

CS 6316 Machine Learning

Support Vector Machines and Kernel Methods

Yangfeng Ji

Department of Computer Science
University of Virginia



ENGINEERING

About Online Lectures

Course Information Update

- ▶ Record the lectures and upload the videos on Collab
- ▶ By default, turn off the video and mute yourself
- ▶ If you have a question
 - ▶ Unmute yourself and chime in anytime
 - ▶ Use the raise hand feature
 - ▶ Send me a private message

Course Information Update

- ▶ Record the lectures and upload the videos on Collab
- ▶ By default, turn off the video and mute yourself
- ▶ If you have a question
 - ▶ Unmute yourself and chime in anytime
 - ▶ Use the raise hand feature
 - ▶ Send me a private message
- ▶ Slack: as a stable communication channel to
 - ▶ send out instant messages if my network connection is unreliable
 - ▶ online discussion

Course Information Update

- ▶ Homework
 - ▶ Subject to change

Course Information Update

- ▶ Homework
 - ▶ Subject to change
- ▶ Final project
 - ▶ Send out my feedback later this week
 - ▶ Continue your collaboration with your teammates
 - ▶ Presentation: record a presentation video and share it with me

Course Information Update

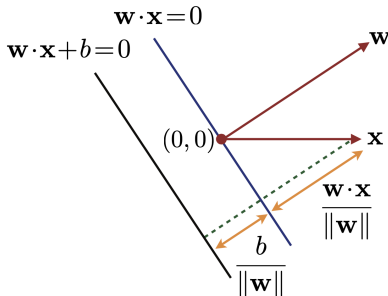
- ▶ Homework
 - ▶ Subject to change
- ▶ Final project
 - ▶ Send out my feedback later this week
 - ▶ Continue your collaboration with your teammates
 - ▶ Presentation: record a presentation video and share it with me
- ▶ Office hour
 - ▶ Wednesday 11 AM: I will be on Zoom
 - ▶ You can also send me an email or Slack message anytime

Separable Cases

Geometric Margin

The geometric margin of a linear binary classifier $h(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$ at a point \mathbf{x} is its distance to the hyper-plane $\langle \mathbf{w}, \mathbf{x} \rangle = 0$

$$\rho_h(\mathbf{x}) = \frac{|\langle \mathbf{w}, \mathbf{x} \rangle + b|}{\|\mathbf{w}\|_2} \quad (1)$$



Geometric Margin (II)

The geometric margin of $h(\mathbf{x})$ for a set of examples $T = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ is the minimal distance over these examples

$$\rho_h(T) = \min_{\mathbf{x}' \in T} \rho_h(\mathbf{x}') \quad (2)$$

[Mohri et al., 2018, Page 80]

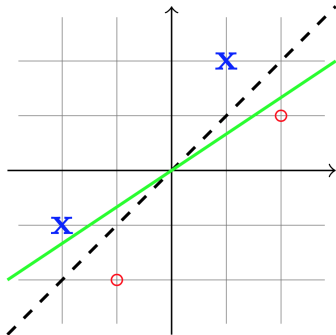
Half-Space Hypothesis Space

- ▶ Training set $S = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$ with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{+1, -1\}$
- ▶ If the training set is linearly separable

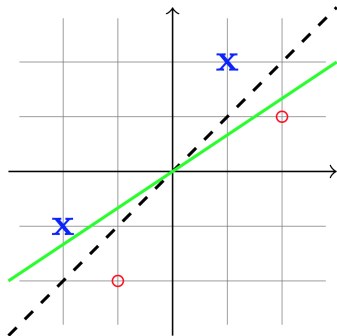
$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) > 0 \quad \forall i \in [m] \quad (3)$$

- ▶ Linearly separable cases
 - ▶ Existence of equation 3
 - ▶ All halfspace predictors that satisfy the condition in equation 3 are ERM hypotheses

Which Hypothesis is Better?



Which Hypothesis is Better?



- ▶ Intuitively, a hypothesis with larger *margin* is better, because it is more robust to noise
- ▶ Final definition of margin will be provided later

Hard SVM/Separable Cases

The mathematical formulation of the previous idea

$$\rho = \max_{(w,b)} \min_{i \in [m]} \frac{|\langle w, x_i \rangle + b|}{\|w\|_2} \quad (4)$$

$$\text{s.t. } y_i(\langle w, x_i \rangle + b) > 0 \quad \forall i \quad (5)$$

- $y_i(\langle w, x_i \rangle + b) > 0 \forall i$: guarantee (w, b) is an ERM hypothesis

