CS 4774 Machine Learning

Support Vector Machines and Kernel Methods

Yangfeng Ji

Information and Language Processing Lab
Department of Computer Science
University of Virginia

Overview

1. Review: Linear Functions
2. Separable Cases
3. Constrained Optimization
4. Non-separable Cases
5. Dual Optimization Problem
6. Kernel Methods

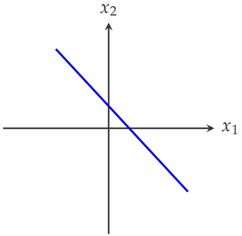
Readings: [Shalev-Shwartz and Ben-David, 2014, Chapter 15 & 16]

Review: Linear Functions

Linear Functions

Consider a two-dimensional case with $\boldsymbol{w} = (1, 1, -0.5)$

$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b = x_1 + x_2 - 0.5 \tag{1}$$



Different values of $f(\boldsymbol{x})$ map to different areas on this 2-D space. For example, the following equation defines the blue line L.

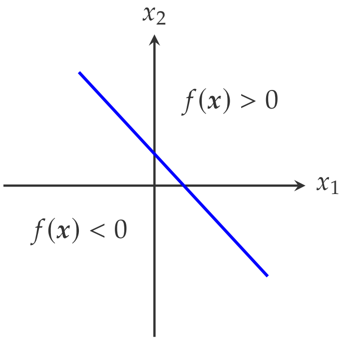
$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b = 0 \tag{2}$$

Properties of Linear Functions (Cont.)

Furthermore,

$$f(\boldsymbol{x}) = x_1 + x_2 - 0.5 = 0 \tag{3}$$

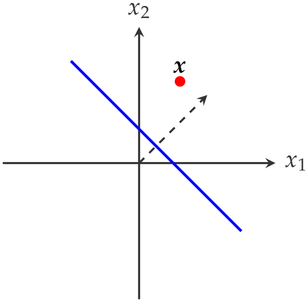
separates the 2-D space $\mathbb{R}^2$ into two half spaces



Properties of Linear Functions (Cont.)

The distance of point $\boldsymbol{x}$ to line $L : f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = 0$ is given by

$$\frac{f(\boldsymbol{x})}{\|\boldsymbol{w}\|_2} = \frac{\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b}{\|\boldsymbol{x}\|_2} = \langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}, \boldsymbol{x} \rangle + \frac{b}{\|\boldsymbol{w}\|_2} \tag{4}$$

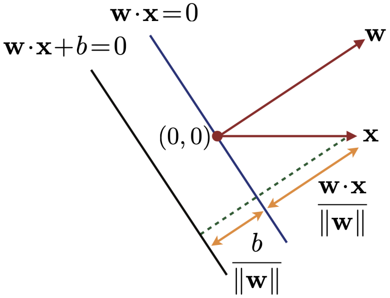


Separable Cases

Geometric Margin

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

$$\rho_h(x) = \frac{|\langle w, x \rangle + b|}{\|w\|_2} \tag{5}$$



Geometric Margin (II)

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

$$\rho_h(T) = \min_{\boldsymbol{x}' \in T} \rho_h(\boldsymbol{x}') \tag{6}$$

[Mohri et al., 2018, Page 80]

Half-Space Hypothesis Space

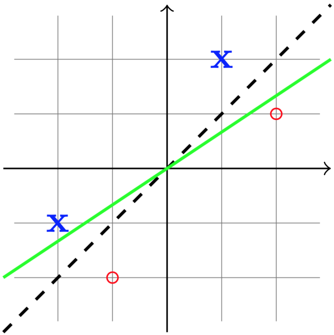
- Training set $S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\}$ with $\boldsymbol{x}_i \in \mathbb{R}^d$ and $y_i \in \{+1, -1\}$
- If the training set is linearly separable

$$y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) > 0 \quad \forall i \in [m] \tag{7}$$

- Linearly separable cases
  - Existence of equation 7
  - All halfspace predictors that satisfy the condition in equation 7 are ERM hypotheses

Which Hypothesis is Better?

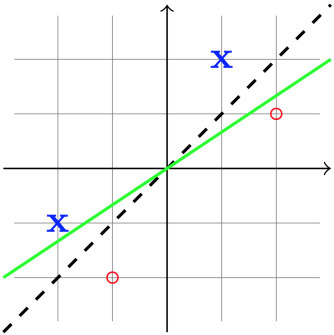
Is the one represented by the green line or the black dashed line?



