} \qquad Y_i\left(\langle h, \Phi(X_i)\rangle + b\right) \geq 1 - \xi_i$$
$$\xi_i \geqq 0$$

正則化問題として表わすと

$$\min_{h,b} \quad \sum_{i=1}^{N} \left(1 - Y^i(\langle h, \Phi(X_i)\rangle + b)\right)_+ + \lambda \| h \|_H^2$$

– 最適な h は次の形

$$h(x) = \sum_{i=1}^{N} w_i \Phi(X_i) = \sum_{i=1}^{N} w_i k(x, X_i)$$

∵） 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，
$h = h_0 + h_1$ の分解で，正則化表現第1項は h_0 のみに依存．第2項は
$\| h \|^2 = \| h_0 \|^2 + \| h_1 \|^2$ により $h_1 = 0$のときが最適．

– Gram行列表現

$$\| h \|_H^2 = w^T K w, \qquad \langle h, \Phi(X_i)\rangle = (Kw)_i. \qquad K_{ij} = k(X_i, X_j).$$

([Exercise] Check them.)

サポートベクターマシン（カーネル化）

$$\min_{w,b,\xi_i} w^T K w + C\sum_{i=1}^{N}\xi_i \qquad \text{subject to} \qquad Y_i((Kw)_i + b) \geq 1 - \xi_i$$
$$\xi_i \geqq 0$$

- この最適化もQP.　標準的ソルバーが使える.
- 実際には，双対問題を考えることが多い.（２日目で述べる）
- 係数 C とカーネル（あるいはカーネル内のパラメータ）の選択は，cross-validationによることが多い.

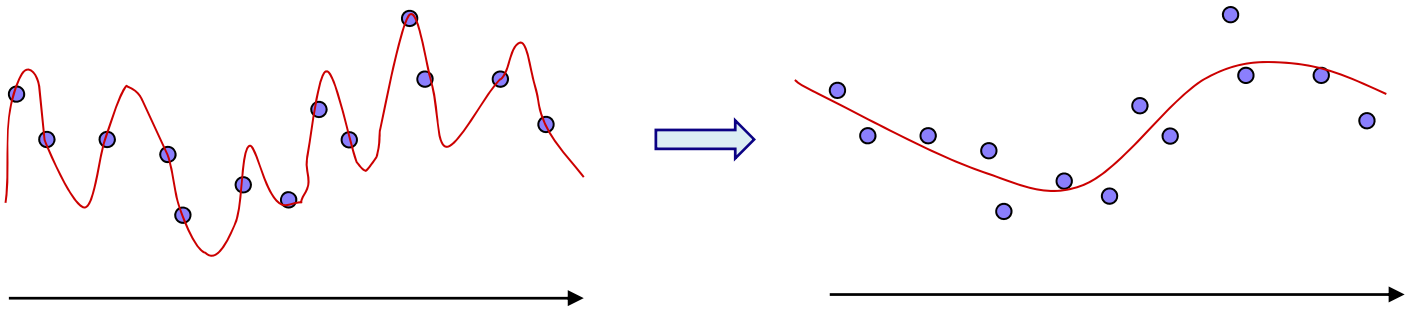
補足：正則化

- 例
  - リッジ回帰：　$\min_{f \in H} \sum_{i=1}^{N} (Y_i - f(X_i))^2 + \lambda \| f \|_H^2$
  - SVM：　$\min_{b, f \in H} \sum_{i=1}^{N} (1 - Y_i (f(X_i) + b))_+ + \lambda \| f \|_H^2$