Hard SVM/Separable Cases

The mathematical formulation of the previous idea

$$\rho = \max_{(w,b)} \min_{i \in [m]} \frac{|\langle w, x_i \rangle + b|}{\|w\|_2} \quad (4)$$

$$\text{s.t. } y_i(\langle w, x_i \rangle + b) > 0 \quad \forall i \quad (5)$$

- ▶ $y_i(\langle w, x_i \rangle + b) > 0 \forall i$: guarantee (w, b) is an ERM hypothesis
- ▶ $\min_{i \in [m]}$: calculate the margin between a hyper-plane and a set of examples

Hard SVM/Separable Cases

The mathematical formulation of the previous idea

$$\rho = \max_{(w,b)} \min_{i \in [m]} \frac{|\langle w, x_i \rangle + b|}{\|w\|_2} \quad (4)$$

$$\text{s.t. } y_i(\langle w, x_i \rangle + b) > 0 \quad \forall i \quad (5)$$

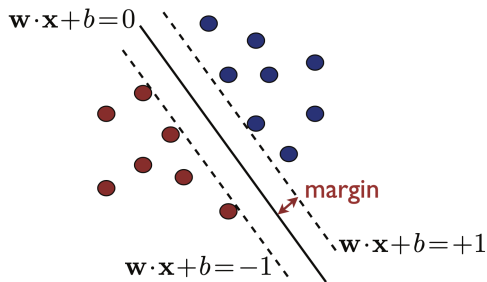
- ▶ $y_i(\langle w, x_i \rangle + b) > 0 \forall i$: guarantee (w, b) is an ERM hypothesis
- ▶ $\min_{i \in [m]}$: calculate the margin between a hyper-plane and a set of examples
- ▶ $\max_{(w,b)}$: maximize the margin

Illustration

Original form

$$\rho = \max_{(w,b)} \min_{i \in [m]} \frac{|\langle w, x_i \rangle + b|}{\|w\|_2} \quad (6)$$

$$\text{s.t. } y_i(\langle w, x_i \rangle + b) > 0 \quad \forall i \quad (7)$$



Alternative Forms

► Original form

$$\rho = \max_{(w,b)} \min_{i \in [m]} \frac{|\langle w, x_i \rangle + b|}{\|w\|_2} \quad (8)$$

$$\text{s.t. } y_i(\langle w, x_i \rangle + b) > 0 \quad \forall i \quad (9)$$

Alternative Forms

- ▶ Original form

$$\rho = \max_{(w,b)} \min_{i \in [m]} \frac{|\langle w, x_i \rangle + b|}{\|w\|_2} \quad (8)$$

$$\text{s.t. } y_i(\langle w, x_i \rangle + b) > 0 \quad \forall i \quad (9)$$

- ▶ Alternative form 1

$$\rho = \max_{(w,b)} \min_{i \in [m]} \frac{y_i(\langle w, x_i \rangle + b)}{\|w\|_2} \quad (10)$$

Alternative Forms

- ▶ Original form

$$\rho = \max_{(w,b)} \min_{i \in [m]} \frac{|\langle w, x_i \rangle + b|}{\|w\|_2} \quad (8)$$

$$\text{s.t. } y_i(\langle w, x_i \rangle + b) > 0 \quad \forall i \quad (9)$$

- ▶ Alternative form 1

$$\rho = \max_{(w,b)} \min_{i \in [m]} \frac{y_i(\langle w, x_i \rangle + b)}{\|w\|_2} \quad (10)$$

- ▶ Alternative form 2

$$\rho = \max_{(w,b): \min_{i \in [m]} y_i(\langle w, x_i \rangle + b) = 1} \frac{1}{\|w\|_2} \quad (11)$$

$$= \max_{(w,b): y_i(\langle w, x_i \rangle + b) \geq 1} \frac{1}{\|w\|_2} \quad (12)$$

Alternative Forms (II)

- ▶ Alternative form 2

$$\rho = \max_{(w,b): y_i(\langle w, x_i \rangle + b) \geq 1} \frac{1}{\|w\|_2} \quad (13)$$

- ▶ Alternative form 3: Quadratic programming (QP)

$$\begin{aligned} \min_{(w,b)} \quad & \frac{1}{2} \|w\|_2^2 \\ \text{s.t.} \quad & y_i(\langle w, x_i \rangle + b) \geq 1, \quad \forall i \in [m] \end{aligned} \quad (14)$$

which is a **constrained** optimization problem that can be solved by standard QP packages

Alternative Forms (II)

- ▶ Alternative form 2

$$\rho = \max_{(w,b): y_i(\langle w, x_i \rangle + b) \geq 1} \frac{1}{\|w\|_2} \quad (13)$$

- ▶ Alternative form 3: Quadratic programming (QP)

$$\begin{aligned} \min_{(w,b)} \quad & \frac{1}{2} \|w\|_2^2 \\ \text{s.t.} \quad & y_i(\langle w, x_i \rangle + b) \geq 1, \quad \forall i \in [m] \end{aligned} \quad (14)$$

which is a **constrained** optimization problem that can be solved by standard QP packages