[Shalev-Shwartz and Ben-David, 2014, Page 203]

Which Hypothesis is Better?

Is the one represented by the green line or the black dashed line?



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

[Shalev-Shwartz and Ben-David, 2014, Page 203]

Hard SVM/Separable Cases

The mathematical formulation of the previous idea

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

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

s.t. means *subject to* in optimization, to introduce constraints

Notations:

- $y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b) > 0\ \forall i$: guarantee $(\boldsymbol{w}, b)$ is an ERM hypothesis

Hard SVM/Separable Cases

The mathematical formulation of the previous idea

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

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

s.t. means *subject to* in optimization, to introduce constraints

Notations:

- $y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) > 0 \; \forall i$: guarantee $(\boldsymbol{w}, b)$ 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

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

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

s.t. means *subject to* in optimization, to introduce constraints

Notations:

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

Hard SVM/Separable Cases

The mathematical formulation of the previous idea

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

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

s.t. means *subject to* in optimization, to introduce constraints

Notations:

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

Overall, the optimization problem is to find a hypothesis that (1) classifies all training example correctly and (2) also has the largest margin.

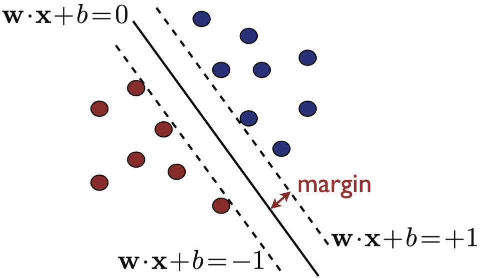
Illustration

Original form

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

$$\text{s.t. } y_i(\langle w, x_i\rangle + b) > 0 \quad \forall i \quad (11)$$

An example with the margin as 1



Alternative Forms

- Original form

$$
\begin{aligned}
\rho &= \max_{(\boldsymbol{w},b)} \min_{i\in[m]} \frac{|\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b|}{\|\boldsymbol{w}\|_2} && (12)\\
&\quad \text{s.t. } y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b) > 0 \quad \forall i && (13)
\end{aligned}
$$

Alternative Forms

- Original form

$$\rho = \max_{(\boldsymbol{w},b)} \min_{i\in[m]} \frac{|\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b|}{\|\boldsymbol{w}\|_2} \tag{12}$$

$$\text{s.t. } y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b) > 0 \quad \forall i \tag{13}$$

- Alternative form 1

$$\rho = \max_{(\boldsymbol{w},b)} \min_{i\in[m]} \frac{y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b)}{\|\boldsymbol{w}\|_2} \tag{14}$$

Alternative Forms

- Original form

$$\rho = \max_{(\boldsymbol{w},b)} \min_{i\in[m]} \frac{|\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b|}{\|\boldsymbol{w}\|_2} \tag{12}$$

$$\text{s.t. } y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b) > 0 \quad \forall i \tag{13}$$

- Alternative form 1

$$\rho = \max_{(\boldsymbol{w},b)} \min_{i\in[m]} \frac{y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b)}{\|\boldsymbol{w}\|_2} \tag{14}$$

- Alternative form 2

$$\rho = \max_{(\boldsymbol{w},b):\ \min_{i\in[m]} y_i(\langle \boldsymbol{w},\boldsymbol{x}_i\rangle + b = 1} \frac{1}{\|\boldsymbol{w}\|_2} \tag{15}$$

$$= \max_{(\boldsymbol{w},b):\ y_i(\langle \boldsymbol{w},\boldsymbol{x}_i\rangle + b \geq 1} \frac{1}{\|\boldsymbol{w}\|_2} \tag{16}$$

Alternative Forms (II)

- Alternative form 2

$$\rho = \max_{(\boldsymbol{w},b):\ y_i(\langle \boldsymbol{w},\boldsymbol{x}_i\rangle+b\geq 1} \frac{1}{\|\boldsymbol{w}\|_2} \tag{17}$$

- Alternative form 3: Quadratic programming (QP)

$$\begin{aligned} \min_{(\boldsymbol{w},b)}\ & \frac{1}{2}\|\boldsymbol{w}\|_2^2 \\ \text{s.t.}\ & y_i(\langle \boldsymbol{w},\boldsymbol{x}_i\rangle + b) \geq 1, \quad \forall i \in [m] \end{aligned} \tag{18}$$

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

Unconstrained Optimization Problem

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

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 - \sum_{i=1}^{m} \alpha_i (y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) - 1) \tag{19}$$

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

Unconstrained Optimization Problem

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

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 - \sum_{i=1}^{m} \alpha_i (y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) - 1) \tag{19}$$

where

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

Can you identify the similarity between Eq. 19 and regularized linear regression?

