カーネル法入門
2. さまざまなカーネル法

福水健次

統計数理研究所／総合研究大学院大学

大阪大学大学院基礎工学研究科・集中講義

2014 September

カーネル法のいくつかの例を通して，カーネル法の考え方を理解する，

- カーネルPCA
- カーネルCCA
- カーネルリッジ回帰
- サポートベクターマシンの初歩

カーネルPCA

再掲: カーネルPCAのアルゴリズム

- 中心化Gram行列 $\tilde{K}_X$ の計算

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

- $\tilde{K}_X$ の固有分解

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

$\lambda_1 \geq \lambda_2 \geq \Lambda \geq \lambda_N \geq 0$　固有値

$u_1, u_2, \mathrm{K}, u_N$　単位固有ベクトル

- 第p主成分方向 $f_p = \sum_j \frac{1}{\sqrt{\lambda_p}} u_{pj} \left\{\Phi\left(X^{(j)}\right) - \frac{1}{n}\sum_b \Phi(X^{(b)})\right\}$
- $X^{(i)}$の第p主成分 $= \sqrt{\lambda_p} u_{pi}$

カーネルPCAの例

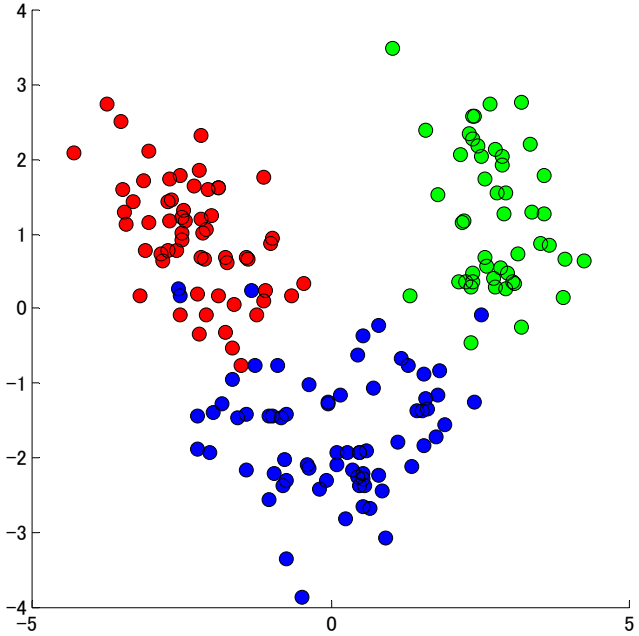
- Wine データ (UCI repository)

  3種類のイタリアワインに関する，13 次元の化学測定値

  178 データ.

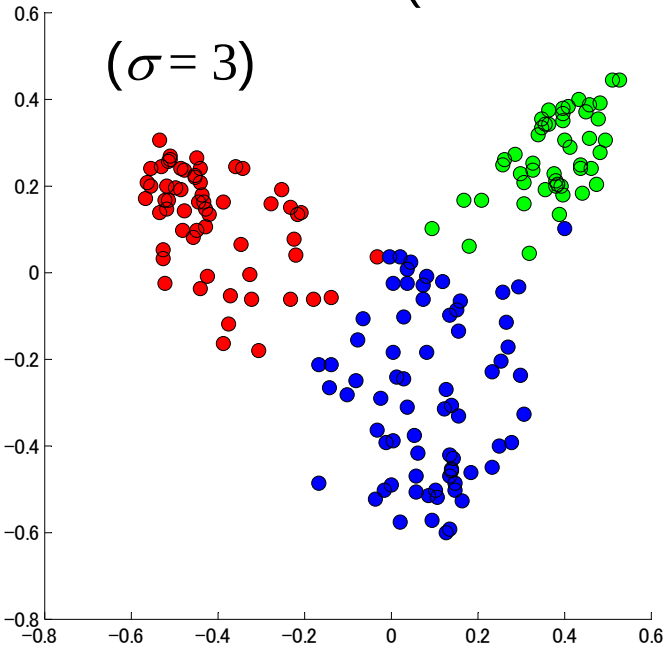
  クラスの情報はカーネルPCAには用いていない

Linear PCA

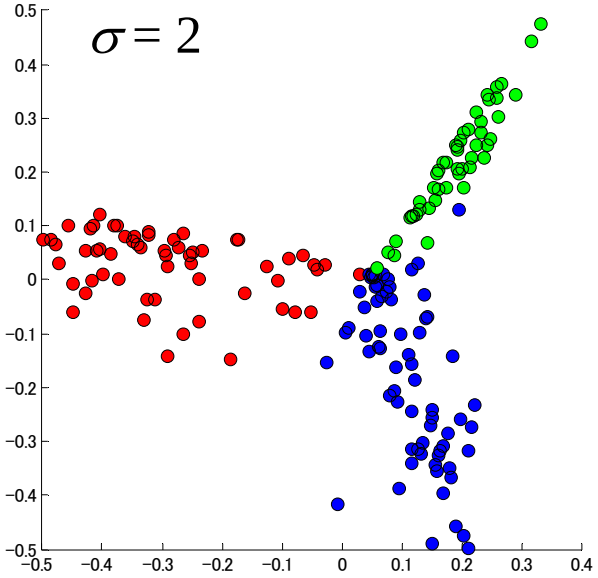


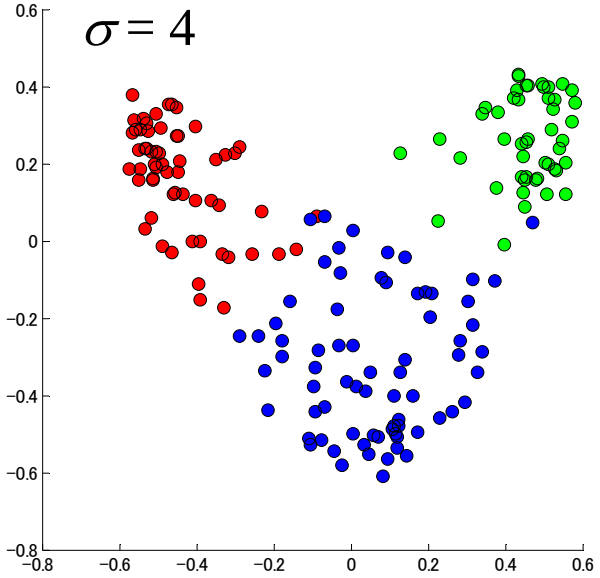
Kernel PCA (Gaussian kernel)

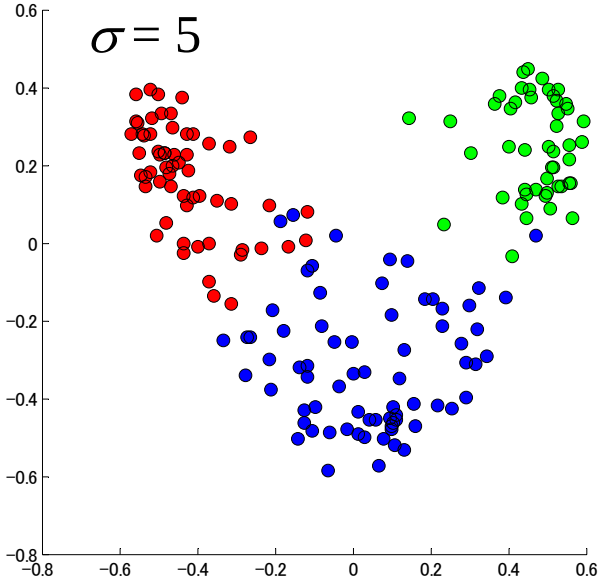
$(\sigma = 3)$



Kernel PCA (Gaussian)