- ▶ *Exercise:* Solve a SVM problem with quadratic programming

Unconstrained Optimization Problem

The quadratic programming problem with constraints can be converted to an unconstrained optimization problem with the Lagrangian method

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2} \|\mathbf{w}\|_2^2 - \sum_{i=1}^m \alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1) \quad (15)$$

where

- ▶ $\boldsymbol{\alpha} = \{\alpha_1, \dots, \alpha_m\}$ is the Lagrange multiplier, and
- ▶ $\alpha_i \geq 0$ is associated with the i -th training example

Constrained Optimization Problems

Constrained Optimization Problems: Definition

- ▶ $\mathcal{X} \subseteq \mathbb{R}^d$ and
- ▶ $f, g_i : \mathcal{X} \rightarrow \mathbb{R}, \forall i \in [m]$

Then, a constrained optimization problem is defined in the form of

$$\min_{x \in \mathcal{X}} \quad f(x) \quad (16)$$

$$\text{s.t.} \quad g_i(x) \leq 0, \forall i \in [m] \quad (17)$$

Constrained Optimization Problems: Definition

- ▶ $\mathcal{X} \subseteq \mathbb{R}^d$ and
- ▶ $f, g_i : \mathcal{X} \rightarrow \mathbb{R}, \forall i \in [m]$

Then, a constrained optimization problem is defined in the form of

$$\min_{x \in \mathcal{X}} \quad f(x) \quad (16)$$

$$\text{s.t.} \quad g_i(x) \leq 0, \forall i \in [m] \quad (17)$$

Comments

- ▶ In general definition, x is the target variable for optimization
- ▶ Special cases of $g_i(x)$: (1) $g_i(x) = 0$, (2) $g_i(x) \geq 0$, and (3) $g_i(x) \leq b$

Lagrangian

The Lagrangian associated to the general constrained optimization problem defined in equation 16 – 17 is the function defined over $\mathcal{X} \times \mathbb{R}_+^m$ as

$$L(\mathbf{x}, \boldsymbol{\alpha}) = f(\mathbf{x}) + \sum_{i=1}^m \alpha_i g_i(\mathbf{x}) \quad (18)$$

where

- ▶ $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m) \in \mathbb{R}_+^m$
- ▶ $\alpha_i \geq 0$ for any $i \in [m]$

Karush-Kuhn-Tucker's Theorem

Assume that $f, g_i : \mathcal{X} \rightarrow \mathbb{R}, \forall i \in [m]$ are **convex and differentiable** and that the constraints are qualified. Then \mathbf{x}' is a solution of the constrained problem **if and only if** there exist $\boldsymbol{\alpha}' \geq 0$ such that

$$\nabla_x L(\mathbf{x}', \boldsymbol{\alpha}') = \nabla_x f(\mathbf{x}') + \boldsymbol{\alpha}' \cdot \nabla_x g(\mathbf{x}') = 0 \quad (19)$$

$$\nabla_{\alpha} L(\mathbf{x}, \boldsymbol{\alpha}) = g(\mathbf{x}') \leq 0 \quad (20)$$

$$\boldsymbol{\alpha}' \cdot g(\mathbf{x}') = \sum_{i=1}^m \alpha'_i g_i(\mathbf{x}') = 0 \quad (21)$$

Equations 19 – 21 are called KKT conditions

[Mohri et al., 2018, Thm B.30]

KKT in SVM

Apply the KKT conditions to the SVM problem

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2} \|\boldsymbol{w}\|_2^2 - \sum_{i=1}^m \alpha_i (y_i (\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) - 1) \quad (22)$$

We have

$$\nabla_{\boldsymbol{w}} L = \boldsymbol{w} - \sum_{i=1}^m \alpha_i y_i \boldsymbol{x}_i = 0 \quad \Rightarrow \quad \boldsymbol{w} = \sum_{i=1}^m \alpha_i y_i \boldsymbol{x}_i$$

KKT in SVM

Apply the KKT conditions to the SVM problem

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2} \|\mathbf{w}\|_2^2 - \sum_{i=1}^m \alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1) \quad (22)$$

We have

$$\nabla_{\mathbf{w}} L = \mathbf{w} - \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i$$

$$\nabla_b L = - \sum_{i=1}^m \alpha_i y_i = 0 \quad \Rightarrow \quad \sum_{i=1}^m \alpha_i y_i = 0$$

KKT in SVM

Apply the KKT conditions to the SVM problem

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2} \|\mathbf{w}\|_2^2 - \sum_{i=1}^m \alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1) \quad (22)$$

We have

$$\nabla_{\mathbf{w}} L = \mathbf{w} - \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i$$

