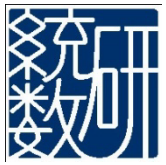
4. サポートベクターマシン

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

- サポートベクターマシンの最適化
  - 双対問題
  - サポートベクターマシンの双対問題とサポートベクター
  - Sequential Minimal Optimization (SMO)

- サポートベクターマシンの拡張
  - 多クラス識別問題

概要

- サポートベクターマシンの最適化
  - 双対問題
  - サポートベクターマシンの双対問題とサポートベクター
  - Sequential Minimal Optimization (SMO)

- サポートベクターマシンの拡張
  - 多クラス識別問題

サポートベクターマシンの最適化

- 主問題

$$\min_{w,b,\xi_i} \sum_{i,j=1}^{N} k(X_i, X_j) w_i w_j + C \sum_{i=1}^{N} \xi_i$$

$$\text{subj. to} \quad \begin{cases} Y_i \left(\sum_{j=1}^{N} w_j k(X_i, X_j) + b \right) \geq 1 - \xi_i \\ \xi_i \geqq 0 \end{cases}$$

- 双対問題

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \sum_{i,j=1}^{N} \alpha_i \alpha_j Y_i Y_j k(X_i, X_j)$$

$$\text{subj. to} \quad \begin{cases} 0 \leq \alpha_i \leq C \\ \sum_{i=1}^{N} Y_i \alpha_i = 0. \end{cases}$$

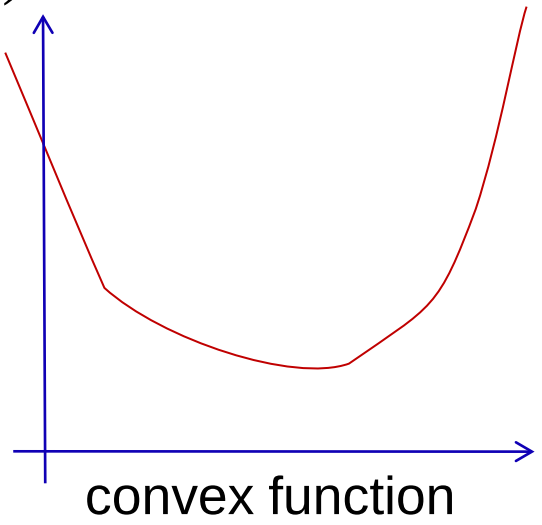
凸最適化の一般論

– 凸関数

$\mathbf{R}^m$の凸集合 C 上で定義された関数 f が凸（convex）であるとは，
任意の $x, y \in C$ と $t \in [0, 1]$ に対して

$$f(tx + (1-t)y) \le tf(x) + (1-t)f(y)$$

が成り立つことをいう.



– 凸最適化問題

D: $\mathbf{R}^m$の凸集合.

f, h_i $(i = 1,...,p)$: D 上の凸関数

$$\min f(x)$$

$$\text{subj. to} \begin{cases} h_i(x) \le 0 \quad (i = 1,..., p) & \text{不等式制約} \\ a_j{}^T x + b_j = 0 \quad (j = 1,..., q) & \text{等式制約} \end{cases}$$

- 凸最適化問題には局所最適解がない（最適解集合も凸）.

Lagrange双対問題

- 主問題

$$\min f(x) \quad \text{subj. to} \quad \begin{cases} h_i(x) \le 0 \quad (i = 1,\ldots, p) \\ r_j(x) = a_j{}^T x + b_j = 0 \quad (j = 1,\ldots, q) \end{cases}$$

- Lagrange関数

$$L(x;\lambda,\nu) \equiv f(x) + \sum_{i=1}^{p} \lambda_i h_i(x) + \sum_{j=1}^{q} \nu_j r_j(x).$$

