Tutorial 5: Support Vector Machine

Rick Lin

The Chinese University of Hongkong, Shenzhen

October 17, 2022

Contents

Mathematical likbez

Coordinate notation

For inner product and vector norm:

- $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n : \langle \boldsymbol{a}, \boldsymbol{b} \rangle = \sum_{i=1}^{n} a_i b_i = \boldsymbol{a}^T \boldsymbol{b}$
- norm: $\|\boldsymbol{a}\| = \sqrt{\langle \boldsymbol{a}, \boldsymbol{a} \rangle} = \sqrt{\boldsymbol{a}^T \boldsymbol{a}}$

For fixed $\boldsymbol{w} \in \mathbb{R}^n, b \in \mathbb{R}$:

- linear hyperplane equation: $\boldsymbol{w}^T \boldsymbol{x} = 0, \boldsymbol{w}$ is a normal vector
- affine hyperplane equation: $\boldsymbol{w}^T \boldsymbol{x} - b = 0$
- projection of vector $\boldsymbol{x} \in \mathbb{R}^n$ onto a line, directed by $\boldsymbol{w}$: normalize $\boldsymbol{w}$ taking $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$ and take scalar product $\left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right)^T \boldsymbol{x}$, hence the projection vector is $\left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right)^T \boldsymbol{x} \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$

Linear separable case

Margin concept

class - class → point - class → point - point → x^2 optimization → dual opt

tricky

Margin is a a characteristic, evaluating how deep an object is "immersed" in its class, how typical a representative of the class it is. The smaller is the value of the margin, the closer the object is to the boundary of classes (and the higher the probability of error becomes). The value of margin is negative if and only if the algorithm makes an error on the object.

Definition. For the linear classifier $a(\boldsymbol{x}) = \operatorname{Sgn}\left(\boldsymbol{w}^T\boldsymbol{x} - b\right)$ hard margin $M_i(\boldsymbol{w}, b)$ of the object $(\boldsymbol{x}_i, y_i)$ is defined as

label

$$M_i(\boldsymbol{w}, b) = y_i \cdot \left(\boldsymbol{w}^T\boldsymbol{x}_i - b\right)$$