SVM Online Demo

Link

Constrained Optimization

Constrained Optimization Problems: Definition

A generic formulation of constrained optimization

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

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

$$\min_{\boldsymbol{x} \in \mathcal{X}} \quad f(\boldsymbol{x}) \tag{20}$$

$$\text{s.t.} \quad g_i(\boldsymbol{x}) \leq 0, \forall i \in [m] \tag{21}$$

Constrained Optimization Problems: Definition

A generic formulation of constrained optimization

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

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

$$\min_{\boldsymbol{x} \in \mathcal{X}} \quad f(\boldsymbol{x}) \tag{20}$$

$$\text{s.t.} \quad g_i(\boldsymbol{x}) \leq 0, \forall i \in [m] \tag{21}$$

Comments

- Unlike a learning problem, here $\boldsymbol{x}$ is the target variable for optimization
- Special cases of $g_i(\boldsymbol{x})$: (1) $g_i(\boldsymbol{x}) = 0$, (2) $g_i(\boldsymbol{x}) \geq 0$, and (3) $g_i(\boldsymbol{x}) \leq b$

Lagrangian

The Lagrangian associated to the general constrained optimization problem defined in equation 20 – 21 is the function defined over $\mathcal{X} \times \mathbb{R}_+^m$ as

$$L(\boldsymbol{x}, \boldsymbol{\alpha}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \alpha_i g_i(\boldsymbol{x}) \tag{22}$$

where

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

Karush-Kuhn-Tucker's Theorem

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

$$
\begin{aligned}
\nabla_{\boldsymbol{x}} L(\boldsymbol{x}', \boldsymbol{\alpha}') &= \nabla_{\boldsymbol{x}} f(\boldsymbol{x}') + \boldsymbol{\alpha}' \cdot \nabla_{\boldsymbol{x}} g(\boldsymbol{x}) = 0 && (23) \\
\nabla_{\boldsymbol{\alpha}} L(\boldsymbol{x}, \boldsymbol{\alpha}) &= g(\boldsymbol{x}') \leq 0 && (24) \\
\boldsymbol{\alpha}' \cdot g(\boldsymbol{x}') &= \sum_{i=1}^{m} \alpha_i' g_i(\boldsymbol{x}') = 0 && (25)
\end{aligned}
$$

Equations 23 – 25 are called KKT conditions

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

KKT in SVM

Apply the KKT conditions to the SVM problem

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 - \sum_{i=1}^{m} \alpha_i (y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) - 1) \tag{26}$$

We have

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

KKT in SVM

Apply the KKT conditions to the SVM problem

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 - \sum_{i=1}^{m} \alpha_i (y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) - 1) \tag{26}$$

We have

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

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

KKT in SVM

Apply the KKT conditions to the SVM problem

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 - \sum_{i=1}^{m} \alpha_i (y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) - 1) \tag{26}$$

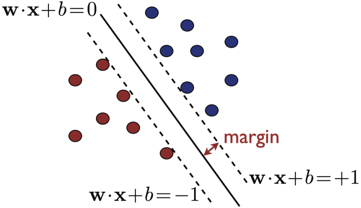
We have

$$\begin{aligned}
\nabla_{\boldsymbol{w}} L = \boldsymbol{w} - \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i = 0 \quad &\Rightarrow \quad \boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i \\
\nabla_b L = -\sum_{i=1}^{m} \alpha_i y_i = 0 \quad &\Rightarrow \quad \sum_{i=1}^{m} \alpha_i y_i = 0 \\
\forall i, \alpha_i (y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) - 1) = 0 \quad &\Rightarrow \quad \alpha_i = 0 \text{ or } y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) = 1
\end{aligned}$$

Support Vectors

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

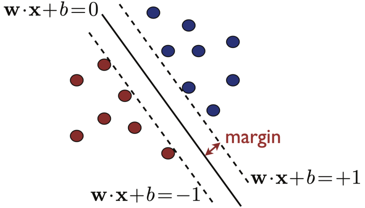
- $\alpha_i > 0$ and $y_i(\langle w, x_i \rangle + b) = 1$
  or