$\lambda_i \geqq 0, \ \nu_j \in \mathbf{R}$: Lagrange乗数（双対問題の変数）

- 双対関数

$$g(\lambda,\nu) = \inf_{x \in D} L(x;\lambda,\nu)$$

制約無しの最適化

双対関数 g は凹（concave）関数

- **双対問題**

$$\max g(\lambda, \nu) \quad \text{subj. to} \quad \lambda \geq 0$$

- **双対性**

$$p^* = \inf f(x) \quad \text{subj. to} \quad \begin{cases} h_i(x) \leq 0 & (i = 1, \ldots, p) \\ r_j(x) = 0 & (j = 1, \ldots, q) \end{cases}$$

$$d^* = \sup g(\lambda, \nu) \quad \text{subj. to} \quad \lambda \geq 0$$

Theorem (strong duality)

主問題が凸最適化であるとし，ある制約想定（e.g. Slater条件）を満たすとする．このとき，双対問題の解 λ^*, ν^* が存在して

$$g(\lambda^*, \nu^*) = d^* = p^*.$$

– Slater条件： D の相対的内点 x_0 があって，主問題の等式制約と $h_i(x) < 0$ を満たす．

Karush-Kuhn-Tucker条件

- KKT条件

$$h_i(x^*) \le 0 \quad (i = 1,\ldots, p) \qquad \text{主不等式制約}$$

$$r_j(x^*) = 0 \quad (j = 1,\ldots, q) \qquad \text{主等式制約}$$

$$\lambda_i^* \ge 0 \quad (i = 1,\ldots, p) \qquad \text{双対不等式制約}$$

$$\lambda_i^* h_i(x^*) = 0 \quad (i = 1,\ldots, p) \qquad \text{相補性条件}$$

$$\nabla_x L(x^*, \lambda^*, \nu^*) = \nabla f(x^*) + \sum_{i=1}^{p} \lambda_i^* \nabla h_i(x^*) + \sum_{j=1}^{q} \nu_j^* \nabla r_j(x^*) = 0.$$

Theorem (KKT condition)

f, h_i が微分可能関数からなる凸最適化問題に対し，x^* と (λ^*, ν^*) がそれぞれ主，双対問題の解であることとKKT条件とは同値である.

サポートベクターマシンの双対問題

- SVMの主問題

$$\min_{w,b,\xi_i} \sum_{i,j=1}^{N} w_i w_j K_{ij} + C\sum_{i=1}^{N} \xi_i \qquad \text{subj. to} \quad \begin{cases} 1-\xi_i - Y_i\left(\sum_{j=1}^{N} w_j K_{ij} + b\right) \le 0 \\ -\xi_i \leqq 0 \end{cases}$$

$$\left(K_{ij} = k(X_i, X_j)\right)$$

- Lagrange関数

$$L(w,b,\xi;\alpha,\beta) = \sum_{i,j=1}^{N} w_i w_j K_{ij} + C\sum_{i=1}^{N} \xi_i + \sum_{i=1}^{N} \alpha_i \left(1-\xi_i - Y_i\left(\sum_{j=1}^{N} w_j K_{ij} + b\right)\right) + \sum_{i=1}^{N} \beta_i\left(-\xi_i\right)$$

- SVMの双対問題

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \sum_{i,j=1}^{N} \alpha_i \alpha_j Y_i Y_j K_{ij} \qquad \text{subj. to} \quad \begin{cases} 0 \le \alpha_i \le C \\ \sum_{i=1}^{N} Y_i \alpha_i = 0. \end{cases}$$

[Exercise] 双対問題を導出せよ. Hint:双対関数 $g(\alpha,\beta)$ を求めて β を消去せよ.

- SVMでは，主問題より双対問題のほうが解きやすい
  - 変数と制約の数が少ない
  - 1個の等式制約を除いては，すべて区間制約

- しかし，データ数Nが大きいと，一般的なQPソルバーでは時間がかかりすぎる.
  → 計算量を減らす工夫が必要
    - Chunking, Sequential Minimal Optimization (SMO)など（後述）.