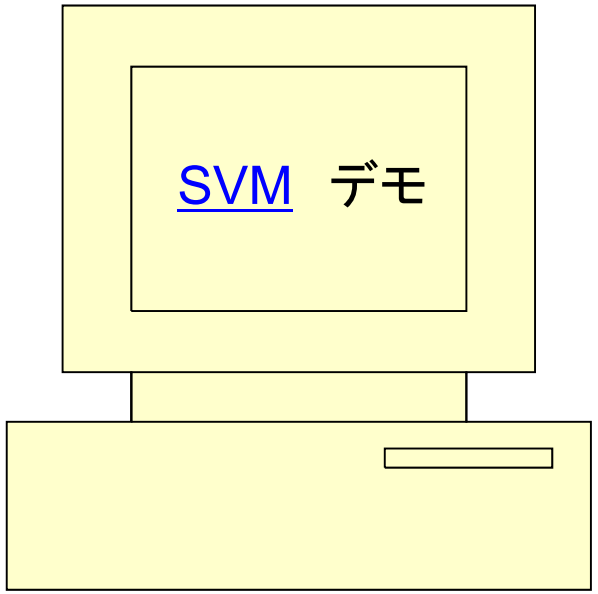
f が多様な関数をとりえると，第1項目（損失）だけでは解が一意でない．

⇒　正則化項（第2項目）の付加

- 正則化項は滑らかさに関係するものが多い（平滑化）



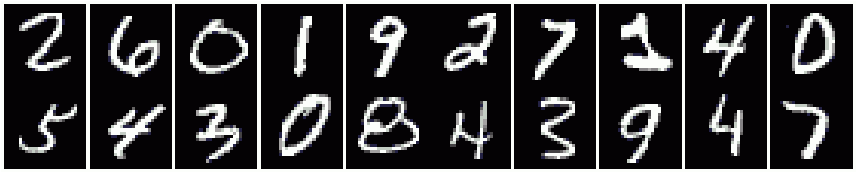
- SVMのJava applet



http://svm.dcs.rhbnc.ac.uk/pagesnew/GPat.shtml

SVMの文字認識への応用

– MNIST: Handwritten digit recognition
  • 28 x 28 binary pixels.
  • 60000 training data
  • 10000 test data



– Some results

	kNN	10PCA +Quad	RBF +Linear	LeNet-4	LeNet-5	SVM poly4	RS-SVM poly5
Test Err (%)	5.0	3.3	3.6	1.1	0.95	1.1	1.0

LeCun et al. 2001

概要

- Kernel PCAの応用
- Kernel CCA（カーネル正準相関分析）
- サポートベクターマシンの基礎
- カーネル法の方法論
  - 共通する方法
  - Representer定理
- その他の話題
  - カーネルの選択
  - 低ランク近似

カーネル法導出に共通する構造

– 正定値カーネルの定める特徴写像によるデータの変換

$$X_1,\ldots,X_N \quad \rightarrow \quad \Phi(X_1),\ldots,\Phi(X_N)$$

高次モーメント／非線形性を抽出

– 特徴ベクトルにRKHS上で「線形の」アルゴリズムを適用.

– 目的関数の最適解は特徴ベクトルの線形和

$$f(x)=\sum_{i=1}^{N} a_i \Phi(X_i)$$

で与えられることが多い.

– 目的関数は, Gram行列で表現される.

その最適化は各アルゴリズムによって異なる.

（KPCA, KCCA→固有値問題； SVM→QP）

– いったんGram 行列が得られれば, 後はデータサイズ N に依る計算量

— 元の空間の次元などに依存しない. 高次元データに有利.

Representer Theorem

- 正則化の問題（復習）
  - リッジ回帰　$\min_{f \in H} \sum_{i=1}^{N} \left(Y^i - f(X^i)\right)^2 + \lambda \| f \|_H^2$
  - 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）

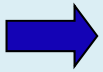
  $h_1(x), \ldots, h_d(x)$ ： 固定された関数（SVMの定数bなど）

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

Representer Theorem

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

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



$(*)$ の解 f は $\widetilde{H}_N$ 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 N 次元の最適化に還元できる

- Representer theorem の証明

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

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

$f = \tilde{f}_N + f_\perp$　　$\langle f_\perp, k(\cdot, x_i)\rangle = 0$　（$\forall i$）

