ARTICLE TITLE

JOHN SMITH* & JAMES SMITH[1]

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

* *Department of Biology, University of Examples, London, United Kingdom*

[1] *Department of Chemistry, University of Examples, London, United Kingdom*

1 INTRODUCTION

My current plan is to write a survey about the screening methods used in identifying non-support vector in solving Support Vector Machine (SVM) problem. Such methods differ from the screening strategy for lasso in that, in SVM the data points are discarded while in lasso the features are discarded.

1.1 Margin

Given a hyperplane $\mathcal{H} = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{w} \cdot \mathbf{x} + b = 0\}$, the distance from a point $\mathbf{z} \in \mathbb{R}^d$ to $\mathcal{H}$ is $\mathbf{D}(\mathbf{z}, \mathcal{H}) = \frac{\mathbf{w} \cdot \mathbf{z} + b}{\|\mathbf{w}\|}$. If we know whether the point is above or below the hyperplane by the sign variable $y \in \{-1, +1\}$, we define the margin to be $\mathbf{M}(\mathbf{z}, \mathcal{H}) = y\frac{\mathbf{w} \cdot \mathbf{z} + b}{\|\mathbf{w}\|}$

Now suppose we are given a dataset $\mathcal{X} = \{y_i, \mathbf{x}_i\}_{i=1}^m$, the margin of $\mathcal{X}$ to $\mathcal{H}$ is defined as

$$\mathbf{M}(\mathcal{X}, \mathcal{H}) = \min_{\mathbf{z} \in \mathcal{X}} \mathbf{M}(\mathbf{z}, \mathcal{H}).$$

Given that the dataset $\mathcal{X}$ is seperable, SVM aims to seek a hyperplane $\mathcal{H}^*$ that gives the largest margin of $\mathcal{X}$ to $\mathcal{H}^*$. Such goal can be formalized as

$$\begin{aligned} \max_{\mathcal{H}} \quad & \rho \\ \text{s.t.} \quad & \mathbf{M}(\mathbf{z}, \mathcal{H}) \geqslant \rho, \forall \mathbf{z} \in \mathcal{X} \end{aligned}$$

Clearly we can scale $\mathbf{w}$ and b such that $\min_{\mathbf{z} \in \mathcal{X}} |\mathbf{z} \cdot \mathbf{w} + b| = 1$, thus the above formulation is equivalent to

$$\begin{aligned} \min \quad & \|\mathbf{w}\| \\ \text{s.t.} \quad & y_i(\mathbf{x}_i \cdot \mathbf{w} + b) \geqslant 1, i = 1, \dots, m \end{aligned} \tag{1}$$

2 SVM: NON-SEPARABLE CASE

We are interested in solving SVM problem where there is no hyperplane that can correctly classify all the data points. There are several formulations of SVM whose primal and dual problems, KKT conditions, and corresponding screening properties will be discussed one by one.

2.1 Formation I

The primal formulation of such problem is

$$\begin{aligned} \min_{\mathbf{w}, b, \boldsymbol{\xi}} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^m \xi_i^\beta \\ \text{s.t.} \quad & y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geqslant 1 - \xi_i \cap \xi_i \geqslant 0, i \in [m]. \end{aligned} \tag{P_β^I}$$

where $\beta = 1$ or 2, corresponding to l_1 and l_2 SVM variants.

It is clear that when $\beta = 1$, the l_1-SVM problem is equivalent to the following problem.

$$\min_{\mathbf{w}, b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^m [1 - y_i(\mathbf{w} \cdot \mathbf{x}_i + b)]_+ \tag{2}$$

Using the standard derivation, we have the following dual formulation of (P^{I}_{β}) when $\beta = 1$

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \|\boldsymbol{\alpha}\|_1 - \frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{Q}\boldsymbol{\alpha} \\ \text{s.t.} \quad & 0 \leqslant \alpha_i \leqslant C \cap \boldsymbol{\alpha} \cdot \mathbf{y} = 0, i \in [m], \end{aligned} \tag{D^{I}_{1}}$$

where $\mathbf{Q}_{i,j} = y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$. Given a primal solution $\mathbf{w}^*, \boldsymbol{\xi}^*, b^*$ and a dual solution $\boldsymbol{\alpha}^*$, the KKT condition states that

1. $\xi_i^* \geqslant 0, i = 1, \dots, m.$
2. $\xi_i^* \geqslant 1 - y_i(\mathbf{w}^* \cdot \mathbf{x}_i + b^*), i = 1, \dots, m$
3. $0 \leqslant \alpha_i^* \leqslant C, i = 1, \dots, m$
4. $\alpha_i^*(1 - y_i(\mathbf{w}^* \cdot \mathbf{x}_i + b^*) - \xi_i^*) = 0, i = 1, \dots, m$
5. $(C - \alpha_i^*)\xi_i^* = 0, i = 1, \dots, m$
6. $\mathbf{w}^* = \mathbf{A}\boldsymbol{\alpha}^*$

where $\mathbf{A} = [\cdots y_i \mathbf{x}_i \cdots]_i$.