SVMのKKT条件

KKT条件

① $1-\xi_i^*-Y_i\left(\sum_{j=1}^N w_j^* K_{ij}+b^*\right)\le 0$ 主不等式制約

② $-\xi_i^*\le 0$ 主不等式制約

③ $\alpha_i^*\ge 0$ 双対不等式制約

④ $\beta_i^*\ge 0$ 双対不等式制約

⑤ $\alpha_i^*\left\{1-\xi_i^*-Y_i\left(\sum_{j=1}^N w_j^* K_{ij}+b^*\right)\right\}=0$ 相補性条件

⑥ $\beta_i^*\xi_i^*=0$ 相補性条件

⑦ $\nabla_w:\quad \sum_{j=1}^N K_{ij}w_j^*-\sum_{j=1}^N K_{ij}Y_j\alpha_j^*=0$

$\nabla_b:\quad \sum_{i=1}^N Y_i\alpha_i^*=0.$

$\nabla_\xi:\quad C-\alpha_i^*-\beta_i^*=0$

SVMの最適解

SVMの最適解

$$f_*(x) = \sum_{i=1}^{N} \alpha_i^* Y_i k(x, X_i) + b^*$$

- KKT条件⑦を用いて w^* を α^* で表わす.
- b^* の求め方 → 後述

サポートベクター

- KKT条件の相補性条件

⑤ $\alpha_i^*\left(1-\xi_i^*-Y_i f_*(X_i)\right)=0,$

⑥ $(C-\alpha_i^*)\xi_i^*=0$

- サポートベクター
  - If $\alpha_i^*=0$, then $\xi_i^*=0$ and by primal constraints

$$Y_i f_*(X_i)\geq 1.$$ well-separated

  - If $0<\alpha_i^*<C$, then $\xi_i^*=0$ and from ⑤

$$Y_i f_*(X_i)=1.$$ マージン境界上　サポートベクター

  - If $\alpha_i^*=C$, then from ⑤

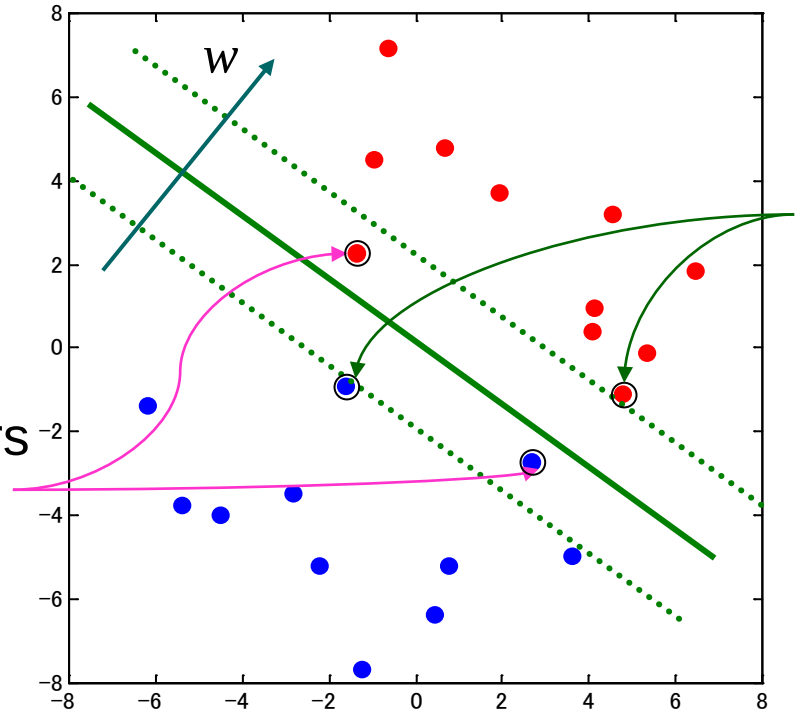
$$Y_i f_*(X_i)\leq 1.$$ マージン境界より誤識別側

サポートベクター

- サポートベクターによる最適解のスパース表現

$$f_*(x) = \sum \alpha_i^* Y_i k(x, X_i) + b^*$$

i: support vector



support vectors
$0 < \alpha_i < C$
$(Y_i f(X_i) = 1)$

support vectors
$\alpha_i = C$
$(Y_i f(X_i) \leq 1)$

定数項 b の求め方

- 相補性条件による b の解法
  - $0 < \alpha_i^* < C$ (マージン境界)を満たす任意の i に対し,

$$Y_i\left(\sum_{j=1}^{N} K_{ij} Y_j \alpha_j^* + b^*\right) = 1.$$

  - 例えば, このような全ての i に対して b の平均を求める.

計算量削減の工夫

- The dual QP problem of SVM has N variables, where N is the sample size.

- If N is very large, say N = 10000, the optimization is very hard.

- Some approaches have been proposed for optimizing subsets of the variables sequentially.