Support Vectors

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

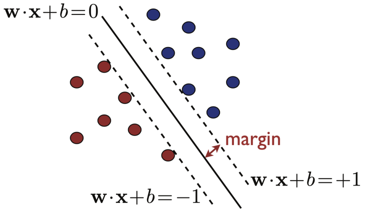
- $\alpha_i > 0$ and $y_i(\langle w, x_i \rangle + b) = 1$
  or
- $\alpha_i = 0$ and $y_i(\langle w, x_i \rangle + b) \geq 1$



Support Vectors

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

- $\alpha_i > 0$ and $y_i(\langle w, x_i \rangle + b) = 1$
  or
- $\alpha_i = 0$ and $y_i(\langle w, x_i \rangle + b) \geq 1$



$$w = \sum_{i=1}^{m} \alpha_i y_i x_i \tag{27}$$

- 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:

$$\begin{aligned} &\min_{(\boldsymbol{w},b)} \frac{1}{2}\|\boldsymbol{w}\|_2^2 \\ &\text{s.t. } y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq 1, \quad \forall i \in [m] \end{aligned} \tag{28}$$

Non-separable Cases

Recall the separable case:

$$\begin{aligned} \min_{(\boldsymbol{w},b)} \; & \frac{1}{2}\|\boldsymbol{w}\|_2^2 \\ \text{s.t. } & y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq 1, \quad \forall i \in [m] \end{aligned} \tag{28}$$

For non-separable cases, there always exists an $\boldsymbol{x}_i$, such that

$$y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \ngeq 1 \tag{29}$$

or, we can formulate it as

$$y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq 1 - \xi_i \tag{30}$$

with $\xi_i \geq 0$

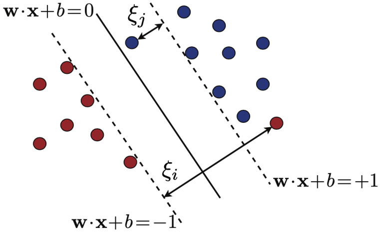
Geometric Meaning of ξ_i

Consider the relaxed constraint

$$y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i \tag{31}$$

and three cases of ξ_i

- $\xi_i = 0$
- $0 < \xi_i < 1$
- $\xi_i \geq 1$



Non-separable Cases (II)

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

$$
\begin{aligned}
&\min_{(\boldsymbol{w},b)} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i^p \\
&\text{s.t. } y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b) \geq 1 - \xi_i, \quad \forall i \in [m] \\
&\quad\quad \xi_i \geq 0
\end{aligned}
\tag{32}
$$

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

$$
\begin{aligned}
L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i \\
& - \sum_{i=1}^{m}\alpha_i(y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1 + \xi_i) \\
& - \sum_{i=1}^{m}\beta_i\xi_i
\end{aligned} \tag{33}
$$

with $\alpha_i, \beta_i \geq 0$

Support Vectors

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

$$\alpha_i + \beta_i = C \tag{34}$$

$$\alpha_i = 0 \quad \text{or} \quad y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) = 1 - \xi_i \tag{35}$$

$$\beta_i = 0 \quad \text{or} \quad \xi_i = 0 \tag{36}$$

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

- $\xi_i = 0$: $\beta_i \geq 0$ and $0 < \alpha_i \leq C$
  - $\boldsymbol{x}_i$ 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

$$\alpha_i + \beta_i = C \tag{34}$$

$$\alpha_i = 0 \quad \text{or} \quad y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) = 1 - \xi_i \tag{35}$$

$$\beta_i = 0 \quad \text{or} \quad \xi_i = 0 \tag{36}$$

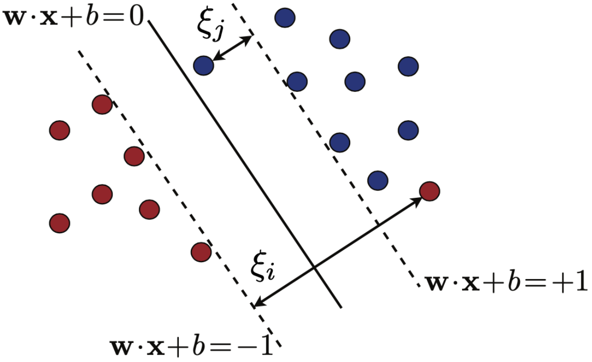
Depending the value of ξ_i, there are two types of support vectors

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

Support Vectors (II)

Two types of support vectors

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



Dual Optimization Problem

Lagrangian

Combine the Lagrangian

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

Lagrangian

Combine the Lagrangian

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

with some of the KKT conditions

$$
\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i \tag{37}
$$

$$
\sum_{i=1}^{m} \alpha_i y_i = 0, \tag{38}
$$

we have ...

