Non-Euclidean Support Vector Classifiers for Sparse Learning

Ying Lin (林颖)
Joint work with and Qi Ye (叶颀)

South China Normal University

CSIAM 2020 Student Forum

Outline

Machine Learning

- Given

$$D := \left\{ (\boldsymbol{x}_i, y_i) \right\}_{i=1}^{N} \subseteq \left(\boldsymbol{X} \times \{\pm 1\} \right)^N \subseteq \left(\mathbb{R}^n \times \{\pm 1\} \right)^N,$$

  our goal is to find a linear function $\boldsymbol{x}^\top \boldsymbol{\omega} + b$ such that for any i, $y_i(\boldsymbol{x}_i^\top \boldsymbol{\omega} + b) \geq 0$, where $\boldsymbol{\omega} \in \mathbb{R}^n$, $b \in \mathbb{R}$.
- Given a special *loss function* $L : \mathbb{R} \times \mathbb{R} \to [0, +\infty]$, it is done by solving

$$\min_{\boldsymbol{\omega} \in \mathbb{R}^n, b \in \mathbb{R}} \sum_{i=1}^{N} L(y_i, \boldsymbol{x}_i^\top \boldsymbol{\omega} + b).$$

- But it is an ill-posed problem.

Regularization and Sparsity

- Regularization is a crucial technique to prevent machine learning algorithms from over-fitting, it has the general form

$$\min_{\boldsymbol{\omega}\in\mathbb{R}^n, b\in\mathbb{R}} \sum_{i=1}^{N} L(y_i, \boldsymbol{x}_i^\top \boldsymbol{\omega} + b) + \lambda\|\boldsymbol{\omega}\|_2^2.$$

- The final target of regularization is to obtain a *sparse* result, meaning that as many of the components of the parameter have values of 0 as possible. It is always done by 1-norm regularization.
- Understanding regularization and sparsity can help us to dive deep into learning theorems. There are many aspects to explore them, such as *functional analysis, convex analysis, statistical learning*, etc..

Support Vector Machine Classifiers

- *Support Vector Machine Classifier (SVM classifier)* is by far one of the most successful binary-classification methods, it can finally be represented by

$$\min_{\boldsymbol{\omega}\in\mathbb{R}^n, b\in\mathbb{R}} \sum_{i=1}^{N}[1-y_i(\boldsymbol{x}_i^\top\boldsymbol{\omega}+b)]_+ + \lambda\|\boldsymbol{\omega}\|_2^2.$$

- The Euclidean distance used by the classical SVM classifier leads to 2-norm regularization.

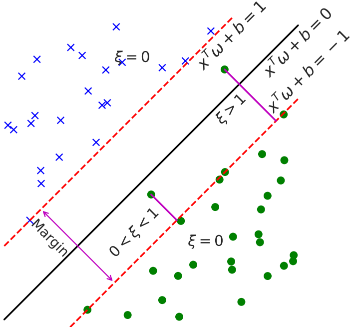


Figure *Support Vector Machine Classifiers*

Outline

Distances between Points and Hyperplanes

- Based on Theorem 2.2 in (O. L. Mangasarian, 1999), the distance derived from a general norm $\| \cdot \|$ from any points to a hyperplane

$$P := \{\boldsymbol{x} : \boldsymbol{x}^\top \boldsymbol{\omega} + b = 0 \, , \boldsymbol{x} \in \mathbb{R}^n\}$$

is given by

$$\text{dist}(\boldsymbol{x}, P) = \frac{|\boldsymbol{x}^\top \omega + b|}{\|\boldsymbol{\omega}\|_*},$$

where $\| \cdot \|_*$, defined as $\|\boldsymbol{z}\|_* = \sup\{\boldsymbol{z}^\top \boldsymbol{x} : \|\boldsymbol{x}\| < 1\}$, is the dual norm of $\| \cdot \|$.

Distances between Points and Hyperplanes: Examples

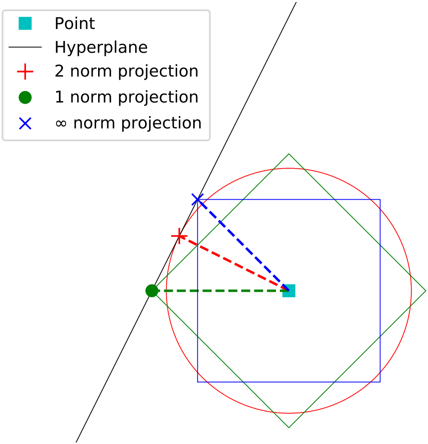


Figure Distances derived from special norms

- For 2 norm $\| \cdot \|_2$,

$$\mathrm{dist}(\boldsymbol{x}, P) = \frac{\|\boldsymbol{x}^\top \boldsymbol{\omega} + b\|}{\|\boldsymbol{\omega}\|_2}.$$

- For ∞ norm $\| \cdot \|_\infty$,

$$\mathrm{dist}(\boldsymbol{x}, P) = \frac{\|\boldsymbol{x}^\top \boldsymbol{\omega} + b\|}{\|\boldsymbol{\omega}\|_1}.$$