  - Chunking [Vapnik 1982]
  - Osuna's method [Osuna et al]
  - Sequential minimal optimization (SMO) [Platt 1999]
  - SVM^{light} (http://svmlight.joachims.org/)

Sequential Minimal Optimization (SMO)

– 方針：2変数(i, j)の組のみに対するQPをとき，これを繰り返す．

- ペアをどう選択するか? – KKT 条件が使える．
- w, ξ, β, を消去すると，SVMのKKT SVM 条件は以下と同値(Platt 1999, see also 福水2010)

$$\sum_{i=1}^{N} Y_i \alpha_i^* = 0 \quad \text{かつ} \quad (*) \begin{cases} \alpha_i^* = 0 \ \text{かつ} \ Y_i f_*(X_i) \geqq 1 \\ \text{or} \\ 0 < \alpha_i^* < C \ \text{かつ} \ Y_i f_*(X_i) = 1 \\ \text{or} \\ \alpha_i^* = C \ \text{かつ} \ Y_i f_*(X_i) \leqq 1 \end{cases}$$

$$Y_i f_*(X_i) = 1$$

- (＊)の条件は，データi毎にチェックすることが可能．
- (i, j) の少なくとも一方が(＊)を満足しないようにペアを選ぶ．

２変数(i, j) のQPは解析的に解ける!

- For simplicity, assume $(i, j) = (1, 2)$.
- Constraint of 1 and 2:

$$Y_1\alpha_1 + Y_2\alpha_2 = -\sum_{i=3}^{N} Y_i\alpha_i = \text{const.}$$

- Objective function:

$$\alpha_1 + \alpha_2 - \frac{1}{2}\alpha_1^2 K_{11} - \frac{1}{2}\alpha_2^2 K_{22} - Y_1 Y_2 \alpha_1 \alpha_2 K_{12} - \frac{Y_1}{2}\alpha_1 \sum_{i=3}^{N} Y_i \alpha_i K_{1i} - \frac{Y_2}{2}\alpha_2 \sum_{i=3}^{N} Y_i \alpha_i K_{2i} + \text{const.}$$

- 区間制約上の２次関数の最小化 → 解析解が求まる.

その他のSVM最適化へのアプローチ

- 主問題で解く
  - Chapelle（2007), SVM^{perf} (Joachims 2006), Shalev-Shwartz et al. (2007), etc.
- オンラインSVM.
  - Tax and Laskov (2003)
  - LaSVM [Buttou et al. 2005] http://leon.bottou.org/projects/lasvm/
- 並列計算
  - Cascade SVM [Graf et al. 2005]
  - Zanni et al (2006)
- Geometric approach
  - Mafrovorakis and Theodoridis (2006)
    （各クラスのデータ点のreduced convex hullによりSVMを解釈する）

SVMのソフトウエア

代表的なもの

- SVM^{light} http://svmlight.joachims.org/
- LIBSVM http://www.csie.ntu.edu.tw/~cjlin/libsvm/