Dual Problem

$$L = \frac{1}{2} \| \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i \|_2^2 - \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \\ - \underbrace{b \sum_{i=1}^{m} \alpha_i y_i}_{=0} + \sum_{i=1}^{m} \alpha_i \tag{39}$$

Dual Problem

$$L = \frac{1}{2} \| \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i \|_2^2 - \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \\ - \underbrace{b \sum_{i=1}^{m} \alpha_i y_i}_{=0} + \sum_{i=1}^{m} \alpha_i \tag{39}$$

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

$$L = -\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + \sum_{i=1}^{m} \alpha_i \tag{40}$$

Dual Problem (II)

The dual optimization problem for SVMs of the separable cases is

$$\max_{\boldsymbol{\alpha}} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \tag{41}$$

$$\text{s.t.} \quad \alpha_i \geq 0 \tag{42}$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \; \forall i \in [m] \tag{43}$$

Dual Problem (II)

The dual optimization problem for SVMs of the separable cases is

$$\max_{\boldsymbol{\alpha}} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \tag{41}$$

$$\text{s.t.} \quad \alpha_i \geq 0 \tag{42}$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \ \forall i \in [m] \tag{43}$$

- Lagrange multiplier $\boldsymbol{\alpha}$ is also called dual variable
- This is an optimization problem only about $\boldsymbol{\alpha}$

Dual Problem (II)

The dual optimization problem for SVMs of the separable cases is

$$\max_{\boldsymbol{\alpha}} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \tag{41}$$

$$\text{s.t.} \quad \alpha_i \geq 0 \tag{42}$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \ \forall i \in [m] \tag{43}$$

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

Primal and Dual Problem

- Primal problem

$$\begin{aligned} \min_{(\boldsymbol{w},b)} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 \\ \text{s.t.} \quad & y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq 1, \quad \forall i \in [m] \end{aligned} \tag{44}$$

- Dual problem

$$\begin{aligned} \max_{\boldsymbol{\alpha}} & \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \\ \text{s.t.} & \sum_{i=1}^{m} \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0 \ \forall i \in [m] \end{aligned} \tag{45}$$

- These two problems are equivalent

[Boyd and Vandenberghe, 2004, Chapter 5]

SVM Hypothesis, revisited

Once we solve the dual problem with $\boldsymbol{\alpha}$, we have the solution of $\boldsymbol{w}$ as

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i \tag{46}$$

and the hypothesis $h(\boldsymbol{x})$ as

$$h(\boldsymbol{x}) \quad = \quad \text{sign}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b) \tag{47}$$

$$\tag{49}$$

SVM Hypothesis, revisited

Once we solve the dual problem with $\boldsymbol{\alpha}$, we have the solution of $\boldsymbol{w}$ as

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i \tag{46}$$

and the hypothesis $h(\boldsymbol{x})$ as

$$\begin{aligned} h(\boldsymbol{x}) &= \operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b) && (47) \\ &= \operatorname{sign}(\langle \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i, \boldsymbol{x} \rangle + b) && (48) \\ & && (49) \end{aligned}$$

SVM Hypothesis, revisited

Once we solve the dual problem with $\boldsymbol{\alpha}$, we have the solution of $\boldsymbol{w}$ as

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i \tag{46}$$

and the hypothesis $h(\boldsymbol{x})$ as

$$\begin{aligned} h(\boldsymbol{x}) &= \operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b) && (47) \\ &= \operatorname{sign}(\langle \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i, \boldsymbol{x} \rangle + b) && (48) \\ &= \operatorname{sign}(\sum_{i=1}^{m} \alpha_i y_i \langle \boldsymbol{x}_i, \boldsymbol{x} \rangle + b) && (49) \end{aligned}$$

SVM Hypothesis, revisited

Once we solve the dual problem with $\boldsymbol{\alpha}$, we have the solution of $\boldsymbol{w}$ as

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i \tag{46}$$

and the hypothesis $h(\boldsymbol{x})$ as

$$\begin{aligned} h(\boldsymbol{x}) &= \text{sign}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b) && (47) \\ &= \text{sign}(\langle \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i, \boldsymbol{x} \rangle + b) && (48) \\ &= \text{sign}(\sum_{i=1}^{m} \alpha_i y_i \langle \boldsymbol{x}_i, \boldsymbol{x} \rangle + b) && (49) \end{aligned}$$