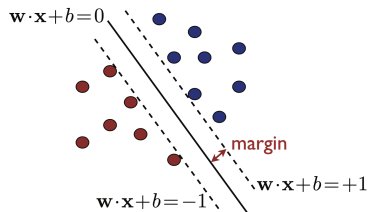
$$\nabla_b L = - \sum_{i=1}^m \alpha_i y_i = 0 \Rightarrow \sum_{i=1}^m \alpha_i y_i = 0$$

$$\forall i, \alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1) = 0 \Rightarrow \alpha_i = 0 \text{ or } y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = 1$$

Support Vectors

Consider the implication of the last equation in the previous page, $\forall i$

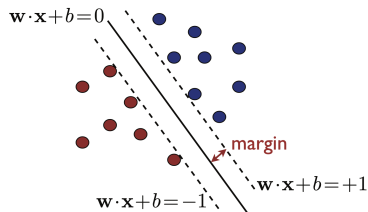
- ▶ $\alpha_i > 0$ and
 $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = 1$ or



Support Vectors

Consider the implication of the last equation in the previous page, $\forall i$

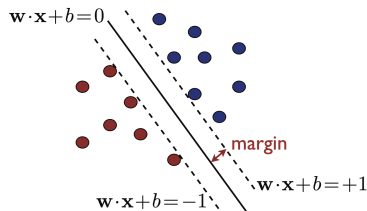
- ▶ $\alpha_i > 0$ and
 $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = 1$ or
- ▶ $\alpha_i = 0$ and
 $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1$



Support Vectors

Consider the implication of the last equation in the previous page, $\forall i$

- ▶ $\alpha_i > 0$ and $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = 1$ or
- ▶ $\alpha_i = 0$ and $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1$



$$\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i \quad (23)$$

- ▶ Examples with $\alpha_i > 0$ are called **support vectors**
- ▶ In \mathbb{R}^d , $d + 1$ examples are sufficient to define a hyper-plane

Non-separable Cases

Non-separable Cases

Recall the separable case:

$$\begin{aligned} \min_{(w,b)} \quad & \frac{1}{2} \|w\|_2^2 \\ \text{s.t.} \quad & y_i(\langle w, x_i \rangle + b) \geq 1, \quad \forall i \in [m] \end{aligned} \tag{24}$$

Non-separable Cases

Recall the separable case:

$$\begin{aligned} \min_{(w,b)} \quad & \frac{1}{2} \|w\|_2^2 \\ \text{s.t.} \quad & y_i(\langle w, x_i \rangle + b) \geq 1, \quad \forall i \in [m] \end{aligned} \tag{24}$$

For non-separable cases, there always exists an x_i , such that

$$y_i(\langle w, x_i \rangle + b) \not\geq 1 \tag{25}$$

or, we can formulate it as

$$y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i \tag{26}$$

with $\xi_i \geq 0$

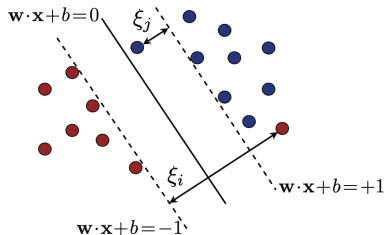
Geometric Meaning of ξ_i

Consider the relaxed constraint

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i \quad (27)$$

and three cases of ξ_i

- ▶ $\xi_i = 0$
- ▶ $0 < \xi_i < 1$
- ▶ $\xi_i > 0$



Non-separable Cases (II)

In general, the SVM problem of non-separable cases can be formulated as

$$\begin{aligned} \min_{(w,b)} \quad & \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^m \xi_i^p \\ \text{s.t.} \quad & y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i, \quad \forall i \in [m] \\ & \xi_i \geq 0 \end{aligned} \tag{28}$$

where $C \geq 0$, $p \geq 1$, and $\{\xi_i\}_{i=1}^m \geq 0$ are known as **slack variables** and are commonly used in optimization to define relaxed versions of constraints.

Lagrangian

Follows the same procedure as the separable cases, the Lagrangian is defined as

$$\begin{aligned} L(\mathbf{w}, b, \xi, \alpha, \beta) = & \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^m \xi_i \\ & - \sum_{i=1}^m \alpha_i (y_i (\mathbf{w}^\top \mathbf{x}_i + b) - 1 + \xi_i) \\ & - \sum_{i=1}^m \beta_i \xi_i \end{aligned} \quad (29)$$

with $\alpha_i, \beta_i \geq 0$

Lagrangian

Follows the same procedure as the separable cases, the Lagrangian is defined as

$$\begin{aligned} L(\mathbf{w}, b, \xi, \alpha, \beta) = & \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^m \xi_i \\ & - \sum_{i=1}^m \alpha_i (y_i (\mathbf{w}^\top \mathbf{x}_i + b) - 1 + \xi_i) \\ & - \sum_{i=1}^m \beta_i \xi_i \end{aligned} \quad (29)$$

with $\alpha_i, \beta_i \geq 0$

Exercise: show the KKT conditions of equation 29

Support Vectors

The first two equations in the KKT conditions are similar to the separable cases, and the rest are

$$\alpha_i + \beta_i = C \quad (30)$$

$$\alpha_i = 0 \quad \text{or} \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) = 1 - \xi_i \quad (31)$$

$$\beta_i = 0 \quad \text{or} \quad \xi_i = 0 \quad (32)$$

Depending the value of ξ_i , there are two types of support vectors

- ▶ $\xi_i = 0$: $\beta_i \geq 0$ and $0 < \alpha_i \leq C$
 - ▶ \mathbf{x}_i may lie on the marginal hyper-planes (as in the separable case)