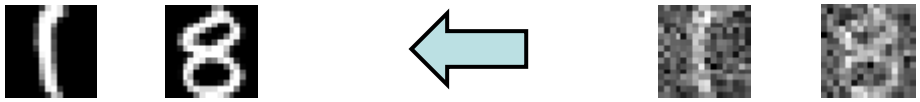


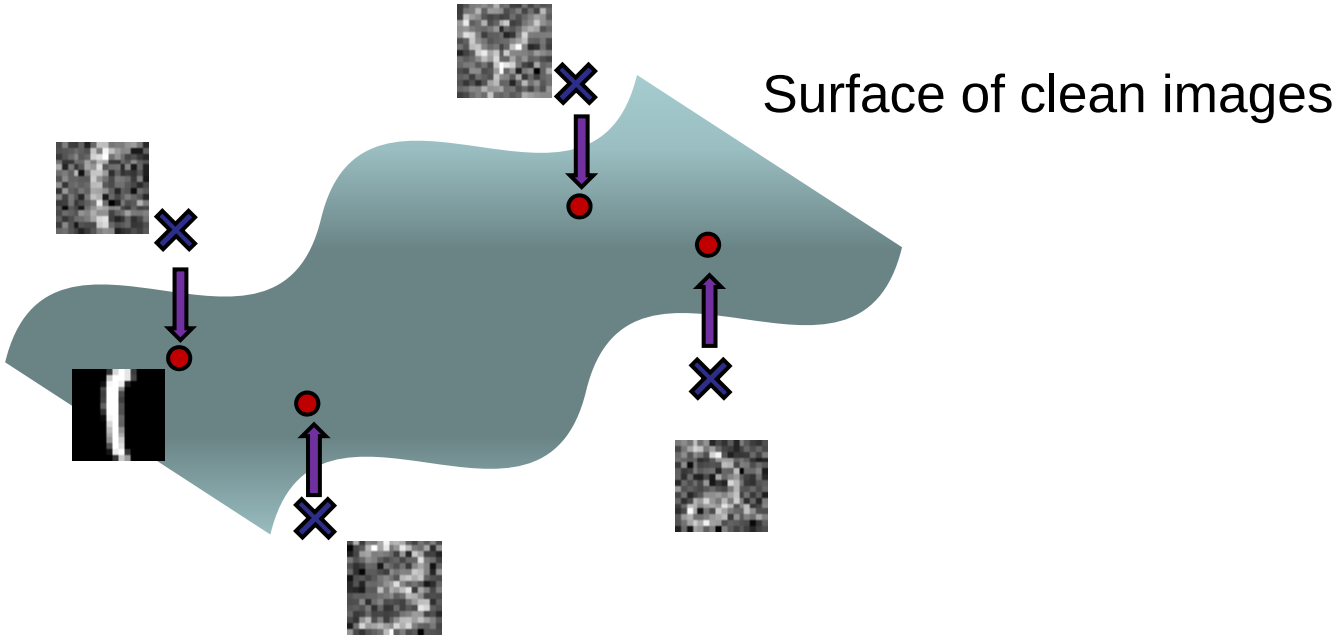
$$k_G(x, y) = \exp\left(-\|x - y\|^2 / \sigma^2\right)$$

応用:ノイズ削減

– カーネルPCA をノイズ削減に応用:

主成分 = 有益な情報, その他の方向 = ノイズ



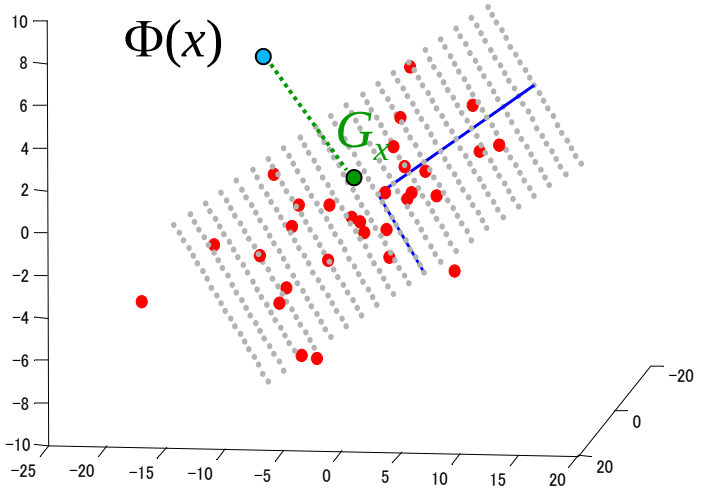


■ カーネルPCAの適用

- V_d: d個の主成分方向が張るd次元部分空間
- G_x: $\Phi(x)$ の V_d への正射影 = ノイズ除去された特徴ベクトル.
- 原像問題: G_x を近似する特徴ベクトル$\Phi(\hat{x})$を与える原像 $\hat{x}$ を求める.

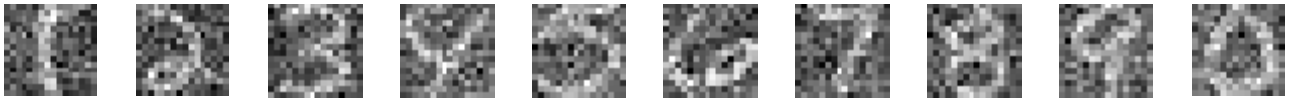
$$\hat{x} = \arg\min_{x'} \left\| \Phi(x') - G_x \right\|^2$$

G_x は Φ の像に入っているとは限らない.

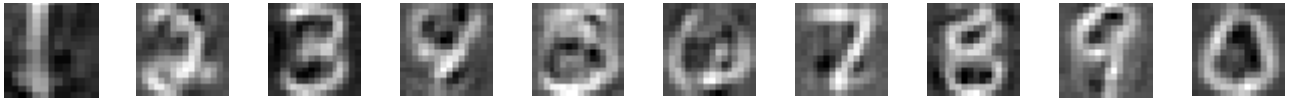


■ USPSデータ

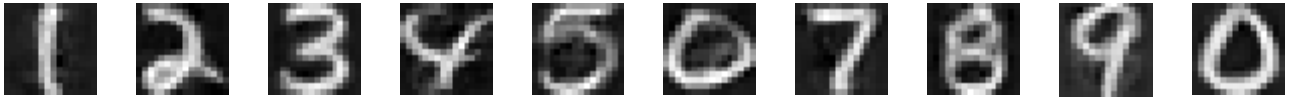
Original data (NOT used for PCA)



Noisy images



Noise reduced images（linear PCA）



Noise reduced images（kernel PCA, Gaussian kernel）

(Generated by Matlab stprtool by V. Franc)

カーネルPCAの特徴

- 非線形な次元削減： 非線形特徴を捉えることが可能.
- 識別などさらなる解析の前処理として用いることも多い（次元削減／特徴抽出）
- 結果はカーネル(あるいはパラメータ)に大きく依存する.
  線形PCAのように，軸に対する解釈を与えることは容易ではない.
- カーネル／パラメータをどのように決めるか?
  - Cross-validationの適用はそのままではうまくいかない.
  - 前処理として用いる場合は，最終的な結果のよさをcross-vaildationなどで測れる場合がある.

カーネル正準相関分析

正準相関分析

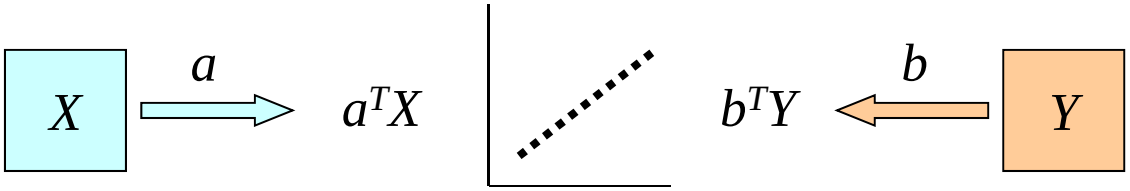
– 正準相関分析　Canonical correlation analysis (CCA, Hotelling 1936)

2つの多次元の変数の線形相関を見る方法.

- Data $(X_1, Y_1), \ldots, (X_N, Y_N)$
- X_i: m-次元,　Y_i: ℓ-次元

$a^T X$ と $b^T Y$ の相関が最大になるように，ベクトル a, b を求める.

$$\rho = \max_{a,b} \mathrm{Corr}[a^T X, b^T Y] = \max_{a,b} \frac{\mathrm{Cov}[a^T X, b^T Y]}{\sqrt{\mathrm{Var}[a^T X]\mathrm{Var}[b^T Y]}},$$



■ CCAの解法

$$\max_{a,b} a^T \hat{V}_{XY} b \quad \text{subject to} \quad a^T \hat{V}_{XX} a = b^T \hat{V}_{YY} b = 1.$$

– 一般化固有問題に還元される:

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

[演習問題: これを導出せよ. (ヒント. Lagrange乗数法を用いよ.)]

– 解:

$$a = V_{XX}^{1/2} u_1, \ b = V_{YY}^{1/2} v_1$$

u_1 (v_1) は $\hat{V}_{XX}^{-1/2} \hat{V}_{XY} \hat{V}_{YY}^{-1/2}$の最大特異値に対する
左(右) 単位固有ベクトル.