- In addition, we also have $b = y_i - \sum_{i=1}^{m} \alpha_i y_i \langle \boldsymbol{x}_i, \boldsymbol{x} \rangle$ for any $\boldsymbol{x}_i$ with $\alpha_i > 0$
- Therefore, everything can be represented in the form of dot product

Kernel Methods

Properties of Inner Product

In the solution of SVMs

$$
\begin{aligned}
h(\boldsymbol{x}) &= \operatorname{sign}\Big(\sum_{i=1}^{m} \alpha_i y_i \langle \boldsymbol{x}_i, \boldsymbol{x} \rangle + b\Big) \\
b &= y_i - \sum_{i=1}^{m} \alpha_i y_i \langle \boldsymbol{x}_i, \boldsymbol{x} \rangle
\end{aligned} \tag{50}
$$

Properties of Inner Product

In the solution of SVMs

$$\begin{aligned} h(\boldsymbol{x}) &= \operatorname{sign}\left(\sum_{i=1}^{m} \alpha_i y_i \langle \boldsymbol{x}_i, \boldsymbol{x} \rangle + b\right) \\ b &= y_i - \sum_{i=1}^{m} \alpha_i y_i \langle \boldsymbol{x}_i, \boldsymbol{x} \rangle \end{aligned} \tag{50}$$

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

$$K(\boldsymbol{x}_i, \boldsymbol{x}) = \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}) \rangle \tag{51}$$

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

SVMs with Kernel Functions

- Problem definition

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$
$$\text{s.t. } \alpha_i \geq 0 \text{ and } \sum_{i=1}^{m} \alpha_i y_i = 0, i \in [m] \tag{52}$$

SVMs with Kernel Functions

- Problem definition

$$
\begin{aligned}
&\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \\
&\text{s.t. } \alpha_i \geq 0 \text{ and } \sum_{i=1}^{m} \alpha_i y_i = 0, i \in [m]
\end{aligned} \tag{52}
$$

- Solution: separable case

$$
h(\boldsymbol{x}) = \operatorname{sign}\left(\sum_{i=1}^{m} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b \right) \tag{53}
$$

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

Examples: Polynomial Kernels

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

$$K(\boldsymbol{x}, \boldsymbol{x}') = (\gamma \langle \boldsymbol{x}, \boldsymbol{x}' \rangle + c)^d, \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^n \tag{54}$$

Examples: Polynomial Kernels

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

$$K(\boldsymbol{x}, \boldsymbol{x}') = (\gamma \langle \boldsymbol{x}, \boldsymbol{x}' \rangle + c)^d, \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^n \tag{54}$$

Special cases

- $d = 1$: $K(\boldsymbol{x}, \boldsymbol{x}') = \gamma \langle \boldsymbol{x}, \boldsymbol{x}' \rangle + c$
- $d = 2$: $K(\boldsymbol{x}, \boldsymbol{x}') = (\gamma \langle \boldsymbol{x}, \boldsymbol{x}' \rangle + c)^2$

Examples: Polynomial Kernels (II)

For the special case with $d = 2$, assume $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^2$ (let $\gamma = 1$ for simplicity)

$$\begin{aligned}
K(\boldsymbol{x}, \boldsymbol{x}') &= (\langle \boldsymbol{x}, \boldsymbol{x}' \rangle + c)^2 && (55)\\
&= (x_1 x_1' + x_2 x_2' + c)^2 && (56)\\
&= x_1^2 x_1'^2 + x_1 x_2 x_1' x_2' + c x_1 x_1' + x_1 x_2 x_1' x_2' \\
&\quad + x_2^2 x_2'^2 + c x_2 x_2' + c x_1 x_1' + c x_2 x_2' + c^2 && (57)
\end{aligned}$$

Examples: Polynomial Kernels (II)

