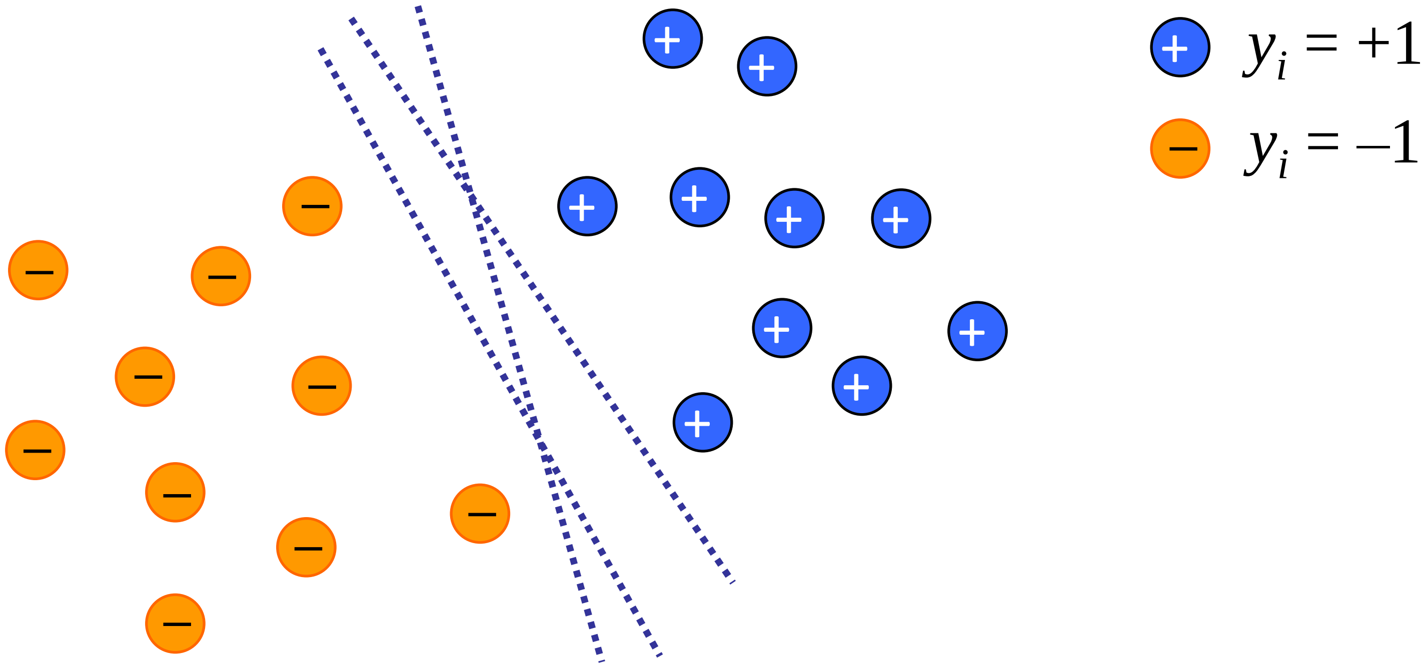
Chapter 7 – SVM

Instructor: Dr. Hongfu Liu
Email: hongfuliu@brandeis.edu

Main source: https://www-users.cs.umn.edu/~kumar001/dmbook/index.php

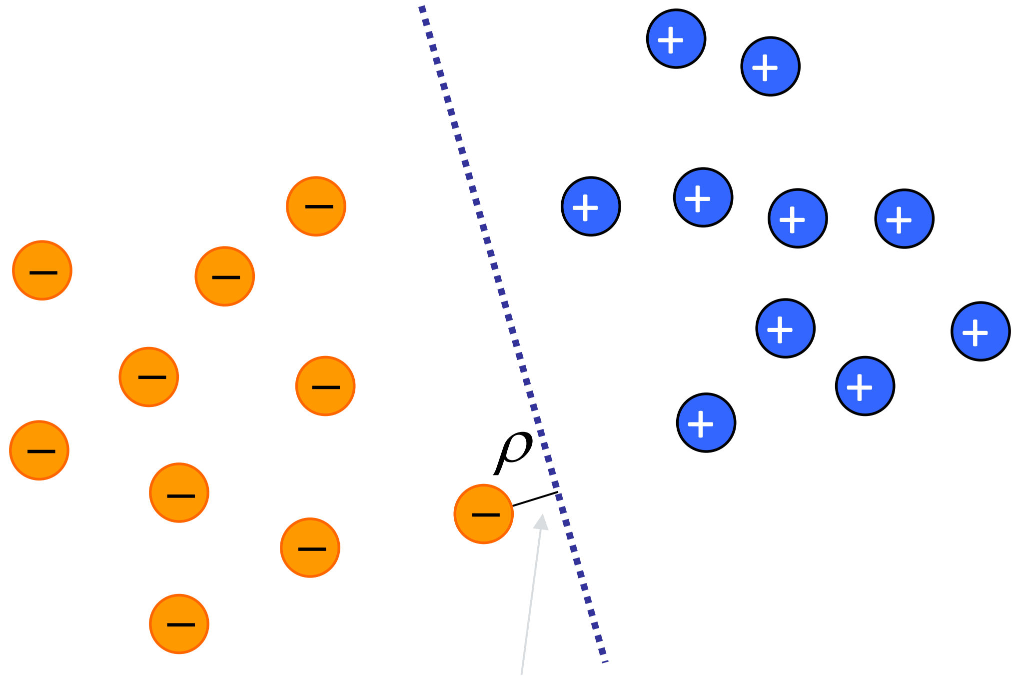
Linearly Separable Patterns

