SMOのまとめ

azimuth-san

概要

SVMの学習アルゴリズムであるSMOについて，以下の文献を元に内容をまとめる．2値分類問題が対象である．

- Platt, J.　Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. 1998.

最適化問題の箇所では以下も参考にした．

- 山下信雄． 非線形計画法．
- 竹内一郎，鳥山昌幸． サポートベクトルマシン．

最適化問題

SVMの学習は以下の最適化問題として定式化される．これを主問題と呼ぶ．

- 主問題

$$
\begin{aligned}
\underset{\boldsymbol{w},\boldsymbol{\xi},b}{\text{minimize}} \quad & P(\boldsymbol{w},\boldsymbol{\xi},b) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\
\text{s.t.} \quad & -\left(y_i(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_i}) + b) - 1 + \xi_i\right) \leq 0, \\
& \xi_i \leq 0, \quad i \in \{1,\cdots,n\}
\end{aligned}
$$

  $\boldsymbol{w} \in \mathrm{R^d}, \boldsymbol{\xi} \in \mathrm{R^n}, \mathrm{b} \in \mathrm{R}$ は決定変数，$\boldsymbol{x}_i \in \mathrm{R^d}$ は特徴ベクトル，$y_i \in \{-1, 1\}$ はクラスラベル，$\boldsymbol{\phi} : \mathrm{R^d} \to D$ は $\boldsymbol{x_i}$ を新たな特徴空間へ写す写像である．

主問題に対しラグランジュ双対問題を導出することで以下が得られる．

- 双対問題

$$
\begin{aligned}
\underset{\boldsymbol{\alpha}}{\text{maximize}} \quad & W(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k(\boldsymbol{x}_i,\boldsymbol{x}_j) + \sum_{i=1}^{n}\alpha_i \\
\text{s.t.} \quad & \sum_{i=1}^{n}\alpha_i y_i = 0, \\
& 0 \leq \alpha_i \leq C, \quad i \in \{1,\cdots,n\}
\end{aligned}
$$

$\boldsymbol{\alpha} \in \mathrm{R}^{\mathrm{n}}$ はラグランジュ定数であり，双対問題に対する決定変数となる．k はカーネル関数であり $k(\boldsymbol{x_i}, \boldsymbol{x_j}) = \boldsymbol{\phi}(\boldsymbol{x_i})^T \boldsymbol{\phi}(\boldsymbol{x_j})$ と定義される．

上の双対問題の解 $\boldsymbol{\alpha}$を用いて，識別関数 f を表すことができる．

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i k(\boldsymbol{x}_i, \boldsymbol{x}) + b$$

$$b = y_j - \sum_{i=1}^{n} \alpha_i y_i k(\boldsymbol{x}_i, \boldsymbol{x}_j), \quad j \in \{1, \cdots, n \mid 0 < \alpha_j < C\}$$

SVMの主問題は，凸2次計画問題と呼ばれるクラスの最適化問題であり，Slaterの制約想定を満たす．この場合，以下が成り立つことが知られている．以下では主問題の決定変数 $(\boldsymbol{w}, \boldsymbol{\xi}, b)$ をまとめて $\boldsymbol{\theta}$ と表記する．

1. KKT条件が主問題と双対問題の最適性の必要十分条件となる．
2. 主問題の最適解 $\boldsymbol{\theta}^*$ と双対問題の最適解 $\boldsymbol{\alpha}^*$ において $P(\boldsymbol{\theta}^*) = W(\boldsymbol{\alpha}^*)$ が成り立つ．

2より，双対問題を解くことで主問題の最適値が得られる．1より，最適解を得るにはKKT条件を満たす点を求めれば良い． また識別関数 f は双対問題の最適解 $\boldsymbol{\alpha}^*$ を用いて表現できる．よって双対問題とKKT条件から $\boldsymbol{\alpha}^*$ を求めることで，識別関数 f を得ることができる．