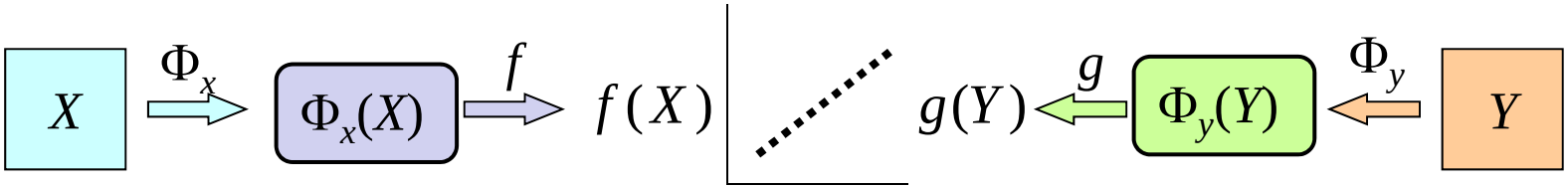
カーネル正準相関分析

– カーネルCCA (Akaho 2000, Melzer et al. 2002, Bach et al 2002)
  - 線形相関だけでなく，依存性／関連性が調べられる．
  - データ: $(X_1, Y_1), \ldots, (X_N, Y_N)$ 任意の確率変数（$\mathbf{R}^m$とは限らない）

  - カーネルk_X, k_Yを決め，特徴ベクトルにCCA を施す．

$$X_1, \ldots, X_N \mapsto \Phi_X(X_1), \ldots, , \Phi_X(X_N) \in H_X,$$
$$Y_1, \ldots, Y_N \mapsto \Phi_Y(Y_1), \ldots, \Phi_Y(Y_N) \in H_Y.$$

$$\max_{f \in H_X, g \in H_Y} \frac{\mathrm{Cov}[f(X), g(Y)]}{\sqrt{\mathrm{Var}[f(X)]\mathrm{Var}[g(Y)]}} = \max_{f \in H_X, g \in H_Y} \frac{\sum_i^N \langle f, \tilde{\Phi}_X(X_i)\rangle \langle \tilde{\Phi}_Y(Y_i), g\rangle}{\sqrt{\sum_i^N \langle f, \tilde{\Phi}_X(X_i)\rangle^2 \sum_i^N \langle g, \tilde{\Phi}_Y(Y_i)\rangle^2}}$$



- $f = \sum_{i=1}^{N} \alpha_i \widetilde{\Phi}_X(X_i), g = \sum_{i=1}^{N} \beta_i \widetilde{\Phi}_Y(Y_i)$としてよい.

(kernel PCAと同じ)

$$\max_{\alpha \in \mathrm{R}^N, \beta \in R^N} \frac{\alpha^T \widetilde{K}_X \widetilde{K}_Y \beta}{\sqrt{\alpha^T \widetilde{K}_X^2 \alpha \; \beta^T \widetilde{K}_Y^2 \beta}}$$

$\widetilde{K}_X, \widetilde{K}_Y$: 中心化Gram行列

- 正則化が必要: 自明な解,

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

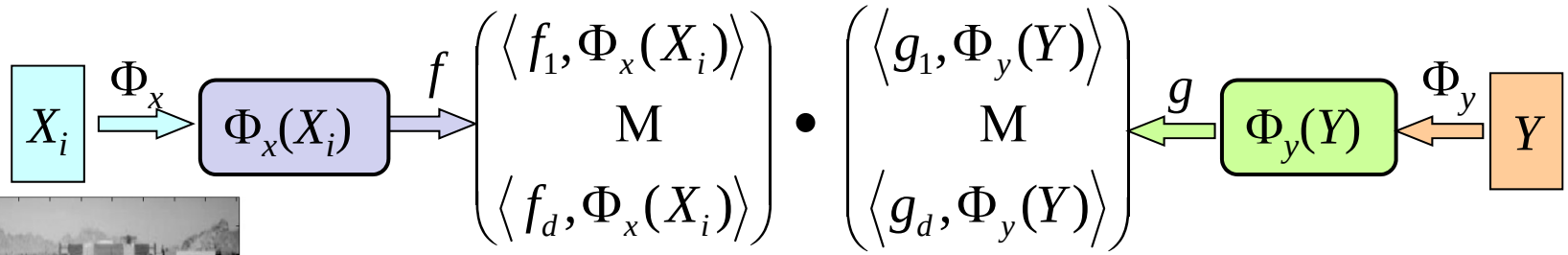
(推定量は一致性を持つ. Fukumizu et al JMLR 2007)

- 解: 一般化固有値問題

$$\begin{pmatrix} O & \widetilde{K}_X \widetilde{K}_Y \\ \widetilde{K}_Y \widetilde{K}_X & O \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \rho \begin{pmatrix} \widetilde{K}_X^2 + \varepsilon_N \widetilde{K}_X & O \\ O & \widetilde{K}_Y^2 + \varepsilon_N \widetilde{K}_Y \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

KCCAの応用

– テキストからの画像検索の問題(Hardoon, et al. 2004).
  - X_i: 画像,
    Y_i: 対応するテキスト (同じウエブページから採取).
  - KCCAによって，d 個のベクトル $f_1,\ldots,f_d$ と $g_1,\ldots,g_d$ を取る．これらは，X と Y が最も依存するような特徴空間の部分空間を張る.
  - 新しい単語 Y_{new}に対し，最も内積の値が大きくなる画像を結果として返す.



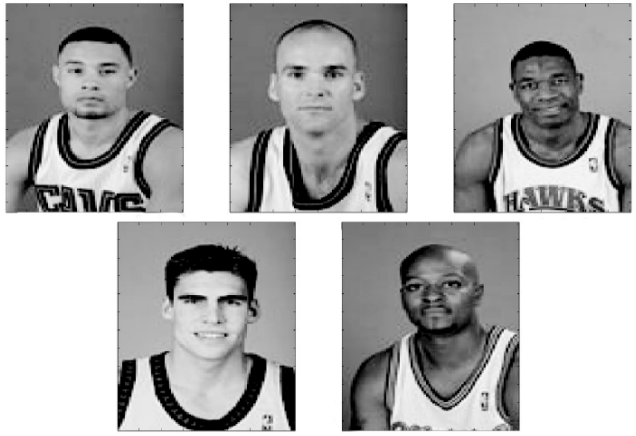
“at phoenix sky harbor on july 6, 1997. 757-2s7, n907wa, ...”

– 例:

- 画像: ガウスカーネル
- テキスト: Bag-of-words カーネル (単語の頻度).

Text -- “height: 6-11, weight: 235 lbs, position: forward,
born: september 18, 1968, split, croatia college: none”

検索された画像の結果（トップ５）



脳信号への応用

カーネルリッジ回帰

リッジ回帰

$(X_1, Y_1), \ldots, (X_n, Y_n)$: $(X_i \in \mathbf{R}^m,\ Y_i \in \mathbf{R})$

リッジ回帰: 線形回帰 +2ノルムによる正則化

$$\min_a \sum_{i=1}^{n} |Y_i^2 - a^T X_i\ |^2 + \lambda \|a\|^2$$

(定数項は簡単のため省略して書く)

– 解 (2次関数の最適化):

$$\hat{a} = (V_{XX} + \lambda I_n)^{-1} X^T Y$$

ここで

$$V_{XX} = X^T X, \quad X = \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix} \in \mathbf{R}^{n \times m}, \quad Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \in \mathbf{R}^n$$

– リッジ回帰は, V_{XX} が特異な (または特異に近い) 場合によく用いられる. -- 共線性への対処

カーネルリッジ回帰

- $(X_1, Y_1), \ldots, (X_n, Y_n)$: X 任意の変数, $Y \in \mathbf{R}$.
- リッジ回帰のカーネル化:　X に対するカーネル k を用いる

$$\min_{f \in H} \sum_{i=1}^{n} |Y_i - \langle f, \Phi(X_i) \rangle_H|^2 + \lambda \|f\|_H^2$$ Ridge regression on H

equivalently,

$$\min_{f \in H} \sum_{i=1}^{n} |Y_i - f(X_i)|^2 + \lambda \|f\|_H^2$$ Nonlinear ridge regr.

– 解は特徴ベクトルの線形和の形　　$f = \sum_{j=1}^{n} c_j \Phi(X_j),$

$f = \sum_{i=1}^{n} c_i \Phi(X_i) + f_\perp = f_\Phi + f_\perp$ とおく.
($f_\Phi \in Span\{\Phi(X_i)\}_{i=1}^{n}$, $f_\perp$: 直交補空間の成分)
このとき，目的関数 $= \sum_{i=1}^{n} | Y_i - \langle f_\Phi + f_\perp, \Phi(X_i) \rangle|^2 + \lambda \| f_\Phi + f_\perp \|^2$
$= \sum_{i=1}^{n} | Y_i - \langle f_\Phi, \Phi(X_i) \rangle|^2 + \lambda (\| f_\Phi \|^2 + \| f_\perp \|^2)$
第1項は$f_\perp$に依存しない．第2項は $f_\perp = 0$ のとき最小.