Support Vectors

The first two equations in the KKT conditions are similar to the separable cases, and the rest are

$$\alpha_i + \beta_i = C \quad (30)$$

$$\alpha_i = 0 \quad \text{or} \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) = 1 - \xi_i \quad (31)$$

$$\beta_i = 0 \quad \text{or} \quad \xi_i = 0 \quad (32)$$

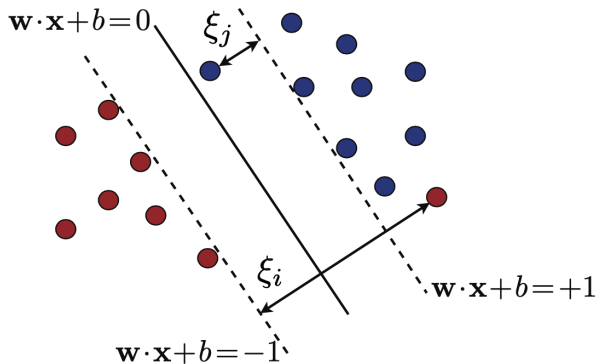
Depending the value of ξ_i , there are two types of support vectors

- ▶ $\xi_i = 0$: $\beta_i \geq 0$ and $0 < \alpha_i \leq C$
 - ▶ \mathbf{x}_i may lie on the marginal hyper-planes (as in the separable case)
- ▶ $\xi_i > 0$: $\beta_i = 0$ and $\alpha_i = C$
 - ▶ \mathbf{x}_i is an outlier

Support Vectors (II)

Two types of support vectors

- ▶ $\alpha_i = C$: \mathbf{x}_i is an outlier
- ▶ $0 < \alpha_i < C$: \mathbf{x}_i lies on the marginal hyper-planes



Dual Optimization Problem

Lagrangian

Combine the Lagrangian

$$\begin{aligned} L &= \frac{1}{2} \|\mathbf{w}\|_2^2 - \sum_{i=1}^m \alpha_i [y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1] \\ &= \frac{1}{2} \|\mathbf{w}\|_2^2 - \sum_{i=1}^m \alpha_i y_i \langle \mathbf{w}, \mathbf{x}_i \rangle - b \sum_{i=1}^m \alpha_i y_i + \sum_{i=1}^m \alpha_i \end{aligned}$$

Lagrangian

Combine the Lagrangian

$$\begin{aligned} L &= \frac{1}{2} \|\mathbf{w}\|_2^2 - \sum_{i=1}^m \alpha_i [y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1] \\ &= \frac{1}{2} \|\mathbf{w}\|_2^2 - \sum_{i=1}^m \alpha_i y_i \langle \mathbf{w}, \mathbf{x}_i \rangle - b \sum_{i=1}^m \alpha_i y_i + \sum_{i=1}^m \alpha_i \end{aligned}$$

with some of the KKT conditions

$$\mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i \quad (33)$$

$$\sum_{i=1}^m \alpha_i y_i = 0, \quad (34)$$

we have ...

Dual Problem

$$\begin{aligned} L = & \frac{1}{2} \left\| \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i \right\|_2^2 - \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ & - b \underbrace{\sum_{i=1}^m \alpha_i y_i}_{=0} + \sum_{i=1}^m \alpha_i \end{aligned} \tag{35}$$

Dual Problem

$$\begin{aligned} L = & \frac{1}{2} \left\| \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i \right\|_2^2 - \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ & - b \underbrace{\sum_{i=1}^m \alpha_i y_i}_{=0} + \sum_{i=1}^m \alpha_i \end{aligned} \quad (35)$$

Given $\left\| \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i \right\|_2^2 = \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$, we have

$$L = -\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \sum_{i=1}^m \alpha_i \quad (36)$$

Dual Problem (II)

The dual optimization problem for SVMs of the separable cases is

$$\max_{\alpha} \quad \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \quad (37)$$

$$\text{s.t.} \quad \alpha_i \geq 0 \quad (38)$$

$$\sum_{i=1}^m \alpha_i y_i = 0 \quad \forall i \in [m] \quad (39)$$

Dual Problem (II)