⇒

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

$\rightarrow L$ の値は $\tilde{f}_N$ だけで決まる

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

$\rightarrow \Psi$ の値は $f_\perp = 0$ が有利

⇒

$f \in \tilde{H}_N$　に最適解がある

（証明終）

概要

- Kernel PCAの応用
- Kernel CCA（カーネル正準相関分析）
- サポートベクターマシンの基礎
- カーネル法の方法論
  - 共通する方法
  - Representer定理
- その他の話題
  - カーネルの選択
  - 低ランク近似

カーネルの選択

- How to choose / design a kernel?
  - 問題の構造に適したカーネルを用いるべし (構造化データ:２日目)
  - 教師付き学習（e.g. SVM）→ cross-validation.
  - 教師無し学習（e.g. kernel PCA, kernel CCA）→ 理論的に裏付けのある方法はほとんど無いのが現状.
    - Suggestion: 関連する教師付学習を作ってCVを使う.
    - Gaussian kernelに対するheuristics

$$\sigma = \mathrm{med}\{\| X_i - X_j \| \mid i \neq j\}$$

  - カーネル学習
    - Multiple kernel learning (MKL):

$$k(x,y) = \sum_{a=1}^{M} c_a k_a(x,y) \qquad \left(\sum_{a=1}^{M} c_a = 1,\ c_a \geq 0\right)$$

の形でカーネルも最適化する.（計算がハード）

補足：教師有り学習と教師無し学習

- 教師有り学習（Supervised learning）
  - Data for input X and output Y are prepared.
  - Y is regarded as “supervisor” or “teacher” of the learning.
    e.g. classification, regression, prediction.

$$X \rightarrow f(X) \approx Y$$

- 教師無し学習（Unsupervised learning）
  - There is no teaching data Y .
    e.g. PCA, CCA, clustering.

- 半教師有り学習（Semisupervised learning）
  - (X, Y) の訓練データと X のテストデータが事前に与えられている.

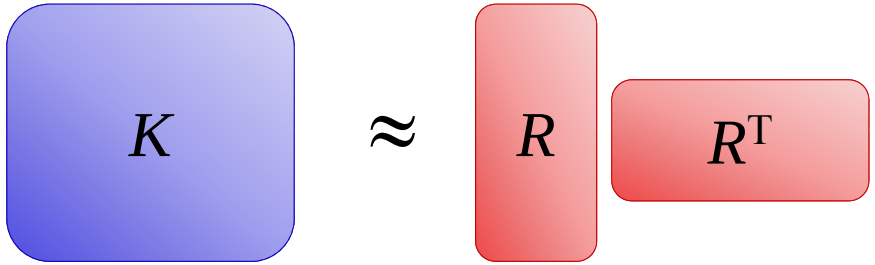
グラム行列の低ランク近似

- カーネル法： Gram行列を求めてしまえば，データ数 N に依る計算量
  -- もとの空間の次元などに依らない．高次元データに有利．

- 逆に，データ数 N が大きいと，Gram 行列 K に関する演算は困難．
  - 逆行列，固有値分解は $O(N^3)$の演算量必要

- 低ランク近似:

$$K \approx RR^T \qquad R\text{: } N \text{ x } r \text{ 行列} \quad (r << N)$$



– 計算量の大きな削減が可能.　　例）kernel ridge regression,

$$Y^T(K+\lambda I_N)^{-1}\mathbf{k}(x) \approx Y^T(RR^T+\lambda I_N)^{-1}\mathbf{k}(x)$$
$$= \frac{1}{\lambda}\left\{Y^T\mathbf{k}(x) - Y^T R(R^T R+\lambda I_r)^{-1}R^T\mathbf{k}(x)\right\},$$

演算量　$O(rN + r^3)$.