– 目的関数:

$$\| Y - K_X c \|^2 + \lambda \, c^T K_X c$$

– 解:　$\hat{c} = (K_X + \lambda I_n)^{-1} Y$

$$\hat{f}(x) = Y^T (K_X + \lambda I_n)^{-1} \mathbf{k}(x)$$

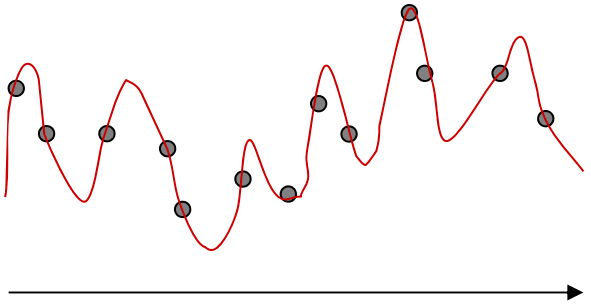
$$\mathbf{k}(x) = \begin{pmatrix} k(x, X_1) \\ \vdots \\ k(x, X_n) \end{pmatrix}$$

正則化

– 最小化問題

$$\min_f \sum_{i=1}^n |Y_i - f(X_i)|^2$$

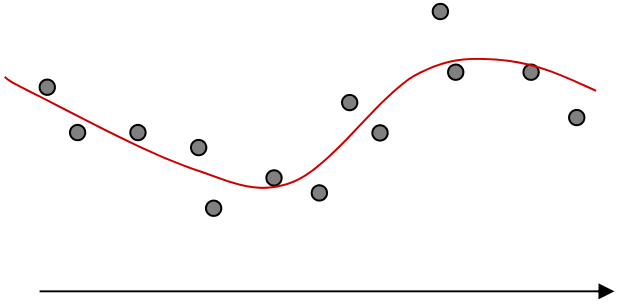
は，誤差0が可能．しかし，無数に解を持ち得る．



– 正則化

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

- 関数が滑らかになりやすい罰則項がよく用いられる．
- 滑らかさとRKHSノルムとの関係はChapter 3 を参照．



比較実験

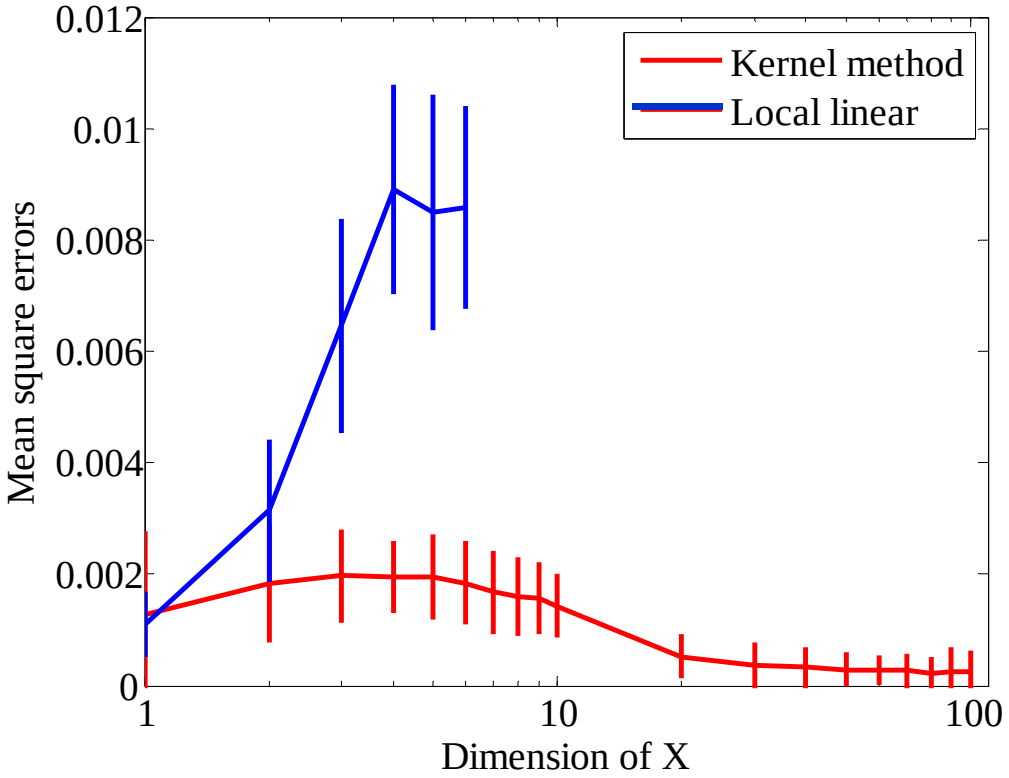
■ カーネルリッジ回帰　vs 局所線形回帰

$Y = 1/(1.5 + ||X||^2) + Z, \quad X \sim N(0, I_d),\ Z \sim N(0, 0.1^2)$

n = 100,　500 runs

カーネルリッジ回帰
　Gaussカーネル
局所線形回帰
　Epanechnikovカーネル*
　(R ‘locfit’)

バンド幅はともに
Cross-vaildationで選択.



*Epanechnikovカーネル = $\frac{3}{4}(1-u^2)\mathbf{1}_{\{|u|\leq 1\}}$

■ 局所線形回帰 (e.g., Fan and Gijbels 1996)

– K: 平滑化カーネル / Parzen ウィンドー $(K(x) \geq 0,\ \int K(x)dx = 1,$ 正定値性は仮定しない)

– 局所線形回帰

$E[Y|X = x_0]$ is estimated by

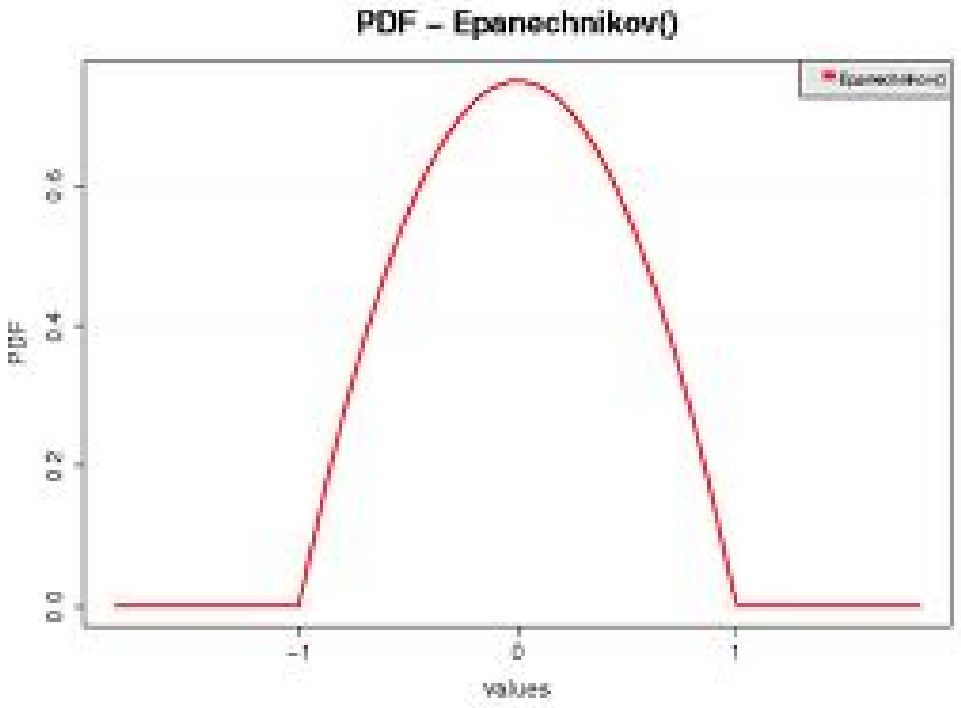
$$K_h(x) = h^{-d} K\left(\frac{x}{h}\right)$$

$$\min_{a,b} \sum_i |Y_i - a - b^T(X_i - x_0)|^2 K_h(X_i - x_0)$$

- 各 x_0 に対して最適化問題を解く．最適化は重み付最小2乗誤差問題として陽に解ける．
- 推定量の統計的性質は詳しく調べられている．
- X が1次元の時，局所線形回帰はよい結果を与える．
  一般に局所多項式回帰は理論的な最適性を有している．
- しかし，高次元データ（5，6次元以上）に弱いことが知られている．

– Epanechnikovカーネル

$$K(u) = \frac{3}{4}(1 - u^2) I_{[-1,1]}(u)$$



Nonparametric regression: theoretical comparison

Kernel ridge regression

$$\hat{E}_{reg}[Y|X=x] \coloneqq \mathbf{k}_X^T(x)(G_X + n\varepsilon_n I_n)^{-1}Y$$

– Convergence rate (Eberts & Steinwart 2011)

If k_X is Gaussian, and $E[Y|X] \in W_2^\alpha(P_X)$, (under some technical assumptions) for any $\rho > 0$,

$$E\left|\hat{E}_{reg}[Y|X] - E[Y|X]\right|^2 = O_p\left(n^{-\frac{2\alpha}{2\alpha+m}+\rho}\right) \quad (n \to \infty)$$

* $W_2^\alpha(P_X)$: Sobolev space of order α.