The dual optimization problem for SVMs of the separable cases is

$$\max_{\alpha} \quad \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \quad (37)$$

$$\text{s.t.} \quad \alpha_i \geq 0 \quad (38)$$

$$\sum_{i=1}^m \alpha_i y_i = 0 \quad \forall i \in [m] \quad (39)$$

- ▶ Lagrange multiplier α is also called dual variable
- ▶ This is an optimization problem only about α

Dual Problem (II)

The dual optimization problem for SVMs of the separable cases is

$$\max_{\alpha} \quad \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \quad (37)$$

$$\text{s.t.} \quad \alpha_i \geq 0 \quad (38)$$

$$\sum_{i=1}^m \alpha_i y_i = 0 \quad \forall i \in [m] \quad (39)$$

- ▶ Lagrange multiplier α is also called dual variable
- ▶ This is an optimization problem only about α
- ▶ The dual problem is defined on the inner product $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$

Primal and Dual Problem

- ▶ Primal problem

$$\begin{aligned} \min_{(w,b)} \quad & \frac{1}{2} \|\mathbf{w}\|_2^2 \\ \text{s.t.} \quad & y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1, \quad \forall i \in [m] \end{aligned} \tag{40}$$

- ▶ Dual problem

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ \text{s.t.} \quad & \sum_{i=1}^m \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0 \quad \forall i \in [m] \end{aligned} \tag{41}$$

- ▶ These two problems are equivalent

[Boyd and Vandenberghe, 2004, Chapter 5]

Primal and Dual Problem

- ▶ Primal problem

$$\begin{aligned} \min_{(w,b)} \quad & \frac{1}{2} \|w\|_2^2 \\ \text{s.t.} \quad & y_i(\langle w, x_i \rangle + b) \geq 1, \quad \forall i \in [m] \end{aligned} \tag{40}$$

- ▶ Dual problem

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\ \text{s.t.} \quad & \sum_{i=1}^m \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0 \quad \forall i \in [m] \end{aligned} \tag{41}$$

- ▶ These two problems are equivalent

[Boyd and Vandenberghe, 2004, Chapter 5]

Primal and Dual Problem

- ▶ Primal problem

$$\begin{aligned} \min_{(w,b)} \quad & \frac{1}{2} \|\mathbf{w}\|_2^2 \\ \text{s.t.} \quad & y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1, \quad \forall i \in [m] \end{aligned} \tag{40}$$

- ▶ Dual problem

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ \text{s.t.} \quad & \sum_{i=1}^m \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0 \quad \forall i \in [m] \end{aligned} \tag{41}$$

- ▶ These two problems are equivalent

[Boyd and Vandenberghe, 2004, Chapter 5]

SVM Hypothesis, revisited

Once we solve the dual problem with α , we have the solution of w as

$$w = \sum_{i=1}^m \alpha_i y_i x_i \quad (42)$$

and the hypothesis $h(x)$ as

$$h(x) = \text{sign}(\langle w, x \rangle + b) \quad (43)$$

$$(45)$$

SVM Hypothesis, revisited

Once we solve the dual problem with α , we have the solution of w as

$$w = \sum_{i=1}^m \alpha_i y_i x_i \quad (42)$$

and the hypothesis $h(x)$ as

$$h(x) = \text{sign}(\langle w, x \rangle + b) \quad (43)$$

$$= \text{sign}(\langle \sum_{i=1}^m \alpha_i y_i x_i, x \rangle + b) \quad (44)$$

$$(45)$$

SVM Hypothesis, revisited

Once we solve the dual problem with α , we have the solution of w as

$$w = \sum_{i=1}^m \alpha_i y_i x_i \quad (42)$$

and the hypothesis $h(x)$ as

$$h(x) = \text{sign}(\langle w, x \rangle + b) \quad (43)$$

$$= \text{sign}(\langle \sum_{i=1}^m \alpha_i y_i x_i, x \rangle + b) \quad (44)$$

$$= \text{sign}(\sum_{i=1}^m \alpha_i y_i \langle x_i, x \rangle + b) \quad (45)$$

SVM Hypothesis, revisited

Once we solve the dual problem with α , we have the solution of w as

$$w = \sum_{i=1}^m \alpha_i y_i x_i \quad (42)$$

and the hypothesis $h(x)$ as

$$h(x) = \text{sign}(\langle w, x \rangle + b) \quad (43)$$

$$= \text{sign}(\langle \sum_{i=1}^m \alpha_i y_i x_i, x \rangle + b) \quad (44)$$

$$= \text{sign}(\sum_{i=1}^m \alpha_i y_i \langle x_i, x \rangle + b) \quad (45)$$

Exercise: Prove $b = y_i - \sum_{i=1}^m \alpha_i y_i \langle x_i, x \rangle$ for any x_i with $\alpha_i > 0$