KKT条件

- KKT条件

$$\begin{aligned} \alpha_i = 0 \ &\Rightarrow \ y_i f(\boldsymbol{x}_i) - 1 \geq 0 \quad (1) \\ 0 < \alpha_i < C \ &\Rightarrow \ y_i f(\boldsymbol{x}_i) - 1 = 0 \quad (2) \\ \alpha_i = C \ &\Rightarrow \ y_i f(\boldsymbol{x}_i) - 1 \leq 0 \quad (3) \end{aligned}$$

- KKT条件の否定

$$\begin{aligned} 0 \leq \alpha_i < C \ \wedge \ y_i f(\boldsymbol{x}_i) - 1 < 0 \\ 0 < \alpha_i \leq C \ \wedge \ y_i f(\boldsymbol{x}_i) - 1 > 0 \end{aligned}$$

$\alpha_i \in [0, \, C]$ が満たされている場合は（SMOではそのように α_i を更新していく），簡略化した以下の条件を用いればKKT条件に違反しているかが分かる．

$$\begin{aligned} \alpha_i < C \ \wedge \ y_i f(\boldsymbol{x}_i) - 1 < 0 \quad (4) \\ \alpha_i > 0 \ \wedge \ y_i f(\boldsymbol{x}_i) - 1 > 0 \quad (5) \end{aligned}$$

SMOではKKT条件に違反している双対変数 α_i を選択し，最適化問題の求解を行う．変数の選択には $(4), (5)$ 式が用いられる．

目的関数の最大化

SMOでは双対変数のうちの2変数を用いて目的関数を最大化することを繰り返す．2変数を用いる理由について説明する．双対問題の１つ目の制約 $\sum_i \alpha_i y_i = 0$ に注目すると，この中の2変数は

$$\alpha_1 y_1 + \alpha_2 y_2 = \text{const}$$

の関係にあるから，これを崩さないよう， α_1, α_2 を以下のように合わせて更新すれば，$\sum_i \alpha_i y_i = 0$ の関係を保持できるためである．

$$\alpha_1^{\text{new}} y_1 + \alpha_2^{\text{new}} y_2 = \alpha_1 y_1 + \alpha_2 y_2 = \text{const}$$

以降では変数を選択した後の目的関数の最大化について説明する．
まず，最適化の目的関数を，選択した変数 α_1, α_2 に関する項と，残りの変数に関する項に分ける．

$$W(\alpha_1, \alpha_2) = \alpha_1 + \alpha_2 - \frac{1}{2}K_{11}\alpha_1^2 - \frac{1}{2}K_{22}\alpha_2^2 - y_1 y_2 K_{12}\alpha_1\alpha_2 - y_1\alpha_1 v_1 - y_2\alpha_2 v_2 + \text{const}$$

$$K_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j), \quad v_i = \sum_{j=3}^{n} y_j \alpha_j K_{ij}$$

上の2次関数を α_2 について最大化することを考える．α_2 について微分し0とおくことで次式が得られる（詳細は参考文献の12.7節）．

$$\alpha_2^{\text{new}} = \alpha_2 - \frac{y_2(E_1 - E_2)}{\eta}, \quad (6)$$

$$E_l = f(\boldsymbol{x}_l) - y_l, \quad (7)$$

$$\eta = 2k(\boldsymbol{x}_1, \boldsymbol{x}_2), -k(\boldsymbol{x}_1, \boldsymbol{x}_1) - k(\boldsymbol{x}_2, \boldsymbol{x}_2) \quad (8)$$

η は目的関数を2階微分した項である．多くの場合 η は負となる（これは目的関数が上に凸であることを表す）．稀に η が負にならないことがある．例えば2つの入力ベクトル x_1, x_2 が同じ値を取る場合 η は 0 となる（目的関数が1次関数）．η が正となる場合もある（目的関数が下に凸）. これらの場合SMOは端点での目的関数値を評価し α_2^{new} を決定する．
α_1^{new} の更新式は，α_2^{new}に依存した形で得られる $\left(\because \alpha_1^{\text{new}} y_1 + \alpha_2^{\text{new}} y_2 = \alpha_1 y_1 + \alpha_2 y_2\right)$.