– Note: $O_p\left(n^{-\frac{2\alpha}{2\alpha+m}}\right)$ is the optimal rate for a linear estimator.

(Stone 1982).

– Local polynomial fitting attains this rate if the degree of polynomial is sufficient.

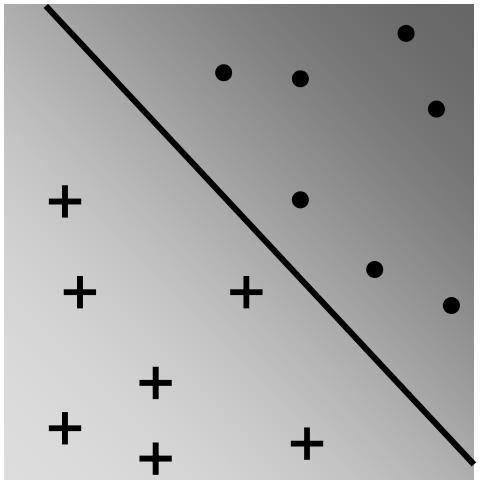
Support Vector Machine

2値識別問題

– 訓練データ

Input data　　Class label

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



– 線形識別

線形関数

$$h_{w,b}(x) = \mathrm{sgn}(w^T x + b)$$

目的関数

$$h_{w,b}(X^{(i)}) = Y^{(i)} \qquad \text{for all (or most) } i.$$

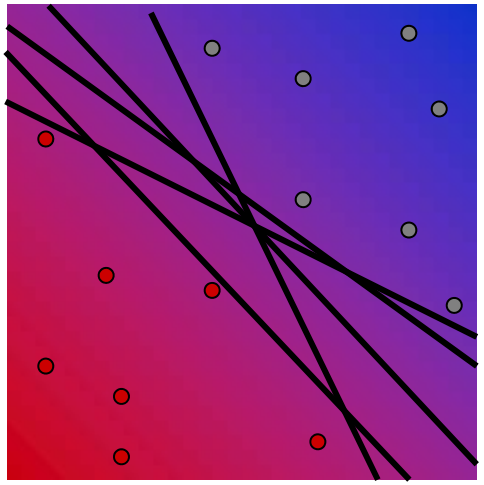
マージン最大化

■ 仮定: 線形識別可能.

ある w, b があって

$$\mathrm{sgn}\left(w^T X^{(i)} + b\right) = Y^{(i)} \qquad \forall i.$$

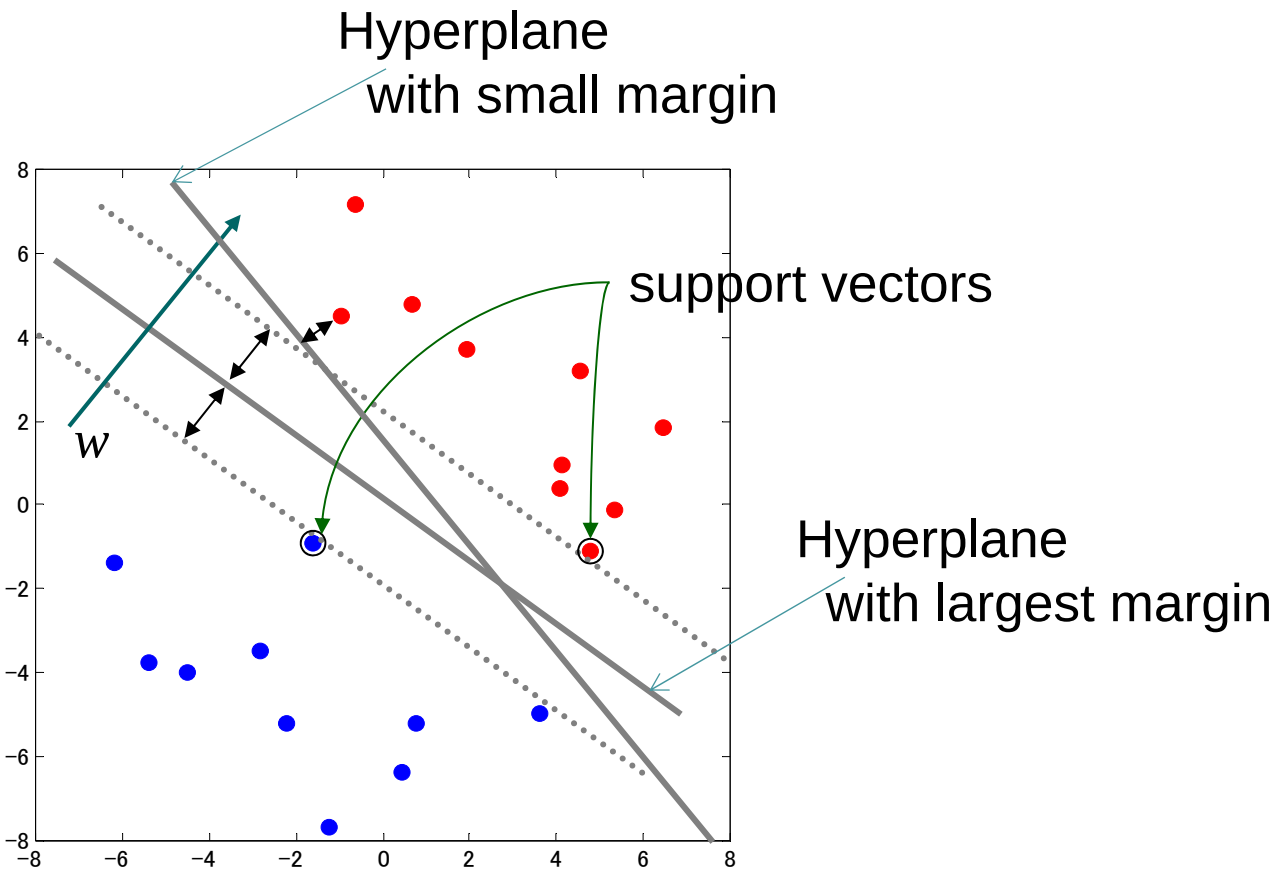
一般に解は一意ではなく無数にある(連続濃度).
どのように決めるか?



■ マージン最大化規準:

マージンを最大にする線形識別関数を選択する.

- マージン = w 方向で測った時の2つのクラスの距離.
- 識別境界は, 2つのクラスまでの距離の中間.


