Non-Euclidean Support Vector Classifiers for Sparse Learning

Ying Lin (林颖), Qi Ye (叶颀)

South China Normal University

CSIAM 2020

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

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

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

- 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

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

where $[\cdot]_+ = \max\{0, \cdot\}$.

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

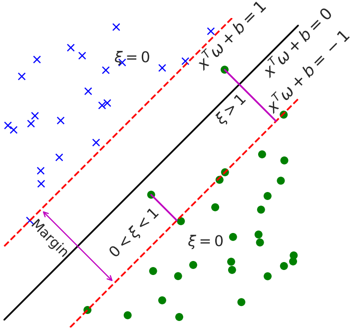


Figure *Support Vector Machine Classifiers*

Kernel-based Learning Methods

- Reproducing Kernel Hilbert Spaces (RKHSs) and Reproducing Kernel Banach Spaces (RKBSs) have been viewed as ideal spaces for *kernel-based learning methods*.
- For example, given a kernel function $K : \boldsymbol{X} \times \boldsymbol{X} \to \mathbb{C}$, there exists a unique RKHS $\mathcal{H}_K$ related to K equipped the norm $\|f\|_{\mathcal{H}_K}$, the learning task on $\mathcal{H}_K$ is

$$\min_{f \in \mathcal{H}_K} \sum_{i=1}^{N} L(y_i, f(\boldsymbol{x}_i)) + \lambda \|f\|_{\mathcal{H}_K}^2,$$

  whose solution has the form $f_h = \sum_{i=1}^{N} c_i K(x_i, \cdot)$ by several celebrated representer theorems.

Motivation and Basic Ideas

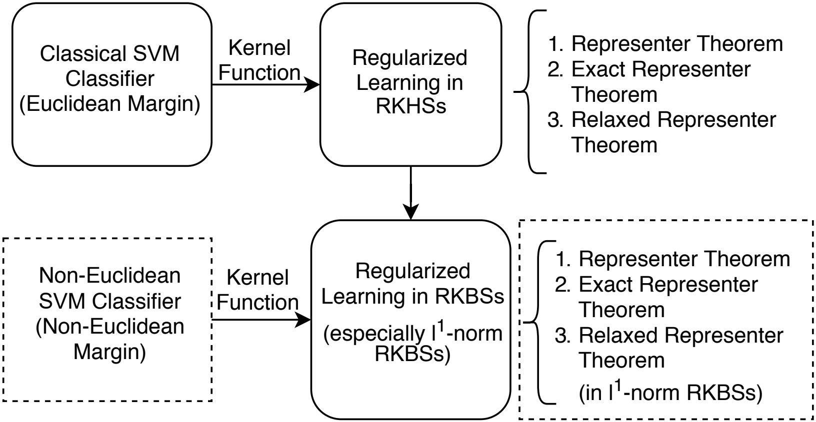


Figure Motivation and Basic Ideas

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

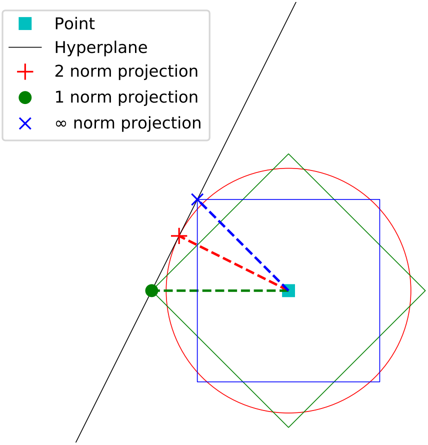


Figure Distances derived from special norms

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

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

- For ∞ norm $\|\cdot\|_\infty$,

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

Non-Euclidean Support Vector Machine Classifiers

- The non-Euclidean SVM classifier has the form

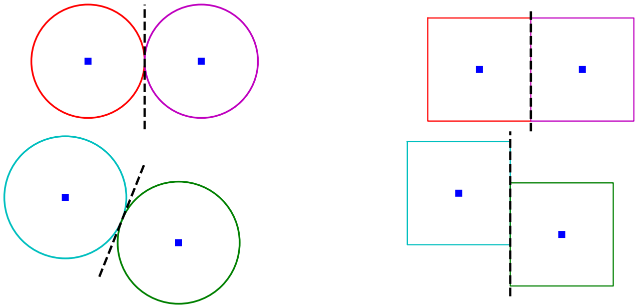
$$\begin{array}{lll} \min\limits_{\boldsymbol{\omega}, b, \boldsymbol{\xi}} & \|\boldsymbol{\omega}\|_* + C\sum\limits_{i=1}^{N} \xi_i & \\ \text{subject to} & y_i(\mathbf{x}_i^T \boldsymbol{\omega} + b) \geq 1 - \xi_i, & \forall i, \\ & \xi_i \geq 0, & \forall i. \end{array} \tag{1}$$

where C is the "cost" parameter, $\boldsymbol{\xi} = \{\xi_i\}_{i=1}^{N}$ are slack variables.