– 低ランク近似の2つの方法:

- Incomplete Cholesky decomposition:
  sample complexity$O(r^2N)$, space complexity $O(rN)$.
- Nyström approximation: random sampling + eigendecomposition.

– Gram行列は，多くの微小な固有値を持つことが多く，低ランク近似の方法が有効.

その他のカーネル法

- カーネルFisher判別分析（kernel FDA） (Mika et al. 1999)
- カーネルロジスティック回帰 （Roth 2001, Zhu&Hastie 2005）
- カーネル Partial Least Square（kernel PLS）（Rosipal&Trejo 2001）
- カーネル K-means クラスタリング （Dhillon et al 2004）
- SVMの仲間
  - Support vector regression (SVR, Vapnik 1995)
  - ν-SVM (Schölkopf et al 2000)
  - one-class SVM (Schölkopf et al 2001)

セクション3のまとめ

- さまざまな線形解析手法のカーネル化 が可能　← 効率的な内積計算
  - Kernel PCA, SVM, kernel CCA, etc.

- 最適解は多くの場合，特徴ベクトルの線形和

$$f(x) = \sum_{i=1}^{N} a_i \Phi(X_i)$$

  で与えられる　(representer theorem).

- 問題は，データサイズ N のGram行列によって表現される.
  - 高次元で，中程度までのデータサイズに適している.
  - データサイズが大きい場合には，低ランク近似が有効.

- 正定値カーネルさえ定義されれば，任意のデータ型に適用可能.
  structured (non-vectorial) data, such as graphs, strings, etc

参考文献

赤穂. (2000) カーネル正準相関分析. 第3回情報論的学習理論ワークショップ予稿集 (IBIS2000).

Bach, F.R. and M.I. Jordan. (2002) Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48.

Dhillon, I. S., Y. Guan, and B. Kulis. (2004) Kernel k-means, spectral clustering and normalized cuts. Proc. 10th ACM SIGKDD Intern. Conf. Knowledge Discovery and Data Mining (KDD), 551–556.

Fukumizu, K., F.R. Bach, and A. Gretton. (2007) Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8:361–383.

Hardoon, D.R., S. Szedmak, and J. Shawe-Taylor. (2004) Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16:2639–2664.

LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner. (2001) Gradient-based learning applied to document recognition. In Simon Haykin and Bart Kosko, eds, *Intelligent Signal Processing*, 306–351. IEEE Press.

Melzer, T., M. Reiter, and H. Bischof. (2001) Nonlinear feature extraction using generalized canonical correlation analysis. *Proc. Intern. Conf. .Artificial Neural Networks* (*ICANN*), 353–360.

Mika, S., G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller. (1999) Fisher discriminant analysis with kernels. In Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, edits, *Neural Networks for Signal Processing, volume IX,* 41–48. IEEE.

Murphy P.M. and D.W. Aha. (1994) UCI repository of machine learning databases. *Tech report*, UC Irvine, Dept Information and Computer Science. http://www.ics.uci.edu/~mlearn/MLRepository.html

Rosipal, R. and L.J. Trejo. (2001) Kernel partial least squares regression in reproducing kernel Hilbert space. *Journal of Machine Learning Research*, 2: 97–123.

Roth, V. Probabilistic discriminative kernel classifiers for multi-class problems. In *Pattern Recognition: Proc. 23rd DAGM Symposium*, 246–253. Springer, 2001.

Schölkopf, B., A. Smola, and K-R. Müller. (1998) Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319.

Schölkopf, B., A. Smola, R.C. Williamson, and P.L. Bartlett. (2000) New support vector algorithms. *Neural Computation, 12(5):1207–1245.*

Schölkopf, B., J.C. Platt, J. Shawe-Taylor, R.C. Williamson, and A.J.Smola. (2001) Estimating the support of a high-dimensional distribution. *Neural Computation,* 13(7):1443–1471.

Vapnik, V.N. *The Nature of Statistical Learning Theory*. Springer 1995.