$$\alpha_1^{\text{new}} = \alpha_1 + y_1 y_2 (\alpha_2 - \alpha_2^{\text{new}})$$

但し，α_2^{new} をそのまま用いると，最適化問題の2つ目の制約 $0 \leq \alpha_1^{\text{new}} \leq C$ を満たさない可能性があるため，α_2^{new}に上下限値を設ける必要がある．

- $y_1 \neq y_2$ の場合
  $\alpha_1^{\text{new}} y_1 + \alpha_2^{\text{new}} y_2 = \alpha_1 y_1 + \alpha_2 y_2$ の両辺にy_1を掛けて，
  $\alpha_1^{\text{new}} = \alpha_1 + \alpha_2^{\text{new}} - \alpha_2$
  $0 \leq \alpha_1^{\text{new}} \leq C \;\Leftrightarrow\; 0 \leq \alpha_1 + \alpha_2^{\text{new}} - \alpha_2 \leq C \;\Leftrightarrow\; \alpha_2 - \alpha_1 \leq \alpha_2^{\text{new}} \leq C + \alpha_2 - \alpha_1$
  よって次のようにα_2を更新すれば良い．
  $L \leq \alpha_2^{\text{new}} \leq H,$
  $L = \max(0,\, \alpha_2 - \alpha_1), \;\; H = \min(C,\, C + \alpha_2 - \alpha_1)$

- $y_1 = y_2$ の場合
  同様にして以下の上下限値が得られる．
  $L \leq \alpha_2^{\text{new}} \leq H,$
  $L = \max(0,\ \alpha_1 + \alpha_2 - C\ ),\ \ H = \min(C,\ \alpha_1 + \alpha_2)$

よって，α_2^{new}についての次の更新式が得られる．

$$\alpha_2^{\text{new}} = \begin{cases} H & (\alpha_2^{\text{new}} \geq H) \\ \alpha_2^{\text{new}} & (L < \alpha_2^{\text{new}} \leq H) \\ L & (\alpha_2^{\text{new}} \leq L) \end{cases}$$

閾値 b と誤差 E の更新

α_2^{new} の更新では以下を計算する必要があった．

$$\alpha_2^{\text{new}} = \alpha_2 - \frac{y_2(E_1 - E_2)}{\eta}, \quad (6)$$

$$E_l = f(\boldsymbol{x}_l) - y_l, \quad (7)$$

$$\eta = 2k(\boldsymbol{x}_1, \boldsymbol{x}_2), -k(\boldsymbol{x}_1, \boldsymbol{x}_1) - k(\boldsymbol{x}_2, \boldsymbol{x}_2) \quad (8)$$

b 及び E_l が必要となる．これらは α_i に依存した量であるため， α_1, α_2 の更新後，合わせて更新する．

閾値 b の更新

- $0 < \alpha_1^{\text{new}} < C$ である場合
  閾値 b を以下のように求めることができる．
  KKT条件の (2) 式より
  $0 < \alpha_1^{\text{new}} < C \ \Rightarrow\ y_1 f(\boldsymbol{x}_1) - 1 = 0 \ \Leftrightarrow f(\boldsymbol{x}_1) - y_1 = 0$
  最後の等式をカーネル関数を用いた形で表すと
  $\sum_i \alpha_i y_i k(\boldsymbol{x}_i, \boldsymbol{x}_1) + b - y_1 = 0$
  となる．また(7) 式より上の左辺は E_1 を表す．
  α_1, α_2 を更新したことで $E_1 \to E_1 + \Delta E_1$ となったとする.
  $E_1 + \Delta E_1 \ = \ E_1 + y_1(\alpha_1^{\text{new}} - \alpha_1)k(x_1, x_1) + y_2(\alpha_2^{\text{new}} - \alpha_2)k(x_2, x_1) + b_1^{\text{new}} - b_1 \ = \ 0$
  よって以下の更新式が得られる．
  $b^{\text{new}} = b_1^{\text{new}} = -E_1 - y_1(\alpha_1^{\text{new}} - \alpha_1)k(x_1, x_1) - y_2(\alpha_2^{\text{new}} - \alpha_2)k(x_2, x_1) + b_1$