Hard-margin Minimization

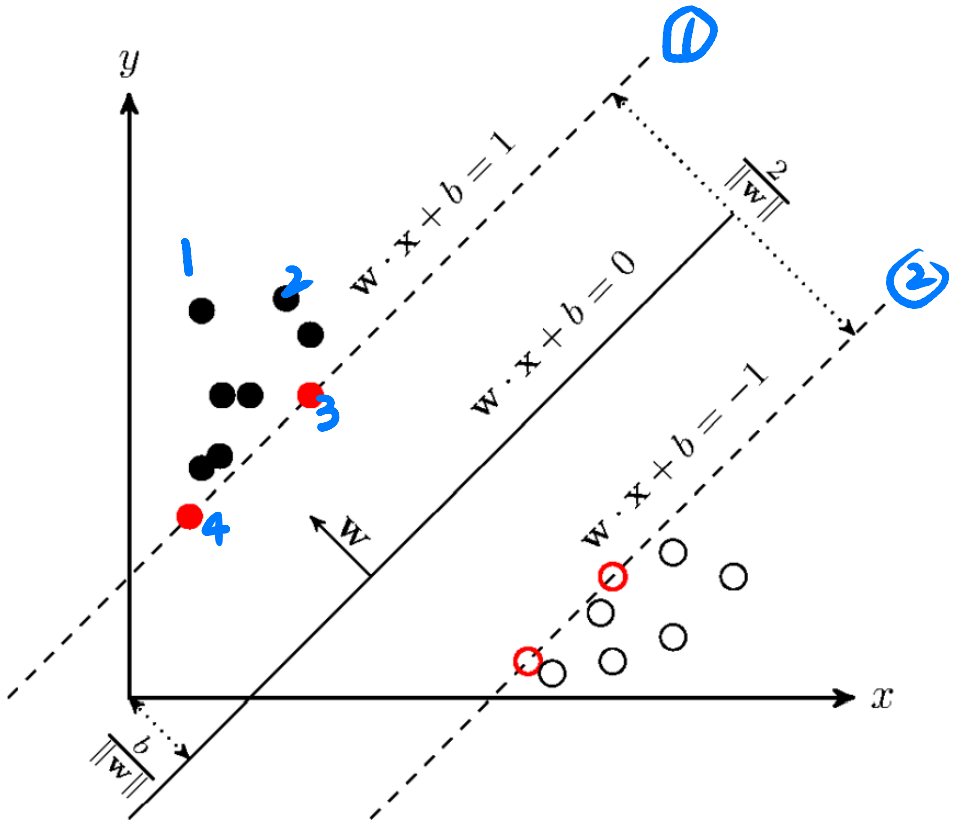


Figure: Support Hyperlanes

Support vectors for hard-margin

class - class → point - class → point – point → x^2 optimization → dual opt

- Hyperplane's equation $\boldsymbol{w}^T\boldsymbol{x} - b = 0$ is homo-genius: $\forall c \neq 0$ an equation $(c\boldsymbol{w})^T\boldsymbol{x} - cb = 0$ defines the same hyperplane
- Choose such c that $\min_i M_i(c\boldsymbol{w}, cb) = 1$
- Each class contains at least one object, minimizing the margin:

$$\exists x_i, x_j : \quad y_i = 1, y_j = -1, M_i(\boldsymbol{w}, b) = M_j(\boldsymbol{w}, b) = 1$$

  x_i and x_j are further referred as $x_+$and x_-respectively
- $x_+$and $x_-$are also called support vectors, colored in red in the preceding figure.

Minimization of the distance

class - class → point - class → point − point → x^2 optimization → dual opt

Distance between hyperplanes could be computed using projection onto the hyperplane normal vector:

$$\frac{\boldsymbol{w}^T\left(\boldsymbol{x}_+ - \boldsymbol{x}_-\right)}{\|\boldsymbol{w}\|} = \frac{\boldsymbol{w}^T\boldsymbol{x}_+ - \boldsymbol{w}^T\boldsymbol{x}_-}{\|\boldsymbol{w}\|} = \frac{(b+1)-(b-1)}{\|\boldsymbol{w}\|} = \frac{2}{\|\boldsymbol{w}\|}.$$

Since we maximize the distance between hyperplanes, we should minimize

$$\frac{1}{2}\|\boldsymbol{w}\|^2.$$

Recall, that no points are between hyperplanes, which means

$$y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i - b\right) \geq 1, \quad \forall i = 1, 2, \ldots, l$$

Quadratic programming problem

class - class ⟶ point - class ⟶ point — point ⟶ x^2 optimization ⟶ dual opt

To sum up, the problem is

$$\begin{aligned} &\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 \\ &\text{s.t. } y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i - b\right) \geq 1 \end{aligned}$$

which is a quadratic programming problem.

KKT conditions

class - class ⟶ point - class ⟶ point — point ⟶ x^2 optimization ⟶ dual opt

Recall Karush-Kuhn-Takker theorem: any optimal regular point of the problem

$$\begin{cases} \min_{\boldsymbol{x}} f(x) & \\ h_i(\boldsymbol{x}) = 0 & i = 1, 2, \ldots, m \\ g_j(\boldsymbol{x}) \leq 0 & j = 1, 2, \ldots, k \end{cases}$$

satisfies the following condition: there are such $\lambda_1, \ldots, \lambda_m, \mu_1, \ldots, m_k$ that for the function $L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i h_i(\boldsymbol{x}) + \sum_{j=1}^{k} \mu_j g_j(\boldsymbol{x})$ holds

$$\begin{cases} \frac{\partial}{\partial x_l}(L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})) = 0 & l = 1, 2, \ldots, n \\ h_i(\boldsymbol{x}) = 0 & i = 1, 2, \ldots, m \\ g_j(\boldsymbol{x}) \leq 0 & j = 1, 2, \ldots, k \\ \mu_j \geq 0 & j = 1, 2, \ldots, k \\ \mu_j g_j(\boldsymbol{x}) = 0 & j = 1, 2, \ldots, k \end{cases}$$

KKT conditions for SVM

class - class ⟶ point - class ⟶ point − point ⟶ x^2 optimization ⟶ dual opt

KKT conditions for the optimal regular point above: since $L(\boldsymbol{w}, b, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{m} \mu_i \left(1 - y_i \left(\boldsymbol{w}^T \boldsymbol{x}_i - b\right)\right)$ we get

$$\begin{cases} 0 = \nabla_{\boldsymbol{w}} L(\boldsymbol{x}, b, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \boldsymbol{w} - \sum_{i=1}^{n} \mu_i y_i \boldsymbol{x}_i & \text{Stationary} \\ 0 = \frac{\partial}{\partial b} L(\boldsymbol{x}, b, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \sum_{i=1}^{n} \mu_i y_i & \text{Stationary} \\ \mu_i \geq 0 & i = 1, 2, \ldots, l \quad \text{Feasibility} \\ \mu_i \left(1 - y_i \left(\boldsymbol{w}^T \boldsymbol{x}_i - b\right)\right) = 0 & i = 1, 2, \ldots, l \quad \text{Complementary Slackness} \end{cases}$$