Kernel Methods

Properties of Inner Product

In the solution of SVMs

$$\begin{aligned} h(\mathbf{x}) &= \text{sign}\left(\sum_{i=1}^m \alpha_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b\right) \\ b &= y_i - \sum_{i=1}^m \alpha_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle \end{aligned} \tag{46}$$

Properties of Inner Product

In the solution of SVMs

$$\begin{aligned}h(\mathbf{x}) &= \text{sign}\left(\sum_{i=1}^m \alpha_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b\right) \\b &= y_i - \sum_{i=1}^m \alpha_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle\end{aligned}\tag{46}$$

Extend the capacity of SVMs by replacing the inner product $\langle \mathbf{x}_i, \mathbf{x} \rangle$ with a kernel function

$$K(\mathbf{x}_i, \mathbf{x}) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}) \rangle\tag{47}$$

where $\Phi(\cdot)$ is a nonlinear mapping function.

Examples: Polynomial Kernels

For any constant $c > 0$, a **polynomial kernel** of degree $d \in \mathbb{N}$ is the kernel K defined over \mathbb{R}^n by

$$K(x, x') = (\langle x, x' \rangle + c)^d, \forall x, x' \in \mathbb{R}^n \quad (48)$$

Examples: Polynomial Kernels

For any constant $c > 0$, a **polynomial kernel** of degree $d \in \mathbb{N}$ is the kernel K defined over \mathbb{R}^n by

$$K(x, x') = (\langle x, x' \rangle + c)^d, \forall x, x' \in \mathbb{R}^n \quad (48)$$

Special cases

- ▶ $d = 1$: $K(x, x') = \langle x, x' \rangle + c$
- ▶ $d = 2$: $K(x, x') = (\langle x, x' \rangle + c)^2$

Examples: Polynomial Kernels (II)

For the special case with $d = 2$, assume $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^2$

$$K(\mathbf{x}, \mathbf{x}') = (\langle \mathbf{x}, \mathbf{x}' \rangle + c)^2 \quad (49)$$

$$= (x_1 x'_1 + x_2 x'_2 + c)^2 \quad (50)$$

Examples: Polynomial Kernels (II)

For the special case with $d = 2$, assume $x, x' \in \mathbb{R}^2$

$$K(x, x') = (\langle x, x' \rangle + c)^2 \quad (49)$$

$$= (x_1 x'_1 + x_2 x'_2 + c)^2 \quad (50)$$

$$= x_1^2 x'^2_1 + x_2^2 x'^2_2 + 2x_1 x'_1 x_2 x'_2 \quad (51)$$

$$+ cx_1 x'_1 + cx_2 x'_2 + c^2 \quad (52)$$

Examples: Polynomial Kernels (II)

For the special case with $d = 2$, assume $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^2$

$$K(\mathbf{x}, \mathbf{x}') = (\langle \mathbf{x}, \mathbf{x}' \rangle + c)^2 \quad (49)$$

$$= (x_1 x'_1 + x_2 x'_2 + c)^2 \quad (50)$$

$$= x_1^2 x'^2_1 + x_2^2 x'^2_2 + 2x_1 x'_1 x_2 x'_2 \quad (51)$$

$$+ cx_1 x'_1 + cx_2 x'_2 + c^2 \quad (52)$$

$$= [x_1^2, x_2^2, \sqrt{2}x_1 x_2, \sqrt{2}cx_1, \sqrt{2}cx_2, c] \begin{bmatrix} x'^2_1 \\ x'^2_2 \\ \sqrt{2}x'_1 x'_2 \\ \sqrt{2}cx'_1 \\ \sqrt{2}cx'_2 \\ c \end{bmatrix}$$

Examples: Polynomial Kernels (III)

Let $K(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle$, then

$$\Phi(\mathbf{x}) = [x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2c}x_1, \sqrt{2c}x_2, c] \quad (53)$$

which maps a 2-D data point \mathbf{x} into a 6-D space as $\Phi(\mathbf{x})$

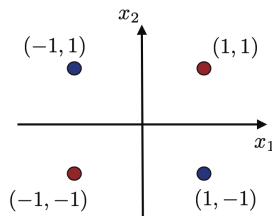
Examples: Polynomial Kernels (III)

Let $K(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle$, then

$$\Phi(\mathbf{x}) = [x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}cx_1, \sqrt{2}cx_2, c] \quad (53)$$

which maps a 2-D data point \mathbf{x} into a 6-D space as $\Phi(\mathbf{x})$

Recall the XOR problem



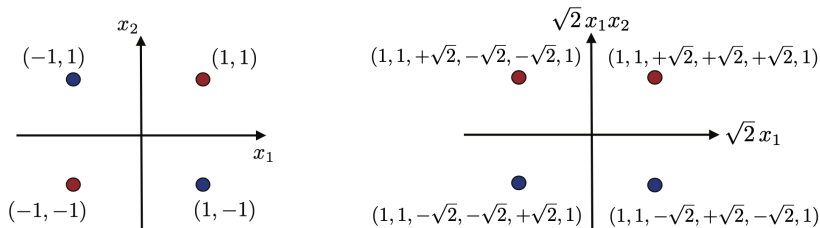
Examples: Polynomial Kernels (III)