For the special case with $d = 2$, assume $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^2$ (let $\gamma = 1$ for simplicity)

$$
\begin{aligned}
K(\boldsymbol{x}, \boldsymbol{x}') &= (\langle \boldsymbol{x}, \boldsymbol{x}' \rangle + c)^2 && (55)\\
&= (x_1 x_1' + x_2 x_2' + c)^2 && (56)\\
&= x_1^2 x_1'^2 + x_1 x_2 x_1' x_2' + c x_1 x_1' + x_1 x_2 x_1' x_2' \\
&\quad + x_2^2 x_2'^2 + c x_2 x_2' + c x_1 x_1' + c x_2 x_2' + c^2 && (57)\\
&= x_1^2 x_1'^2 + x_2^2 x_2'^2 + 2 x_1 x'_1 x_2 x'_2 && (58)\\
&\quad + 2c x_1 x'_1 + 2c x_2 x'_2 + c^2 && (59)
\end{aligned}
$$

Examples: Polynomial Kernels (II)

For the special case with $d = 2$, assume $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^2$ (let $\gamma = 1$ for simplicity)

$$
\begin{aligned}
K(\boldsymbol{x}, \boldsymbol{x}') &= (\langle \boldsymbol{x}, \boldsymbol{x}' \rangle + c)^2 && (55) \\
&= (x_1 x_1' + x_2 x_2' + c)^2 && (56) \\
&= x_1^2 x_1'^2 + x_1 x_2 x_1' x_2' + c x_1 x_1' + x_1 x_2 x_1' x_2' \\
&\quad + x_2^2 x_2'^2 + c x_2 x_2' + c x_1 x_1' + c x_2 x_2' + c^2 && (57) \\
&= x_1^2 x_1'^2 + x_2^2 x_2'^2 + 2 x_1 x'_1 x_2 x'_2 && (58) \\
&\quad + 2c x_1 x'_1 + 2c x_2 x'_2 + c^2 && (59) \\
&= [x_1^2, x_2^2, \sqrt{2} x_1 x_2, \sqrt{2c} x_1, \sqrt{2c} x_2, c] \begin{bmatrix} x_1'^2 \\ x_2'^2 \\ \sqrt{2} x'_1 x'_2 \\ \sqrt{2c} x'_1 \\ \sqrt{2c} x'_2 \\ c \end{bmatrix}
\end{aligned}
$$

Examples: Polynomial Kernels (II)

For the special case with $d = 2$, assume $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^2$ (let $\gamma = 1$ for simplicity)

$$\begin{aligned}
K(\boldsymbol{x}, \boldsymbol{x}') &= (\langle \boldsymbol{x}, \boldsymbol{x}' \rangle + c)^2 && (55)\\
&= (x_1 x_1' + x_2 x_2' + c)^2 && (56)\\
&= x_1^2 x'^2_1 + x_1 x_2 x_1' x_2' + c x_1 x_1' + x_1 x_2 x_1' x_2' \\
&\quad + x_2^2 x'^2_2 + c x_2 x_2' + c x_1 x_1' + c x_2 x_2' + c^2 && (57)\\
&= x_1^2 x'^2_1 + x_2^2 x'^2_2 + 2 x_1 x'_1 x_2 x'_2 && (58)\\
&\quad + 2c x_1 x'_1 + 2c x_2 x'_2 + c^2 && (59)\\
&= [x_1^2, x_2^2, \sqrt{2} x_1 x_2, \sqrt{2c} x_1, \sqrt{2c} x_2, c] \begin{bmatrix} x'^2_1 \\ x'^2_2 \\ \sqrt{2} x'_1 x'_2 \\ \sqrt{2c} x'_1 \\ \sqrt{2c} x'_2 \\ c \end{bmatrix}
\end{aligned}$$

Exercise: Find out the $\Phi(\boldsymbol{x})$ function in $K(\boldsymbol{x}, \boldsymbol{x}') = (\langle \boldsymbol{x}, \boldsymbol{x}' \rangle + c)^3$

Examples: Polynomial Kernels (III)

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

$$\Phi(\boldsymbol{x}) = [x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2c}x_1, \sqrt{2c}x_2, c]^\top \tag{60}$$

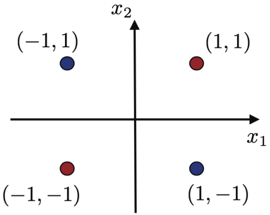
which maps a 2-D data point $\boldsymbol{x}$ into a 6-D space as $\Phi(\boldsymbol{x})$

Examples: Polynomial Kernels (III)

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

$$\Phi(\boldsymbol{x}) = [x_1^2, x_2^2, \sqrt{2} x_1 x_2, \sqrt{2c} x_1, \sqrt{2c} x_2, c]^\top \tag{60}$$

which maps a 2-D data point $\boldsymbol{x}$ into a 6-D space as $\Phi(\boldsymbol{x})$ Recall the XOR problem