Hyperplane
with small margin
support vectors
Hyperplane
with largest margin
w
8
6
4
2
0
−2
−4
−6
−8
−8
−6
−4
−2
0
2
4
6
8

– マージンの測り方

スケールを固定するため以下のように正規化する

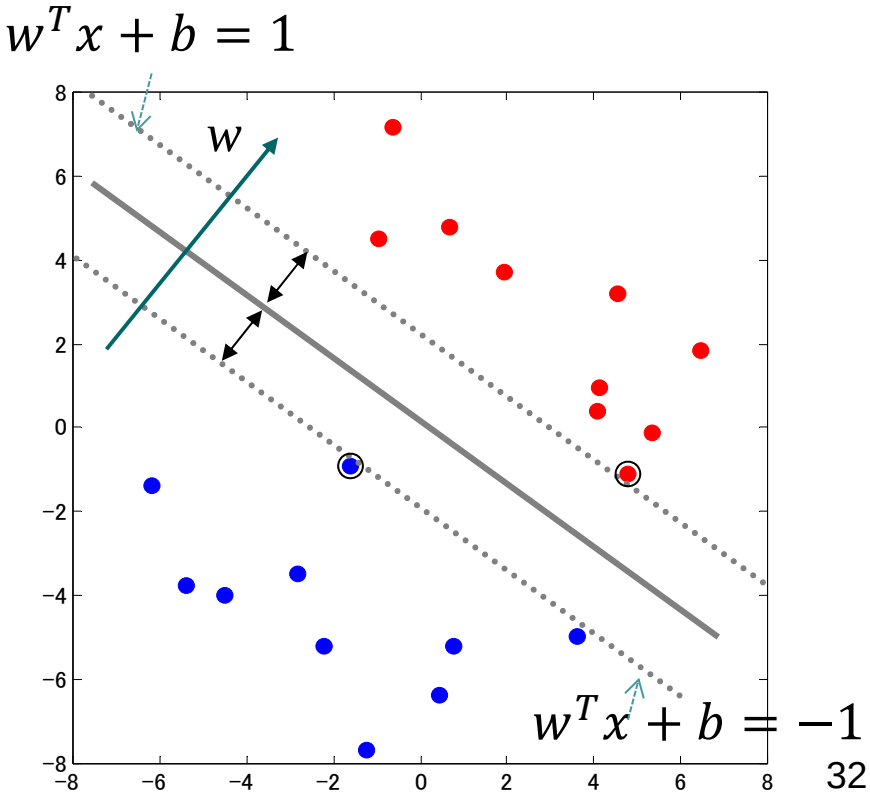
$\min(w^T X^{(i)} + b) = 1$ for the closest point with $Y^{(i)} = +1$,

$\min(w^T X^{(i)} + b) = -1$ for the closest point with $Y^{(i)} = -1$.

このとき,

$$\text{マージン} = \frac{2}{\|w\|}$$

[演習問題: これを示せ]



– マージン最大化識別器:

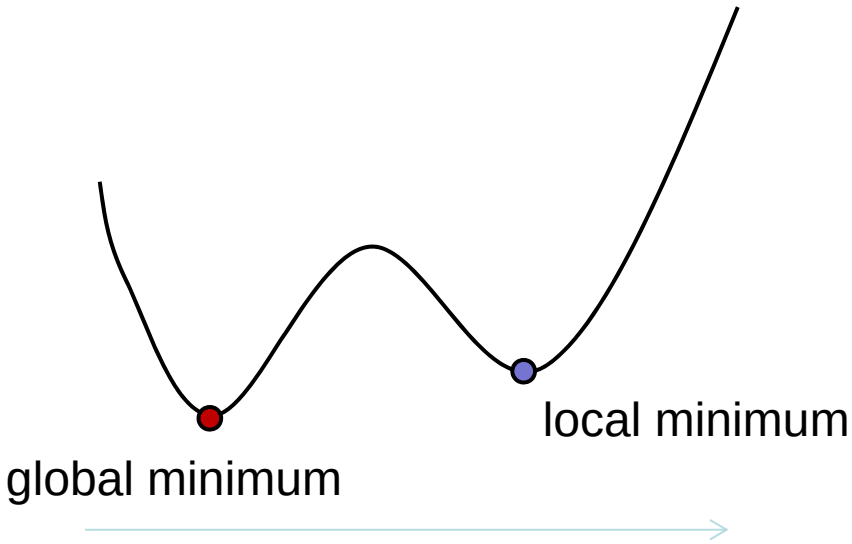
$$\max \frac{1}{\|w\|} \quad \text{subject to} \quad \begin{cases} w^T X^{(i)} + b \geq 1 \quad if \quad Y^{(i)} = +1 \\ w^T X^{(i)} + b \leq -1 \quad if \quad Y^{(i)} = -1 \end{cases}$$

以下と同値

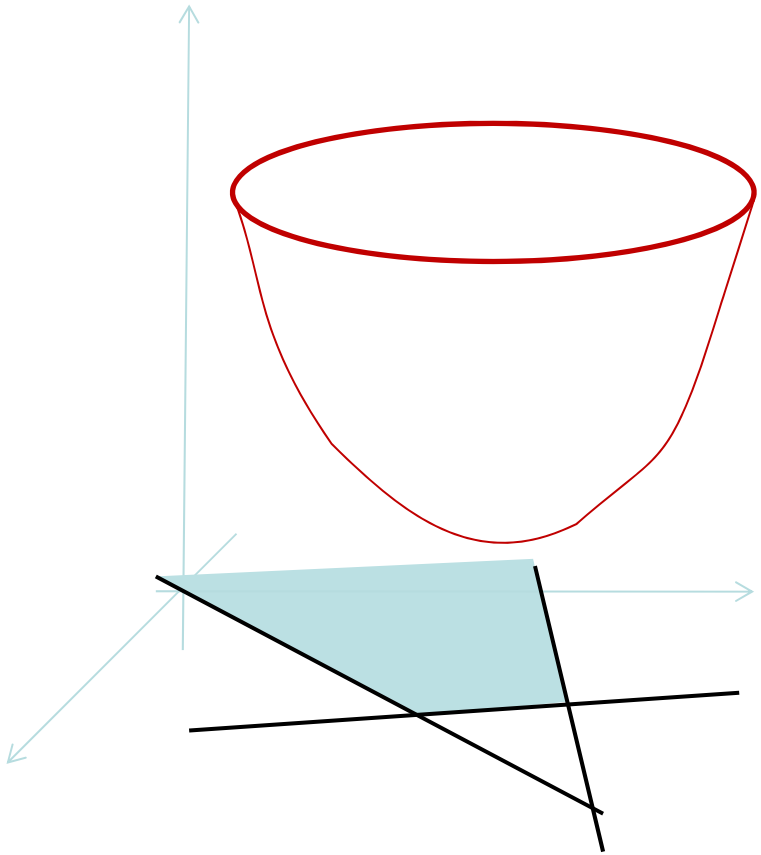
$$\min_{w,b} \|w\|^2 \qquad \text{subject to} \quad Y^{(i)}\left(w^T X^{(i)} + b\right) \geq 1 \ \ (\forall\, i).$$

線形サポートベクターマシン (ハードマージン)

- 2次計画(QP): 線形不等式制約のもとでの2次関数の最小化.
  → 凸最適化. 局所解がない!
- QP ソルバーは多くのソフトウエアパッケージで提供されている.



Non-convex function



Quadratic Program
(convex program)

ソフトマージンSVM

– 線形識別可能という仮定は非現実的

→緩和

ハードな制約: $Y^{(i)}\left(w^T X^{(i)} + b\right) \geq 1$

ソフトな制約: $Y^{(i)}\left(w^T X^{(i)} + b\right) \geq 1 - \xi_i, \quad \xi_i \geq 0.$

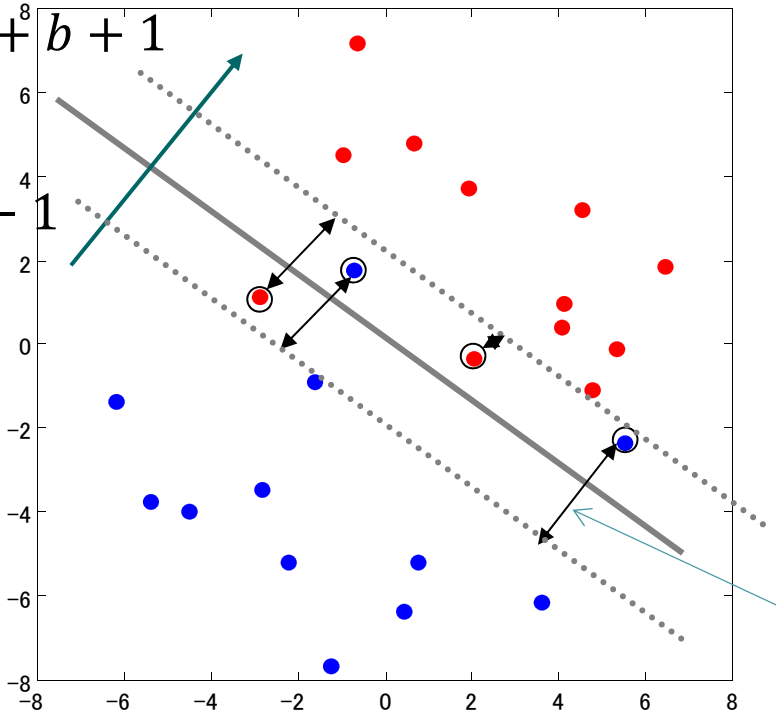
マージン最大化線形識別器 (ソフトマージン)

$$\min_{w,b,\xi} \|w\|^2 + C\sum_i \xi_i \quad \text{subj. to} \quad Y^{(i)}\left(w^T X^{(i)} + b\right) \geq 1 - \xi_i \ (\forall\, i), \quad \xi_i \geq 0.$$

- 最適化はやはりQP.
- C は ハイパーパラメータ. ユーザが決定する必要がある.