Closed form

Using the previous result, we reduce the problem to follows:

$$\begin{cases} \min_{\boldsymbol{\lambda}} -\sum_{i=1}^{l} \lambda_i + \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \lambda_i \lambda_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \\ \sum_{i=1}^{l} \lambda_i y_i = 0 \end{cases} \quad i = 1, 2, \ldots, l$$

After solving those equations, the corresponding classifier parameters could be expressed as follows:

$$\begin{cases} \boldsymbol{w} = \sum_{i=1}^{l} \lambda_i y_i \boldsymbol{x}_i \\ b = \boldsymbol{w}^T \boldsymbol{x}_i - y_i \quad \forall i : \lambda_i > 0 \end{cases}$$

and the classifier itself:

$$a(\boldsymbol{x}) = \operatorname{Sgn}\left(\sum_{i=1}^{l} \lambda_i y_i \boldsymbol{x}_i^T \boldsymbol{x} - b \right)$$

Exercises

Q1: If the data is not linearly separable, then there is no solution to the hard-margin SVM.

a) True

b) False

Q2: The geometric margin in a hard-margin Support Vector Machine is

a) $\frac{\|w\|^2}{2}$

b) $\frac{1}{\|w\|^2}$

c) $\frac{2}{\|w\|}$

d) $\frac{2}{\|w\|^2}$

Exercises

Q3: Suppose we train a hard-margin linear SVM on $\mathrm{n} > 100$ data points in $\mathrm{R}2$, yielding a hyperplane with exactly 2 support vectors. If we add one more data point and retrain the classifier, what is the maximum possible number of support vectors for the new hyperplane (assuming the $n+1$ points are linearly separable)?

a) 2

b) 3

c) n

d) $n+1$

Exercises

Q4: For a hard margin SVM, give an expression to calculate b given the solutions for w and the Lagrange multipliers $\{\alpha_i\}_{i=1}^{N}$.

Exercises

Q4: For a hard margin SVM, give an expression to calculate b given the solutions for w and the Lagrange multipliers $\{\alpha_i\}_{i=1}^N$.

Using the KKT conditions $\alpha_i \left(y_i \left(w^T x_i + b\right) - 1\right) = 0$, we know that for support vectors, $\alpha_i \geq 0$. Thus for some $\alpha_i \geq 0, y_i \left(w^T x_i + b\right) = 1$ and thus

$$b = y_i - w^T x_i$$

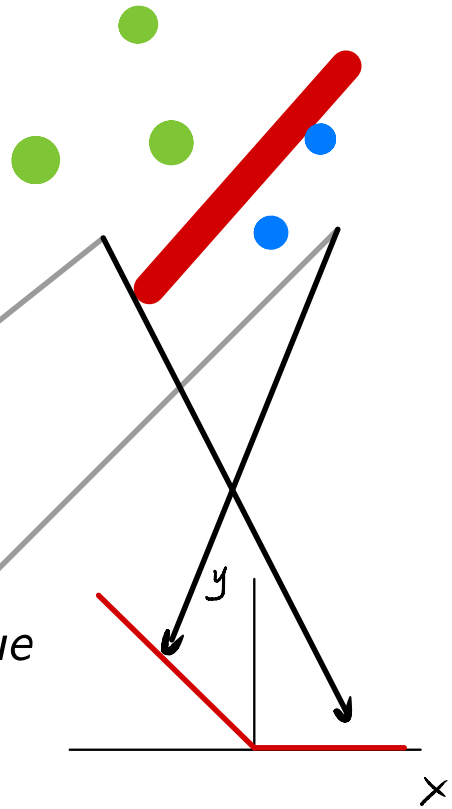
For numerical stability, we can take an average over all the support vectors.

$$b = \sum_{x_i \in S_v} \frac{y_i - w^T x_i}{|S_v|}$$

Non-linearly separable case

Soft margin

- Sometimes the hard separating hyperplane doesn't exist
- We need somehow redefine the concept of margin
- We only want punishment but do not want bonus.