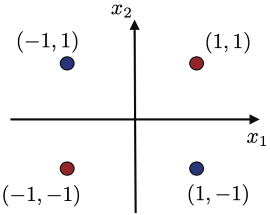


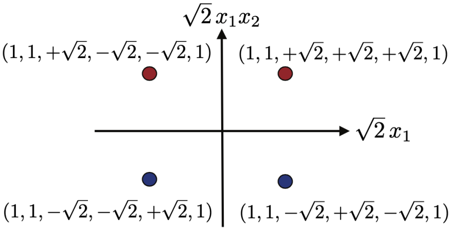
Examples: Polynomial Kernels (III)

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

$$\Phi(\boldsymbol{x}) = [x_1^2, x_2^2, \sqrt{2}x_1 x_2, \sqrt{2c}x_1, \sqrt{2c}x_2, c]^\mathsf{T} \tag{60}$$

which maps a 2-D data point $\boldsymbol{x}$ into a 6-D space as $\Phi(\boldsymbol{x})$ Recall the XOR problem



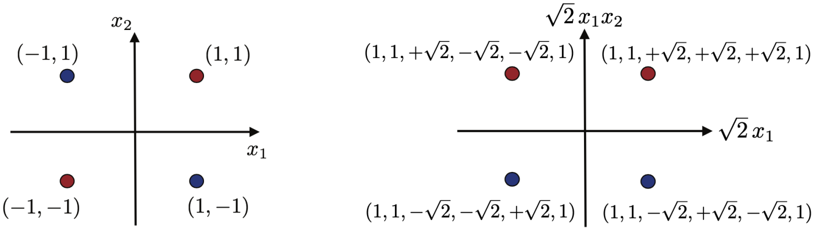


Examples: Polynomial Kernels (III)

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

$$\Phi(\boldsymbol{x}) = [x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2c}x_1, \sqrt{2c}x_2, c]^\top \tag{60}$$

which maps a 2-D data point $\boldsymbol{x}$ into a 6-D space as $\Phi(\boldsymbol{x})$ Recall the XOR problem

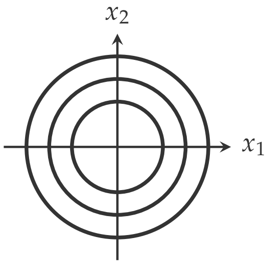


Try the online demo

Gaussian Kernels

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

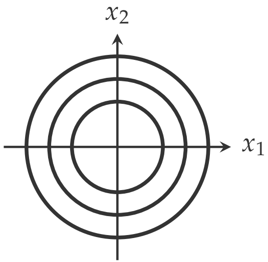
$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\gamma \|\boldsymbol{x}' - \boldsymbol{x}\|_2^2\right) \tag{61}$$



Gaussian Kernels

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

$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\gamma \|\boldsymbol{x}' - \boldsymbol{x}\|_2^2\right) \tag{61}$$



- What $\Phi(\boldsymbol{x})$ looks like in this case?
- What the effect of γ? (demo)

The Choice of Kernels

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

[Mohri et al., 2018, Section 6.1 - 6.2]

The Choice of Kernels

- The choice of $K(\boldsymbol{x}, \boldsymbol{x}')$ 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 $K(\boldsymbol{x}, \boldsymbol{x}')$ is *positive definite symmetric* (PDS)
  - A kernel K is PDS if for any $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ the matrix $\mathbf{K}$ is symmetric positive semi-definite

$$\mathbf{K} = [K(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j} \in \mathbb{R}^{m\times m} \tag{62}$$

[Mohri et al., 2018, Section 6.1 - 6.2]

The Choice of Kernels

- The choice of $K(\boldsymbol{x}, \boldsymbol{x}')$ 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 $K(\boldsymbol{x}, \boldsymbol{x}')$ is *positive definite symmetric* (PDS)
  - A kernel K is PDS if for any $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_m\}$ the matrix $\mathbf{K}$ is symmetric positive semi-definite

$$\mathbf{K} = [K(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j} \in \mathbb{R}^{m \times m} \tag{62}$$

- A symmetric positive semi-definite matrix is defined as

$$\boldsymbol{c}^{\top} \mathbf{K} \boldsymbol{c} \geq 0 \tag{63}$$

[Mohri et al., 2018, Section 6.1 - 6.2]

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.