Let $K(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle$, then

$$\Phi(\mathbf{x}) = [x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}cx_1, \sqrt{2}cx_2, c] \quad (53)$$

which maps a 2-D data point \mathbf{x} into a 6-D space as $\Phi(\mathbf{x})$

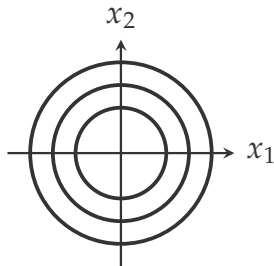
Recall the XOR problem



Gaussian Kernels

For any constant $\sigma > 0$, a **Gaussian kernel** or **radial basis function** (RBF) is the kernel K defined over \mathbb{R}^d by

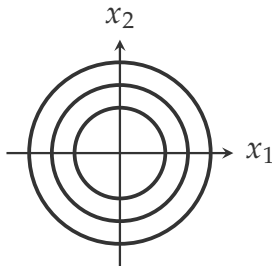
$$K(x, x') = \exp\left(-\frac{\|x' - x\|_2^2}{2\sigma^2}\right) \quad (54)$$



Gaussian Kernels

For any constant $\sigma > 0$, a **Gaussian kernel** or **radial basis function** (RBF) is the kernel K defined over \mathbb{R}^d by

$$K(x, x') = \exp\left(-\frac{\|x' - x\|_2^2}{2\sigma^2}\right) \quad (54)$$



Question: What $\Phi(x)$ looks like in this case?

SVMs with Kernel Functions

► Problem definition

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ \text{s.t.} \quad & \alpha_i \geq 0 \text{ and } \sum_{i=1}^m \alpha_i y_i = 0, i \in [m] \end{aligned} \tag{55}$$

SVMs with Kernel Functions

- ▶ Problem definition

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ \text{s.t.} \quad & \alpha_i \geq 0 \text{ and } \sum_{i=1}^m \alpha_i y_i = 0, i \in [m] \end{aligned} \quad (55)$$

- ▶ Solution: separable case

$$h(x) = \text{sign} \left(\sum_{i=1}^m \alpha_i y_i K(x_i, x) + b \right) \quad (56)$$

with $b = y_i - \sum_{j=1}^m \alpha_j y_j K(x_j, x_i)$ for any x_i with $\alpha_i > 0$

The Choice of Kernels

- ▶ The choice of $K(x, x')$ can be arbitrary, as long as the existence of $\Phi(\cdot)$ is guaranteed
 - ▶ For many cases, $\Phi(\cdot)$ cannot be found explicitly

The Choice of Kernels

- ▶ The choice of $K(\mathbf{x}, \mathbf{x}')$ can be arbitrary, as long as the existence of $\Phi(\cdot)$ is guaranteed
 - ▶ For many cases, $\Phi(\cdot)$ cannot be found explicitly
- ▶ Alternatively, we only need to make sure $K(\mathbf{x}, \mathbf{x}')$ is *positive definite symmetric* (PDS)
 - ▶ A kernel K is PDS if for any $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ the matrix \mathbf{K} is symmetric positive **semi-definite**

$$\mathbf{K} = [K(\mathbf{x}_i, \mathbf{x}_j)]_{i,j} \in \mathbb{R}^{m \times m} \quad (57)$$

The Choice of Kernels

- ▶ The choice of $K(\mathbf{x}, \mathbf{x}')$ can be arbitrary, as long as the existence of $\Phi(\cdot)$ is guaranteed
 - ▶ For many cases, $\Phi(\cdot)$ cannot be found explicitly
- ▶ Alternatively, we only need to make sure $K(\mathbf{x}, \mathbf{x}')$ is *positive definite symmetric* (PDS)
 - ▶ A kernel K is PDS if for any $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ the matrix \mathbf{K} is symmetric positive **semi-definite**

$$\mathbf{K} = [K(\mathbf{x}_i, \mathbf{x}_j)]_{i,j} \in \mathbb{R}^{m \times m} \quad (57)$$

- ▶ A symmetric positive semi-definite matrix is defined as

$$\mathbf{c}^\top \mathbf{K} \mathbf{c} \geq 0 \quad (58)$$

Reference



Boyd, S. and Vandenberghe, L. (2004).
Convex optimization.
Cambridge university press.



Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018).
Foundations of machine learning.
MIT press.



Shalev-Shwartz, S. and Ben-David, S. (2014).
Understanding machine learning: From theory to algorithms.
Cambridge university press.