Definition. *Hinge-loss function of i^{th} object $l_i(\boldsymbol{w}, b)$ is a value*

$$l_i(\boldsymbol{w}, b) = \max\left(0, 1 - y_i\left(\boldsymbol{w}^T \boldsymbol{x}_i - b\right)\right)$$

Reformulation in terms of soft-margin

We now add a penalty to objective function, which is a soft-margin with some weight. Introducing additional variable vector $\boldsymbol{\xi} \in \mathbb{R}_+^l$ we get

$$\begin{cases} \min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{l} \xi_i & \\ \xi_i \geq 1 - y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i - b\right) & i = 1, 2, \ldots, l \\ \xi_i \geq 0 & i = 1, 2, \ldots, l \end{cases}$$

KKT conditions

KKT conditions for the optimal regular point above: since $L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\eta}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\lambda_i\left(y_i\left(\boldsymbol{w}^T\boldsymbol{x} - b\right) + \xi_i - 1\right) - \sum_{i=1}^{l}\eta_i\xi_i$ we get

$$
\begin{cases}
0 = \nabla_{\boldsymbol{w}} L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\eta}) = \boldsymbol{w} - \sum_{i=1}^{l}\lambda_i y_i \boldsymbol{x}_i & \Longrightarrow \quad \boldsymbol{w} = \sum_{i=1}^{l}\lambda_i y_i \boldsymbol{x}_i \\
0 = \frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\eta})}{\partial b} = -\sum_{i=1}^{l}\lambda_i y_i & \Longrightarrow \quad \sum_{i=1}^{l}\lambda_i y_i = 0 \\
0 = \nabla_{\boldsymbol{\xi}} L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\eta}) = -\lambda_i - \eta_i + C & \Longrightarrow \quad \eta_i + \lambda_i = C \\
\lambda_i, \eta_i, \xi_i \geq 0 & i = 1, 2, \ldots, l \\
\lambda_i\left(y_i\left(\boldsymbol{w}^T\boldsymbol{x} - b\right) - 1 + \xi_i\right) = \eta_i\xi_i = 0 & i = 1, 2, \ldots, l
\end{cases}
$$

Closed form

Using the previous result, we reduce the problem to follows:

$$\begin{cases} \min_{\boldsymbol{\lambda}} -\sum_{i=1}^{l} \lambda_i + \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \lambda_i \lambda_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \\ 0 \leq \lambda_i \leq C \\ \sum_{i=1}^{l} \lambda_i y_i = 0 \end{cases} \qquad i = 1, 2, \ldots, l$$

After solving those equations, the corresponding classifier parameters could be expressed as follows:

$$\begin{cases} \boldsymbol{w} = \sum_{i=1}^{l} \lambda_i y_i \boldsymbol{x}_i \\ b = \boldsymbol{w}^T \boldsymbol{x}_i - y_i \quad \forall i : \lambda_i > 0 \end{cases}$$

and the classifier itself:

$$a(\boldsymbol{x}) = \operatorname{Sgn}\left(\sum_{i=1}^{l} \lambda_i y_i \boldsymbol{x}_i^T \boldsymbol{x} - b \right)$$

Exercises

Q1: Choose the correct statement(s) about Support Vector Machines (SVMs).

a) If a finite set of training points from two classes is linearly separable, a hard-margin SVM will always find a decision boundary correctly classifying every training point.

b) If a finite set of training points from two classes is linearly separable, a soft-margin SVM will always find a decision boundary correctly classifying every training point.

c) Every trained two-class hard-margin SVM model has at least one point of each class at a distance of exactly $1/\|w\|$ (the margin width) from the decision boundary.

d) Every trained two-class soft-margin SVM model has at least one point of each class at a distance of exactly $1/\|w\|$ (the margin width) from the decision boundary.

Exercises

Q2: Which of the following changes would commonly cause an SVM's margin $1/\|w\|$ to shrink?

a) Soft margin SVM: decreasing the value of C

b) Soft margin SVM: increasing the value of C

c) Hard margin SVM: adding a sample point that violates the margin

d) Hard margin SVM: adding a new feature to each sample point

Q3: In a soft-margin support vector machine, if we increase C, which of the following are likely to happen?

a) The margin will grow wider

b) Most nonzero slack variables will shrink

c) There will be more points inside the margin

d) The norm $|w|$ will grow larger

Exercises

Q4: The soft margin SVM formulation is as follows:

$$\min \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N}\xi_i$$

$$\text{subject to} \quad y_i\left(\boldsymbol{w}^T\boldsymbol{x_i} + b\right) \geq 1 - \xi_i \quad \forall i$$

$$\xi_i \geq 0 \quad \forall i$$

What is the behavior of the width of the margin $\left(\frac{2}{\|w\|}\right)$ as $C \to 0$?