$y = w^T x + b + 1$

$y = w^T x + b - 1$



ξ_i

SVMのカーネル化

■ 線形SVMのカーネル化

- $\left(X^{(1)}, Y^{(1)}\right), \ldots, \left(X^{(N)}, Y^{(N)}\right)$: 訓練データ
  - $X^{(i)}$: 集合Sの元．（任意の型，ベクトルでなくてもよい)
  - $Y^{(i)} \in \{+1, -1\}$: 2値
- k: S上の正定値カーネル．H: kの定めるRKHS.
- $\Phi\left(X^{(i)}\right) = k(\cdot, X^{(i)})$: 特徴ベクトル.

- 特徴空間上の線形識別器:

$$f(x) = \mathrm{sgn}(\langle h, \Phi(x)\rangle_H + b) = sgn\,(h(x) + b)$$

$$\textit{c.f.} \quad f(x) = \mathrm{sgn}(w^T x + b)$$

– SVMの目的関数

$$\min_{h,b,\xi_i} \|h\|_H^2 + C\ \sum_i \xi_i \quad \text{subj. to} \quad Y^{(i)}\left(h(X^{(i)}) + b\right) \geq 1 - \xi_i$$

$$\xi_i \geq 0.$$

– Representer定理（後述）

最適な h は以下の形で探せば十分

$$h = \sum_i c_i k(\cdot, X^{(i)})$$

Note:

$$\|h\|_H^2 = \sum_i c_i c_j k(X^{(i)}, X^{(j)}), \quad h\left(X^{(i)}\right) = \sum_j c_j\, k(X^{(i)}, X^{(j)}).$$

■ SVM（カーネル版）

$$\min_{c,b,\xi} \sum_i c_i c_j \, k(X^{(i)}, X^{(j)}) + C \sum_i \xi_i$$

$$\text{subj. to} \quad Y^{(i)}\left(\sum_j c_j \, k(X^{(i)}, X^{(j)}) + b\right) \geq 1 - \xi_i$$

$$\xi_i \geq 0.$$

- 最適化はやはりQP.
- 双対問題のほうが容易に解ける(Chap. 5).
- ハイパーパラメータ C とカーネルのパラメータは，cross-validationで選択する場合が多い.

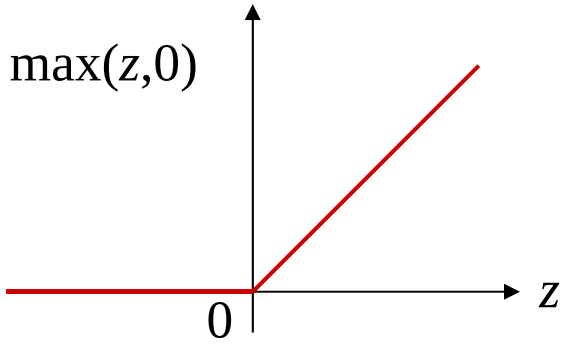
SVMと正則化

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

$$\min_{w,b} \sum_i \left(1 - Y^{(i)}\left(h(X^{(i)}) + b\right)\right)_+ + \lambda \|h\|_H^2$$

where

$$(z)_+ = \max(z, 0)\,.$$



– 証明

SVMの最適化問題

$$\min_{h,\xi} \|h\|_H^2 + C\sum_i \xi_i \qquad \text{subj. to } \xi_i \geq 1 - Y^{(i)}\left(h(X^{(i)}) + b\right),$$
$$\xi_i \geq 0.$$

$\Leftrightarrow$

$$\min_{h,\xi} \|h\|_H^2 + C\sum_i \xi_i \qquad \text{subj. to } \xi_i \geq \max\{1 - Y^{(i)}\left(h\left(X^{(i)}\right) + b\right), 0\}.$$

$\Leftrightarrow$

$$\min_{h,\xi} \|h\|_H^2 + C\sum_i \left(1 - Y^{(i)}\left(h\left(X^{(i)}\right) + b\right)\right)_+$$

$\Leftrightarrow$

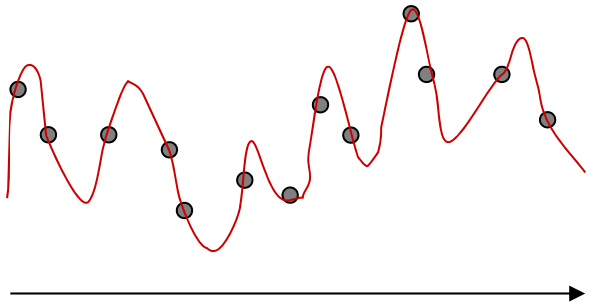
$$\min_{h} \sum_i \left(1 - Y^{(i)}\left(h\left(X^{(i)}\right) + b\right)\right)_+ + \lambda\|h\|_H^2 \qquad (\lambda = 1/C)$$

正則化

- 不良設定問題:

$$\min_f \sum_i \left(Y^{(i)} - f\left(X^{(i)}\right)\right)^2$$

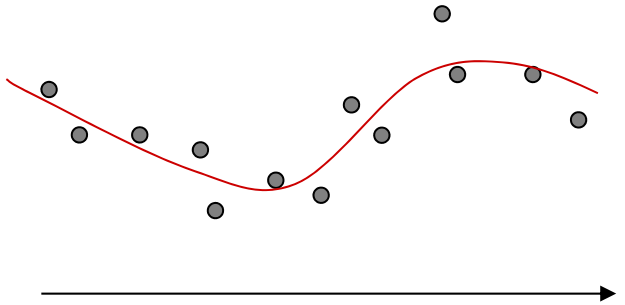
f が広い関数クラスから選ばれると
多くの f が誤差0を与える,



- 正則化

$$\min_f \sum_i \left(Y^{(i)} - f\left(X^{(i)}\right)\right)^2 + \lambda \|f\|_H{}^2$$

解が一意になる.
滑らかな関数のノルムが小さくなるように
正則化することが多い.



SVMのデモ

多くのソフトウエアが公開されている.

LibSVM etc

– JavaScriptの例

http://mine-weblog.blogspot.jp/2012/09/svm-demo.html

応用: 手書き数字認識

– MNIST: 手書き数字データベース
  - 28 x 28 ピクセルの2値画像.
  - 60000 訓練データ
  - 10000 テストデータ

Methods	Error rates
K-NN	5.0
10PCA + Quad	3.3
RBF network	3.6
LeNet 4	1.1
LeNet 5	0.95
SVM Poly 4	1.1
RS-SVM Poly 5	1.0