SVMの特徴

- カーネル法のひとつ

  マージン最大化による線形識別機のカーネル化

- 高次元の特徴空間によるパターン認識

- 凸最適化

  2次計画．局所最適解の問題がない．現代的な凸最適化の機械学習への応用の機運となった．　*c.f.* ニューラルネットワーク

- スパース表現

  サポートベクターによる少数のデータによる解の表現

- 正則化

  スプライン平滑化などとの類似性

概要

- サポートベクターマシンの最適化
  - 双対問題
  - サポートベクターマシンの双対問題とサポートベクター
  - Sequential Minimal Optimization (SMO)

- サポートベクターマシンの拡張
  - 多クラス識別問題

多クラス識別

- Multiclass classification:
  - Data: $(X_1, Y_1), \ldots, (X_N, Y_N)$
    - X: 説明変数
    - $Y \in \{C_1, ..., C_L\}$: labels for L classes.
      - e.g. 10種の数字
  - Make a classifier: $h: \Omega \rightarrow \{C_1, ..., C_L\}$
- SVMの多クラス識別問題への適用
  - オリジナルのSVMは2値識別のみに適用可能.
  - Some approaches for extension to multiclass classification.
    - 多クラス識別におけるマージン最大化基準によるSVMの拡張.
      - Crammer and Singer (2001) など
      - LN変数の最適化が必要 → 計算量の削減が重要
    - 2値識別器の組み合わせ.

2値識別器の組み合わせ

多クラス問題を，2クラス識別問題に分解し，あとで結果を統合する.

- 1-vs-rest

  i-class vs the other classes : L problems

- 1-vs-1

  i-class vs j-class : L(L－1)/2 problems

- More general approach = Error correcting output code (ECOC, [DB95]). ECOC attributes a code for each class.

1-vs-rest の例

	f_1	f_2	f_3	f_4
C_1	+1	-1	-1	-1
C_2	-1	+1	-1	-1
C_3	-1	-1	+1	-1
C_4	-1	-1	-1	+1

- Hamming decoding：
  ベースの2値識別器の出力結果が符号と最も近いクラスに決定

セクション4のまとめ

- SVMの最適化
  - 双対問題を解くほうがよい.
  - KKT条件 → サポートベクターによるスパース表現
  - 大きいデータサイズでは，汎用QPソルバーでは計算困難
    - Sequential Minimal Optimization (SMO)など

- 他クラス識別問題への拡張
  - 2つのアプローチ
    - マージン最大化基準の他クラス拡張
    - 2値識別器の組み合わせ

参考文献

Bottou, L., O. Chapelle, D. DeCoste, and J. Weston. *Large-Scale Kernel Machines. MIT Press, 2007.*

Boyd, S. and L. Vandenberghe*. Convex Optimization*. Cambridge University Press, 2004. http://www.stanford.edu/ boyd/cvxbook/.

Chapelle, O. (2007) Training a support vector machine in the primal. *Neural Computation*, 19:1155–1178.

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

Dietterich, T.G. and G. Bakiri. (1995) Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286.

Graf, H.P., E. Cosatto, L. Bottou, I. Dourdanovic, and V. Vapnik. (2005) Parallel support vector machines: The Cascade SVM. *Advances in Neural Information Processing Systems, volume 17*. MIT Press.

Joachims,T . (2006*)* Training linear SVMs in linear time. In *Proc. ACM Conf. Knowledge Discovery and Data Mining (KDD).*

Mavroforakis, M.E. and S. Theodoridis. (2996) A geometric approach to support vector machine (SVM) classification. *IEEE Trans. Neural Networks*, 17(3), 2006.

Tax, D.M.J. and P. Laskov. (2003) Online svm learning: from classification to data description and back. *Proc. IEEE 13th Workshop on Neural Networks for Signal Processing (NNSP2003), 499–508.*

Zanni, L., T. Serafini, and G. Zanghirati. (2006) Parallel software for training large scale support vector machines on multiprocessor systems. *Journal of Machine Learning Research*, 7:1467–1492.