a) Behaves like hard margin

b) Goes to zero

c) None of the above

d) Goes to infinity

Exercises

Q5: For the dual version of soft margin SVM, the λ_i's for support vectors satisfy $\lambda_i > C$.

a) True

b) False

Kernel

Kernel function

Now cost function: $f = \langle \boldsymbol{w}, \psi(x) \rangle - b$. Vector $\boldsymbol{w}$ is as earlier a linear combination of training dataset vectors, but now represented by $\psi(\boldsymbol{x}_i)$. So,

$$\boldsymbol{w} = \sum_{i=1}^{l} \lambda_i y_i \psi(\boldsymbol{x}_i)$$

However, we need to redefine the scalar product.

Definition. Kernel function $K(\boldsymbol{x}, \boldsymbol{y})$ is a scalar product in H :

$$K(\boldsymbol{x}, \boldsymbol{y}) = \langle \psi(\boldsymbol{x}), \psi(\boldsymbol{y}) \rangle_H.$$

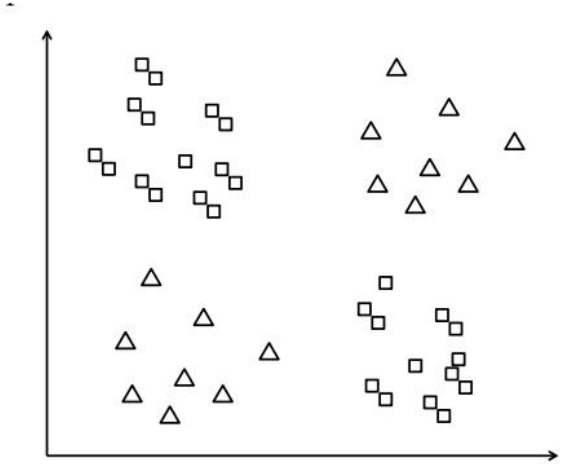
SVM with kernel

$$\begin{cases} \min -\sum_{i=1}^{l} \lambda_i + \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \lambda_i \lambda_j y_i y_j K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) \\ 0 \leq \lambda_i \leq C \\ \sum_{i=1}^{l} \lambda_i y_i = 0 \end{cases} \qquad i = 1, 2, \ldots, l$$

and the corresponding classifier is $\operatorname{Sgn}\left(\sum_{i=1}^{l} K\left(\boldsymbol{x}_i, \boldsymbol{x}\right) - b\right)$

Exercises

Q1: Given the following data samples (square and triangle mean two classes), which one(s) of the following kernels can we use in SVM to separate the two classes?



a) Linear kernel

b) Gaussian RBF (radial basis function) kernel

c) Polynomial kernel

d) None of the above

Exercises

Q2: The polynomial kernel is defined to be

$$k(\boldsymbol{x}, \boldsymbol{y}) = \left(\boldsymbol{x}^T \boldsymbol{y} + c\right)^d$$

where $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, and $c \geq 0$. When we take $d = 2$, this kernel is called the quadratic kernel. Find the feature mapping $\Phi(\boldsymbol{z})$ that corresponds to the quadratic kernel.

Exercises

Q2: ...Find the feature mapping $\Phi(\boldsymbol{z})$ that corresponds to the quadratic kernel.

First we expand the dot product inside, and square the entire sum. We will get a sum of the squares of the components and a sum of the cross products.

$$\begin{aligned}\left(\boldsymbol{x}^T\boldsymbol{y}+c\right)^2 &= \left(c+\sum_{i=1}^n x_i y_i\right)^2 \\ &= c^2+\sum_{i=1}^n x_i^2 y_i^2+\sum_{i=2}^n \sum_{j=1}^{i-1} 2x_i y_i x_j y_j+\sum_{i=1}^n 2x_i y_i c\end{aligned}$$

Pulling this sum into a dot product of x components and y components, we have

$$\begin{aligned}\Phi(x) = [&c, x_1^2, \ldots, x_n^2, \sqrt{2}x_1x_2, \ldots, \sqrt{2}x_1x_n, \\ &\sqrt{2}x_2x_3, \ldots, \sqrt{2}x_{n-1}x_n, \sqrt{2c}x_1, \ldots, \sqrt{2c}x_n]\end{aligned}$$

In this feature mapping, we have c, the squared components of the vector $\boldsymbol{x}$, $\sqrt{2}$ multiplied by all of the cross terms, and $\sqrt{2c}$ multiplied by all of the components.

Thank you!