If $1 - y_i(\mathbf{w}^* \cdot \mathbf{x}_i + b^*) < 0$ then $(1 - y_i(\mathbf{w}^* \cdot \mathbf{x}_i + b^*) - \xi_i^*) < 0$, we have $\alpha_i^* = 0$ by the complementary slackness. Similarly, if $1 - y_i(\mathbf{w}^* \cdot \mathbf{x}_i + b^*) > 0$, we have $\alpha_i^* = C$. By categorizing the m training instances into three sets

- $\mathcal{R} := \{i \in [m] | y_i(\mathbf{w}^* \cdot \mathbf{x}_i + b) > 1\}$
- $\mathcal{E} := \{i \in [m] | y_i(\mathbf{w}^* \cdot \mathbf{x}_i + b) = 1\}$
- $\mathcal{L} := \{i \in [m] | y_i(\mathbf{w}^* \cdot \mathbf{x}_i + b) < 1\}$

we summarize such property by that

- $i \in \mathcal{R} \to \alpha_i = 0$
- $i \in \mathcal{E} \to \alpha_i \in [0, C]$
- $i \in \mathcal{L} \to \alpha_i = C$

2.2 Formulation II

Another way to write SVM is by

$$\begin{aligned} \min_{\mathbf{w}, b, \boldsymbol{\xi}} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 \\ \text{s.t.} \quad & y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geqslant 1 - \xi_i \cap \xi_i \geqslant 0, i \in [m] \\ & \sum_{i=1}^{m} \xi_i^\beta \leqslant s, \end{aligned} \tag{P^{II}_{β}}$$

which is clearly equivalent to the following problem when we take $\beta = 1$

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{s.t.} \quad & \sum_{i=1}^{m} [1 - y_i(\mathbf{w} \cdot \mathbf{x}_i + b)]_+ \leqslant s. \end{aligned} \tag{3}$$

where $[a]_+ = \max\{a, 0\}$ denotes the hinge loss.

Using the standard derivation, we have the following dual formulation of (P_β^{II}) when $\beta = 1$

$$\begin{aligned} \max_{\boldsymbol{\alpha},C} \quad & \|\boldsymbol{\alpha}\|_1 - \frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{Q}\boldsymbol{\alpha} - Cs \\ \text{s.t.} \quad & 0 \leqslant \alpha_i \leqslant C \cap \boldsymbol{\alpha} \cdot \mathbf{y} = 0, i \in [m], \end{aligned} \tag{D_1^{II}}$$

Given a primal solution $\mathbf{w}^*, \boldsymbol{\xi}^*, b^*$ and a dual solution $\boldsymbol{\alpha}^*, C^*$, the KKT condition states that

1. $\xi_i^* \geqslant 0, i = 1, \dots, m.$
2. $\xi_i^* \geqslant 1 - y_i(\mathbf{w}^* \cdot \mathbf{x}_i + b^*), i = 1, \dots, m$
3. $\sum_{i=1}^m \xi_i^* \leqslant s$
4. $0 \leqslant \alpha_i^* \leqslant C^*, i = 1, \dots, m$
5. $C^* \geqslant 0$
6. $(\sum_{i=1}^m \xi_i^* - s)C^* = 0$
7. $\alpha_i^*(1 - y_i(\mathbf{w}^* \cdot \mathbf{x}_i + b^*) - \xi_i^*) = 0, i = 1, \dots, m$
8. $(C^* - \alpha_i^*)\xi_i^* = 0, i = 1, \dots, m$
9. $\mathbf{w}^* = \mathbf{A}\boldsymbol{\alpha}^*$

where $\mathbf{A} = [\cdots y_i \mathbf{x}_i \cdots]_i$.

I have not yet figure out the screening properties in this formulation.

2.3 With Kernel

3 METHODS

4 RESULTS AND DISCUSSION

REFERENCES