- It is equivalent to the following unconstrained optimization problem

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

Sparsity of 1-norm SVM Classifiers I

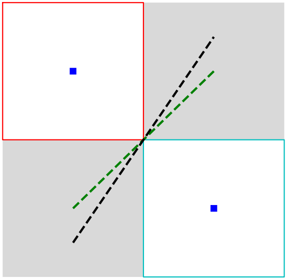


(a) 2 norm.

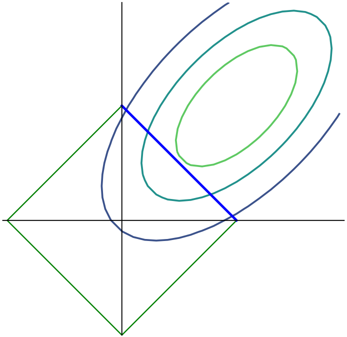
(b) ∞ norm.

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

Sparsity of 1-norm SVM Classifiers II



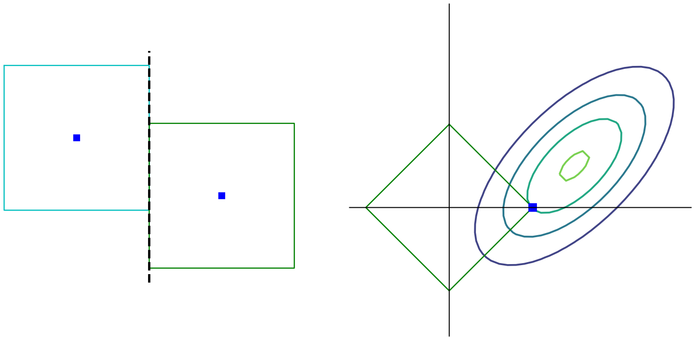
(a) Infinite solutions in data space.



(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.

Special Examples: Tensor Form of Non-Euclidean SVM Classifiers I

- Consider m norm where m is an even number, the primal Lagrange function of (1) is

$$\begin{aligned}\mathcal{L}_P = \quad & \frac{m-1}{m}\|\boldsymbol{\omega}\|_{\frac{m}{m-1}}^{\frac{m}{m-1}} + C\sum_{i=1}^{N}\xi_i \\ & - \sum_{i=1}^{N}\alpha_i[y_i(\boldsymbol{x}_i^\top\boldsymbol{\omega} + b) - (1-\xi_i)] - \sum_{i=1}^{N}\mu_i\xi_i,\end{aligned}$$

  where $\alpha_i \geq 0, \mu_i \geq 0, i = 1, 2, \ldots, N$ are Lagrange multipliers.

- The Lagrange (Wolfe) dual function is

$$\mathcal{L}_D = \sum_{i=1}^{N}\alpha_i - \frac{1}{m}\sum_{i_1, i_2, \ldots, i_m = 1}^{N, N, \ldots, N}\prod_{k=1}^{m}\Big(\alpha_{i_k}y_{i_k}(\sum_{j=1}^{d}\prod_{k=1}^{m}x_{i_k,j})\Big),$$

  where $x_{i_k,j}$ is the jth element of $\boldsymbol{x}_{i_k}$.

Special Examples: Tensor Form of Non-Euclidean SVM Classifiers II

- Let $\mathbf{1} = (1)_{i=1}^{N}$ and

$$\boldsymbol{A}_m := \Big(\prod_{k=1}^{m} y_{i_k}\big(\sum_{j=1}^{d}\prod_{k=1}^{m} x_{i_k,j}\big)\Big)_{i_1,i_2,\ldots,i_m=1}^{N,N,\ldots,N},$$

  which is an mth order Nth dimension tensor.

- The optimization problem (1) where $\|\cdot\|_*$ is $\frac{m}{m-1}$ norm can be solved by alternatively solving the following tensor-form optimization problem

$$\min_{\boldsymbol{\alpha}\in[0,\infty)^N} \mathbf{1}^\top\boldsymbol{\alpha} - \frac{1}{m}\boldsymbol{A}_m\boldsymbol{\alpha}^m,$$

  where $\boldsymbol{A}_m\boldsymbol{\alpha}^m$ is the m-mode product.