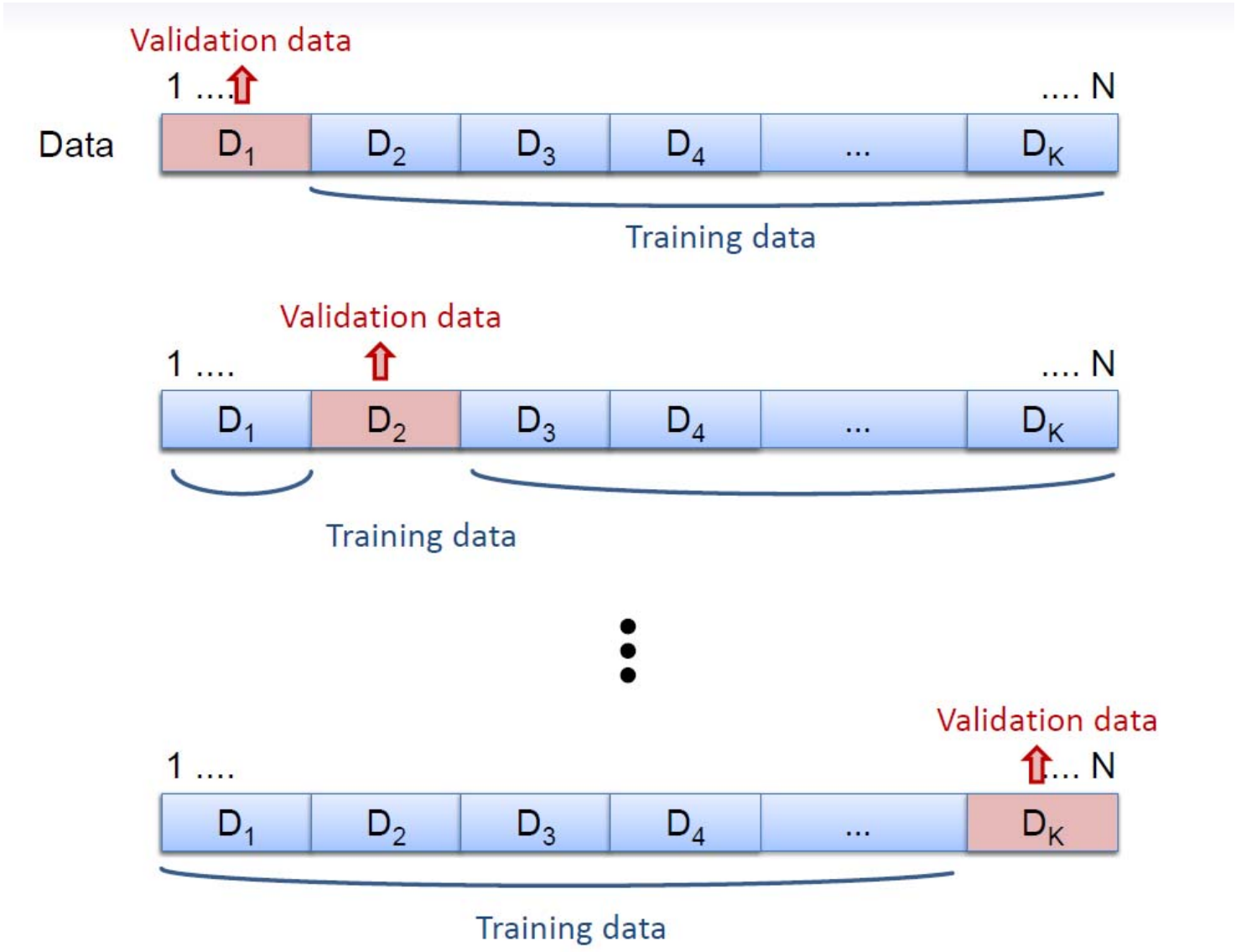
Y. LeCun et al. (2001) in *Intelligent Signal Processing*.

交差検証法Cross-validation

- クロスバリデーション (CV): 未知データに対する識別／予測の誤差を推定する手法.

- K-fold CV
  - データを(ランダムに) K個に分割する.
  - For $i = 1, \ldots, K$
    - 第iグループをテストデータに残し，他のデータで学習.
    - 第iグループに対するテスト誤差を測る.
  - K個の誤差を平均.

- Leave-one-out CV (LOOCV)
  - $K = N$. $i = 1, \ldots, N$に対し，第iデータ以外で学習し，第iデータに対する誤差を測る.
  - N個の誤差を平均.



K-fold cross-validation

SVMのまとめ

- マージン最大化

  ベイズ最適ではないが, 他のよい点を持っている.

- カーネル法

  非線形識別器への発展が容易.

- 2次計画:

  最適化問題は標準的なQP．ソフトウエアパッケージが利用可.

- スパース表現:

  識別器は，少数のサポートベクターにより表現される(Chap.4で説明)

- 正則化

  ソフトマージンSVMの目的関数は正則化として表現できる.

Representer定理

Representer定理

- カーネル法での最適化問題
  - カーネルリッジ回帰

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

  - SVM

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

  - カーネルPCA

$$\min_{f\in H} -\sum_{i=1}^{n} \left(f\left(X^i\right) - \frac{1}{n}\sum_{j=1}^{n} f\left(X^j\right)\right)^2 + \Omega(\|f\|_H)$$

$$\Omega(s) = \begin{cases} 0, & \text{if } s \in [0,1] \\ +\infty, & \text{if } s > 1 \end{cases}$$

- 解は

$$f = \sum_{i=1}^{n} c_i k\left(\cdot, X^i\right)$$

の形をとることをすでに見た。

■ 一般的定式化

- $\mathbf{X}_n = \{X^1, \ldots, X^n\} \subset \Omega_x$; $\mathbf{Y}_n = \{Y^1, \ldots, Y^n\} \subset \Omega_y$: データ
- $k_x: \Omega_x$ 上の正定値カーネル, $H: k_x$ の定めるRKHS
- $h_1(x), \ldots, h_M(x)$: 固定の関数（定数部分など）
- $\Psi(s): [0, \infty) \to \mathbf{R} \cup \{+\infty\}$: 正則化のための単調増加関数

一般的最適化問題

$$\min_{f \in H, a \in \mathbf{R}^M} L\left(\left\{f(X^i) + \sum_{m=1}^{M} a_m h_m(X^i)\right\}_{i=1}^{n}; \mathbf{X}_n, \mathbf{Y}_n\right) + \Psi(\|f\|_H)$$

定理2．1（Representer 定理）

上の仮定のもと，最小化問題の解は

$$f = \sum_{i=1}^{n} c_i k(\cdot, X^i)$$

の形で十分である．さらに Ψ が狭義の増加関数の場合，この形に限られる．

Representer定理により，関数空間での最適化が有限次元（データ数）の最適化に還元される．

– 証明

$H_0 := \mathrm{Span}\{k(\cdot, X^1), \ldots, k(\cdot, X^n)\}$, $H_\perp$ をその直交補空間として,

$$H = H_0 \oplus H_\perp$$

と直交分解する. この分解にしたがって, $f \in H$ を

$$f = f_0 + f_\perp$$

と分解すると, $\langle f_\perp, k(\cdot, X^i)\rangle = 0$ により, 目的関数 L の値は f を f_0 に置き換えても変化しない.

一方, $\|f_0\|_H \leq \|f\|_H$ より,

$$\Psi(\|f_0\|_H) \leq \Psi(\|f\|_H)$$

が成り立つ.

したがって, H_0 の元によって解が与えられる.

カーネルの選択

- カーネル法の結果の良否は，用いるカーネルに強く依存する.
  - カーネルの選択： ガウス，ラプラス，多項式？
  - パラメータの選択： e.g. $\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$ のバンド幅 σ

- 問題の性質に適したカーネルを用いる
  周波数特性，構造化データ，etc

- 教師あり学習 （SVM, カーネルリッジ回帰, etc）
  - Cross-validationが有効 （計算量が許せば）

- 教師なし学習 （カーネルPCAなど）
  - 標準的方法はない.
  - 関連する教師あり学習を作成すれば，cross-validationが使える.

– カーネルの学習

- Multiple Kernel Learning

  複数のカーネルの凸結合（線形結合）を考え，その係数も最適化する

$$k(x, y) = a_1 k_1(x, y) + \cdots + a_M k_M(x, y)$$

まとめ

カーネル法の一般的性質

- 効率的計算による非線形特徴の抽出

  元の空間が高次元でも，計算量が増大しない特別な特徴写像を用いる

- カーネルトリック

  正定値カーネルを特徴写像に用いることより，特徴ベクトルの内積が容易に計算できる.

- Representer定理

  殆どのカーネル法は，解が特徴ベクトルの線形和の形で表現可能．従って，データ数の個数のパラメータ最適化問題に変換される.

- グラム行列による計算

  カーネルトリックとRepresenter定理により，必要な計算はグラム行列の処理に還元される．従って，データ数に依存する計算量となる.

- 非ベクトルデータ

  カーネル法は，正定値カーネルが定義されれば，任意の集合に対して適用できる．→ 構造化データ解析（Chap. 5）

その他のカーネル法

- カーネルK-means クラスタリング
  特徴空間でK-meansクラスタリングを行う.
- カーネル Partial Least Square
  非線形回帰であるPLSのカーネル化
- カーネルロジスティック回帰
  ロジスティック回帰のカーネル化
- サポートベクター回帰
  回帰問題へのSVMの拡張
- 1ークラスサポートベクターマシン
  密度関数のレベル集合の推定へのSVMの拡張

etc.

References

福水 「カーネル法入門」 3章 朝倉書店 2010

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

Akaho. (2000) Kernel Canonical Correlation Analysis. *Proc. 3rd Workshop on Induction-based Information Sciences (IBIS2000).* (in Japanese)

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

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

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.

Fan, J. and I. Gijbels. *Local Polynomial Modelling and Its Applications*. Chapman Hall/CRC, 1996.

Boser, B.E., I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Fifth Annual ACM Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, PA, 1992. ACM Press.

Boser, B.E., I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Fifth Annual ACM Workshop on Computational Learning Theory*, pages 144–152, Pittsburgh, PA, 1992. ACM Press.

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

Cristianini, N., J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge Univ. Press. 2000

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.

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.

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

Crammer, K. and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

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.