- $0 < \alpha_2^{\text{new}} < C$ である場合
  同様にして次の更新式が得られる．
  $b^{\text{new}} = b_2^{\text{new}} = -E_2 - y_1(\alpha_1^{\text{new}} - \alpha_1)k(x_1, x_2) - y_2(\alpha_2^{\text{new}} - \alpha_2)k(x_2, x_2) + b_2$

- $\alpha_1^{\text{new}}, \alpha_2^{\text{new}}$ が共に境界の 0 か C の値を取る場合
  $b^{\text{new}} = \frac{(b_1^{\text{new}} + b_2^{\text{new}})}{2}$ とおく．

誤差 E の更新

- E_k に対応する α_k が最適化の2変数として選ばれており，かつ境界にない $(0 < \alpha_k < C)$ 場合
  $E_k = 0$ と更新する．

- 上記以外の E_k
  $E_k = E_k + y_1(\alpha_1^{\text{new}} - \alpha_1)k(x_1, x_k) + y_2(\alpha_2^{\text{new}} - \alpha_2)k(x_2, x_k)$ と更新する．

2つの双対変数(ラグランジュ乗数)の選び方

最適化に用いる2つのラグランジュ乗数はヒューリスティックに選択する．方針を以下にまとめる．

- 1つ目の変数には，KKT条件に違反している α_i を選択する．
- 2つ目の変数には， α_2 に与える更新量 $\frac{y_2(E_1 - E_2)}{\eta}$ が最大となる α_i を選択する．η 内のカーネル関数の計算に時間がかかるため，$|E_1 - E_2|$で更新量の大きさを見積もる．E_1 が正であれば最小誤差 E_2 に対応する α_i が，E_1 が負であれば最大誤差 E_2 に対応する α_i を選択する．

- KKT条件に許容値 ϵ を設ける． $\epsilon = 10^{-2} \sim 10^{-3}$
  KKT条件に違反しているかの確認は，$(4), (5)$ 式に許容値 ϵ を設けた上で行う．
  これについて補足する．まず，KKT条件の $(1), (2)$ 式は以下の通りであった．
  $$\alpha_i = 0 \;\Rightarrow\; y_i f(\boldsymbol{x}_i) - 1 \geq 0$$
  $$0 < \alpha_i < C \;\Rightarrow\; y_i f(\boldsymbol{x}_i) - 1 = 0$$

  上の2つをまとめる．
  $$0 \leq \alpha_i < C \;\Rightarrow\; y_i f(\boldsymbol{x}_i) - 1 \geq 0$$

  許容値 ϵ を設ける．
  $$0 \leq \alpha_i < C \;\Rightarrow\; y_i f(\boldsymbol{x}_i) - 1 \geq -\epsilon$$

  KKT条件に違反している場合に真となるよう，上記の否定をとる．
  $$0 \leq \alpha_i < C \;\wedge\; y_i f(\boldsymbol{x}_i) - 1 < -\epsilon$$

  $\alpha_i \in [0,\ C]$ となるよう更新するから簡略化でき，(4) 式に許容値 ϵ を設けた以下が得られる．
  $$\alpha_i < C \;\wedge\; y_i f(\boldsymbol{x}_i) - 1 < -\epsilon \quad (4')$$

  同様の操作をKKT条件の $(2), (3)$ 式に対し行うことで，もう一つの条件式が得られる．
  $$\alpha_i > 0 \;\wedge\; y_i f(\boldsymbol{x}_i) - 1 > \epsilon \qquad (5')$$

- 高速化のため，変数選択と最適化処理は常時全ての双対変数を対象とするわけではない．
- 全変数に対するアルゴリズムの実行が一度終わった後は，$\alpha_i \neq 0, \neq C$ (non-bound examples) を満たす変数に限定してアルゴリズムを実行する．
- 全ての non-bound examples がKKT条件を満たせば，もう一度全変数を対象としてアルゴリズムを実行する．

以上