Special Examples: Tensor Kernel Functions

- If we use the basis functions $\boldsymbol{h}$ to obtain the nonlinear function $\boldsymbol{h}(\boldsymbol{x})^\top \boldsymbol{\omega} + b$, $\mathcal{L}_D$ has the form

$$\mathcal{L}_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{m} \sum_{i_1, i_2, \ldots, i_m = 1}^{N, N, \ldots, N} \prod_{k=1}^{m} \Big(\alpha_{i_k} y_{i_k} \big(\sum_{j=1}^{n} \prod_{k=1}^{m} h_j(\boldsymbol{x}_{i_k}) \big) \Big).$$

- Combining with $\mathcal{L}_P$, we can write the nonlinear function as

$$\boldsymbol{h}(\boldsymbol{x})^\top \boldsymbol{\omega} + b = \sum_{i_2, i_3, \ldots, i_m}^{N, N, \ldots, N} \prod_{k=2}^{m} \Big(\alpha_{i_k} y_{i_k} \big(\sum_{j=1}^{n} \prod_{k=2}^{m} h_j(\boldsymbol{x}_{i_k}) h_j(\boldsymbol{x}) \big) \Big) + b.$$

- Letting

$$K_m(\boldsymbol{x}, \boldsymbol{x}_{i_2}, \ldots, \boldsymbol{x}_{i_m}) := \sum_{j=1}^{n} \prod_{k=2}^{m} h_j(\boldsymbol{x}_{i_k}) h_j(\boldsymbol{x}),$$

  we can obtain the mth order *tensor kernel function*.

Outline

Regularization Networks in ℓ^1-norm RKBSs

- Given an admissible kernel K on $\boldsymbol{X}$, the related RKBS $\mathcal{B}_K$ is defined by

$$\mathcal{B}_K := \left\{ \sum_{t \in \mathrm{supp}\, c} c_t K(t, \cdot) : c \in \ell^1(X) \right\}$$

  with the norm $\left\| \sum_{t \in \mathrm{supp}\, c} c_t K(t, \cdot) \right\|_{\mathcal{B}_K} := \|c\|_{\ell^1(X)}$, where for any nonempty set X, we denote

$$\ell^1(X) := \left\{ c = (c_t \in \mathbb{R} : t \in X) : \|c\|_{\ell^1(X)} := \sum_{t \in \mathrm{supp}\, c} |c_t| < +\infty \right\}.$$

- The regularization network in $\mathcal{B}_K$ is

$$\min_{f \in \mathcal{B}_K} \sum_{i=1}^{N} L(y_i, f(\mathbf{x}_i)) + \lambda \|f\|_{\mathcal{B}_K}, \tag{3}$$

  whose solution is of form $f_b = \sum_{i=1}^{N} c_i K(\mathbf{x}_i, \cdot)$ by representer theorems.

Sparse Representer Theorems

Theorem

Let $K : \boldsymbol{X} \times \boldsymbol{X} \to \mathbb{R}$ *be an admissible kernel. If* f_b *is an extreme point of the solution set of regularization network (3) in* $\mathcal{B}_K$*, then* f_b *is of form*

$$f_b(\boldsymbol{x}) = \sum_{k=1}^{M} c_k K(\boldsymbol{z}_k, \boldsymbol{x}), \ \ \boldsymbol{x} \in \boldsymbol{X}$$

where $M \leq N$*,* $\boldsymbol{z}_k \in \boldsymbol{X}$ *and* $c_k \in \mathbb{R}, k = 1, 2, \ldots, M$*.*

Proof tips:

1. Transfer (3) to an equivalent minimal norm interpolation problem.
2. Use Theorem 3.1 in (Boyer et al. 2019) to show the special form of f_b.
3. Use special properties of extreme points of a ball in $\mathcal{B}_K$ to obtain the final form.

Sparse Representer Theorems

Theorem

Let $K : \boldsymbol{X} \times \boldsymbol{X} \to \mathbb{R}$ be an admissible kernel. If f_b is an extreme point of the solution set of regularization network (3) in $\mathcal{B}_K$, then f_b is of form

$$f_b(\boldsymbol{x}) = \sum_{k=1}^{M} c_k K(\mathbf{z}_k, \boldsymbol{x}), \ \ \boldsymbol{x} \in \boldsymbol{X}$$

where $M \leq N$, $\mathbf{z}_k \in \boldsymbol{X}$ and $c_k \in \mathbb{R}, k = 1, 2, \ldots, M$.

Proof tips:

1. Transfer (3) to an equivalent minimal norm interpolation problem.
2. Use Theorem 3.1 in (Boyer et al. 2019) to show the special form of f_b.
3. Use special properties of extreme points of a ball in $\mathcal{B}_K$ to obtain the final form.

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 and ℓ^1-norm RKBSs.
3. We presented several special examples to show the potential of the generalization.
4. We proposed a sparse representer theorem to show the power of sparse learning in ℓ^1-norm RKBSs.

Thank You!

国家自然科学
基金委员会
National Natural Science
Foundation of China
N S F C

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