SMOのまとめ

azimuth-san

概要

SMOを用いたSVMの学習方法について，以下の文献を元に理解した内容をまとめる．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．

- 久野, 繁野, 後藤.　IT Text 数理最適化.
- 山下.　非線形計画法.

最適化問題

SVMの学習は最適化問題として定式化される．主問題となる最適化問題に対し，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．

$$\underset{\boldsymbol{\alpha}}{\text{maximize}} \quad W(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_{i=1}^{n}\alpha_i$$

$$\text{s.t.} \quad \sum_{i=1}^{n}\alpha_i y_i = 0,$$
$$0 \le \alpha_i \le C, \quad i \in \{1, \cdots, n\}$$

$\alpha_i \in \mathrm{R}$ は双対変数（ラグランジュ乗数），$\boldsymbol{x}_i \in \mathrm{R}^{\mathrm{d}}$ は特徴ベクトル，$y_i \in \{-1, 1\}$はクラスラベル，$k(\cdot,\cdot)$はカーネル関数を表す．双対問題の解 α_iを用いて，識別関数は以下のように表すことができる．

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

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．

1. KKT条件を満たすことが主問題と双対問題の最適性の必要十分条件となる．
2. 主問題に最適解が存在すれば双対問題にも解が存在し，最適値が一致する．

1より，双対問題の制約を満たし，かつKKT条件を満たす点を求めることができれば，それが双対問題の最適解となる．2より，双対問題の最適値は主問題の最適値に等しい．また，識別関数 f は双対変数を用いて表現できる．よって最適解となる双対変数を代入した f は主問題の最適値を与える，つまりマージンを最大化する識別

関数となる．

KKT条件

- KKT条件

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

- 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．

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

  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．変数の選択には上の2つの条件式が用いられる．

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．最適化問題の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

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

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．

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

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．

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

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

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）．

$$\alpha_2^{\mathrm{new}} = \alpha_2 - \frac{y_2(E_1 - E_2)}{\eta}, \quad E_l = f(\boldsymbol{x}_l) - y_l$$
$$\eta = 2k(\boldsymbol{x}_1, \boldsymbol{x}_2), -k(\boldsymbol{x}_1, \boldsymbol{x}_1) - k(\boldsymbol{x}_2, \boldsymbol{x}_2)$$

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 $(\because \alpha_1^{\mathrm{new}} y_1 + \alpha_2^{\mathrm{new}} y_2 = \alpha_1 y_1 + \alpha_2 y_2)$.

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

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

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

よって，α_2^{new} についての次の更新式が得られる．

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

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．

$$\alpha_2^{\mathrm{new}} = \alpha_2 - \frac{y_2(E_1 - E_2)}{\eta},\ \ E_l = f(\boldsymbol{x}_l) - y_l,$$
$$f(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i y_i k(\boldsymbol{x}_i, \boldsymbol{x}) + b$$

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

- $0 < \alpha_1^{\mathrm{new}} < C$ 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条件より
  $0 < \alpha_1^{\mathrm{new}} < C \;\Rightarrow\; y_1 f(\boldsymbol{x}_1) - 1 = 0 \;\Leftrightarrow f(\boldsymbol{x}_1) - y_1 = 0$
  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．上式の左辺を E_1 と定義していたことを思い出す．
  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．
  $E_1 + \Delta E_1 \;=\; E_1 + y_1(\alpha_1^{\mathrm{new}} - \alpha_1)k(x_1, x_1) + y_2(\alpha_2^{\mathrm{new}} - \alpha_2)k(x_2, x_1) + b_1^{\mathrm{new}} - b_1 \;=\; 0$
  よって以下の更新式が得られる．
  $b^{\mathrm{new}} = b_1^{\mathrm{new}} = -E_1 - y_1(\alpha_1^{\mathrm{new}} - \alpha_1)k(x_1, x_1) - y_2(\alpha_2^{\mathrm{new}} - \alpha_2)k(x_2, x_1) + b_1$

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

- $\alpha_1^{\mathrm{new}}, \alpha_2^{\mathrm{new}}$ が共に境界の 0 か C の値を取る場合
  $b^{\mathrm{new}} = \frac{(b_1^{\mathrm{new}} + b_2^{\mathrm{new}})}{2}$ 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
  $E_k = E_k + y_1(\alpha_1^{\mathrm{new}} - \alpha_1)k(x_1, x_k) + y_2(\alpha_2^{\mathrm{new}} - \alpha_2)k(x_2, x_k)$ 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ばれる．

- 1つ目の変数には，KKT条件に違反している α_i が選ばれる．
- 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 が選ばれる．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 が選ばれる．

- KKT条件の確認には許容値 ϵ が設けられる． $\epsilon = 10^{-2} \sim 10^{-3}$
  許容値を設けた場合のKKT条件について説明する．KKT条件の内の2つの条件式は以下で表された．
  $\alpha_i = 0 \;\;\Rightarrow\;\; y_i f(\boldsymbol{x}_i) - 1 \geq 0$
  $0 < \alpha_i < C \;\Rightarrow\; y_i f(\boldsymbol{x}_i) - 1 = 0$

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

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

KKT条件に違反しているかを知るには，上記の否定をとって
$0 \leq \alpha_i < C \ \wedge \ y_i f(\boldsymbol{x}_i) - 1 < -\epsilon$

$\alpha_i \in [0, \ C]$ となるよう更新するから簡略化できて，以下が得られる．
$\alpha_i < C \ \wedge \ y_i f(\boldsymbol{x}_i) - 1 < -\epsilon$

同様の手順で，違反しているかを知るための，もう一つの条件式が得られる．
$\alpha_i > 0 \ \wedge \ y_i f(\boldsymbol{x}_i) - 1 > \epsilon$

- 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は実行される．
- 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が実行される．

以上