Non-Euclidean Support Vector Machine Classifiers

- The non-Euclidean SVM classifier has the form

$$
\begin{array}{lll}
\min\limits_{\boldsymbol{\omega}, b, \boldsymbol{\xi}} & \|\boldsymbol{\omega}\|_* + C\sum\limits_{i=1}^{N} \xi_i & \\
\text{subject to} & y_i(\boldsymbol{x}_i^T \boldsymbol{\omega} + b) \geq 1 - \xi_i, & \forall i, \\
& \xi_i \geq 0, & \forall i.
\end{array} \tag{1}
$$

  where C is the Lagrangian multiplier, $\boldsymbol{\xi} = \{\xi_i\}_{i=1}^{N}$ are slack variables.
- It is equivalent to the following unconstrained optimization problem

$$
\min_{\boldsymbol{\omega}, b} \underbrace{\sum_{i=1}^{N} [1 - y_i(\boldsymbol{x}_i^T \boldsymbol{\omega} + b)]_+}_{\text{Hinge Loss}} + \underbrace{\lambda \|\boldsymbol{\omega}\|_*}_{\text{Arbitrary Norm Regularization}} . \tag{2}
$$

Solutions for Optimization Problems

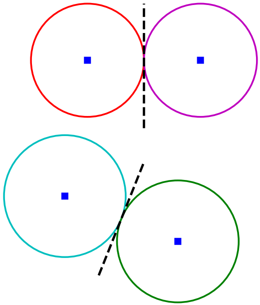
Proposition (Existences and Conditional Uniqueness of Solutions for Optimization Problems)

1. *Solution sets for constrained optimization problems (1), unconstrained optimization problem (2) are non-empty, compact and convex.*
2. *If $\|\cdot\|_*$ is strictly convex, constrained optimization problems (1), unconstrained optimization problem (2) have one and only one solution.*

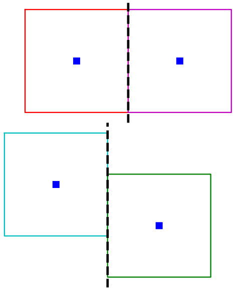
Theorem (Equivalence Between Solution Sets for Constrained and Unconstrained Optimization Problems)

1. *Suppose $\boldsymbol{v}^* = (\boldsymbol{\omega}^*, b^*, \boldsymbol{\xi}^*)$ is one of the solutions of (1), then $\boldsymbol{v}^* = (\boldsymbol{\omega}^*, b^*)$ is also one of the solutions of (2).*
2. *Suppose $\boldsymbol{v}^* = (\boldsymbol{\omega}^*, b^*)$ is one of the solutions of (2), then $\boldsymbol{v}^* = (\boldsymbol{\omega}^*, b^*, \boldsymbol{\xi}^*)$ is also one of the solutions of (1).*

Sparsity of 1-norm SVM Classifiers I



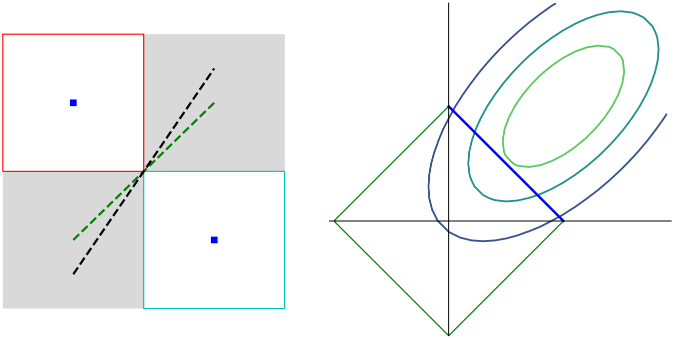
(a) 2 norm.



(b) ∞ norm.

Figure Maximal margin classifiers by 2 norm and ∞ norm in the case of 2 distinct points.

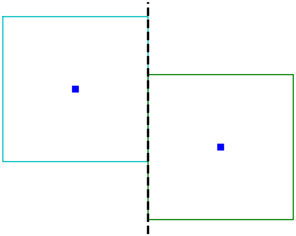
Sparsity of 1-norm SVM Classifiers II



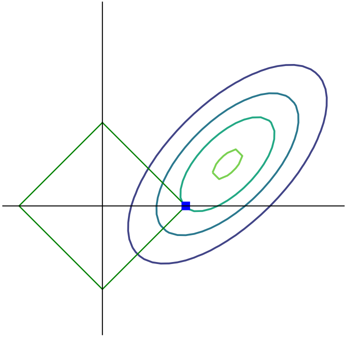
(a) Infinite solutions in data space. (b) Infinite solutions in parameter space.

Figure The geometric explanation for infinite solutions for the SVM classifier by ∞-norm margin.

Sparsity of 1-norm SVM Classifiers III



(a) Unique solution in data space.



(b) Unique solution in parameter space.

Figure The geometric explanation for unique solution for the SVM classifier by ∞-norm margin.

Outline

Summary

1. We extended the classical SVM classifiers to help us understand 1-norm regularization and provide a new view to study sparse learning.
2. We supplemented the mathematical backgrounds of 1-norm SVM classifiers.
3. We presented several special examples to show the sparsity by the 1-norm SVM classifiers.

Thank You!

国家自然科学基金委员会
National Natural Science Foundation of China
NSFC

广东省科学技术厅
Department of Science and Technology of Guangdong Province

广东省教育厅
DEPARTMENT OF EDUCATION OF GUANGDONG PROVINCE

1933
華南師範大學
SOUTH CHINA NORMAL UNIVERSITY