$\{\langle \vec{x}_i, y_i \rangle\}_{i=1}^{N}$



Potentially, indefinite hyperplane

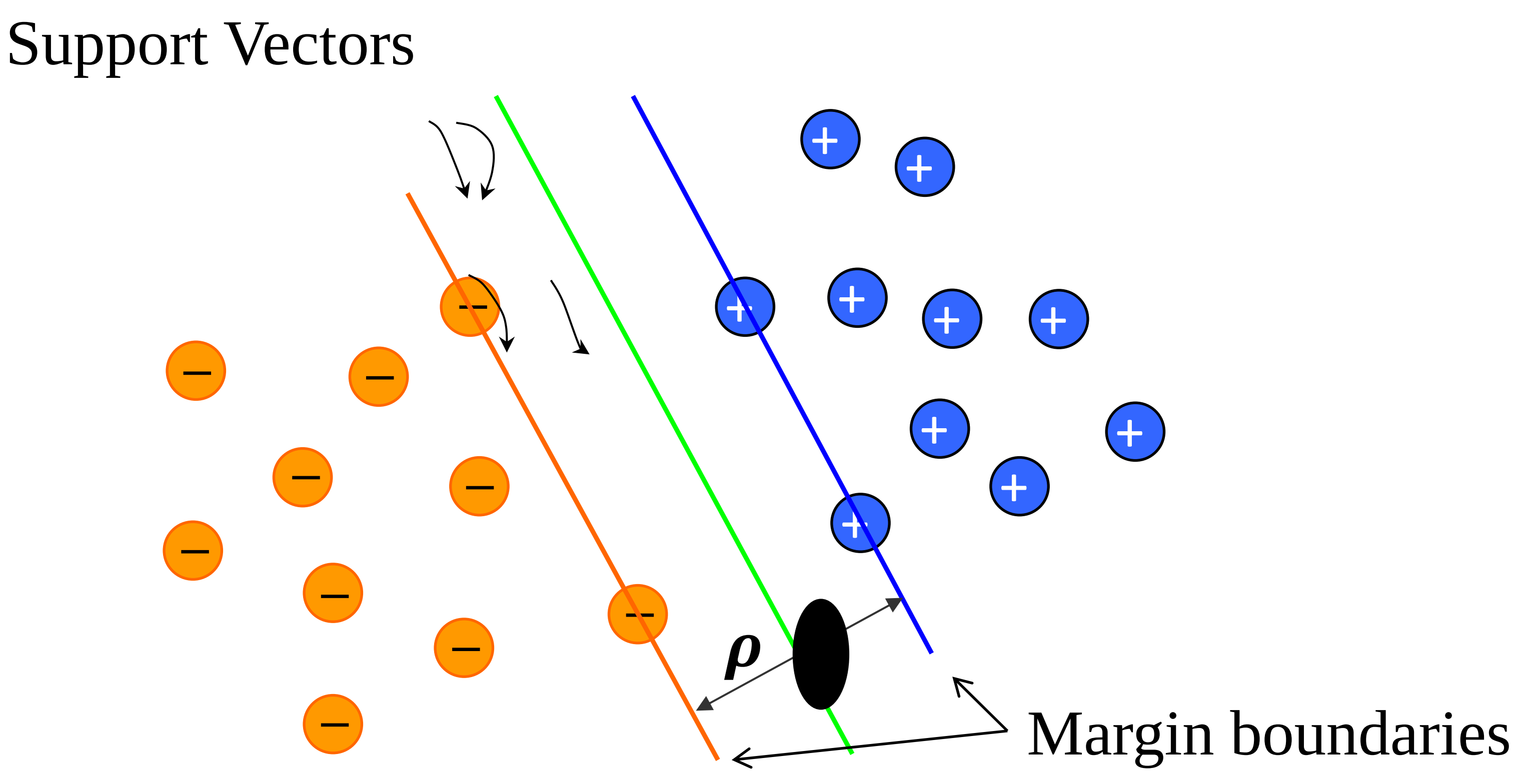
Linearly Separable Patterns



Margin of separation: the separation between the hyperplane and the closest data point.

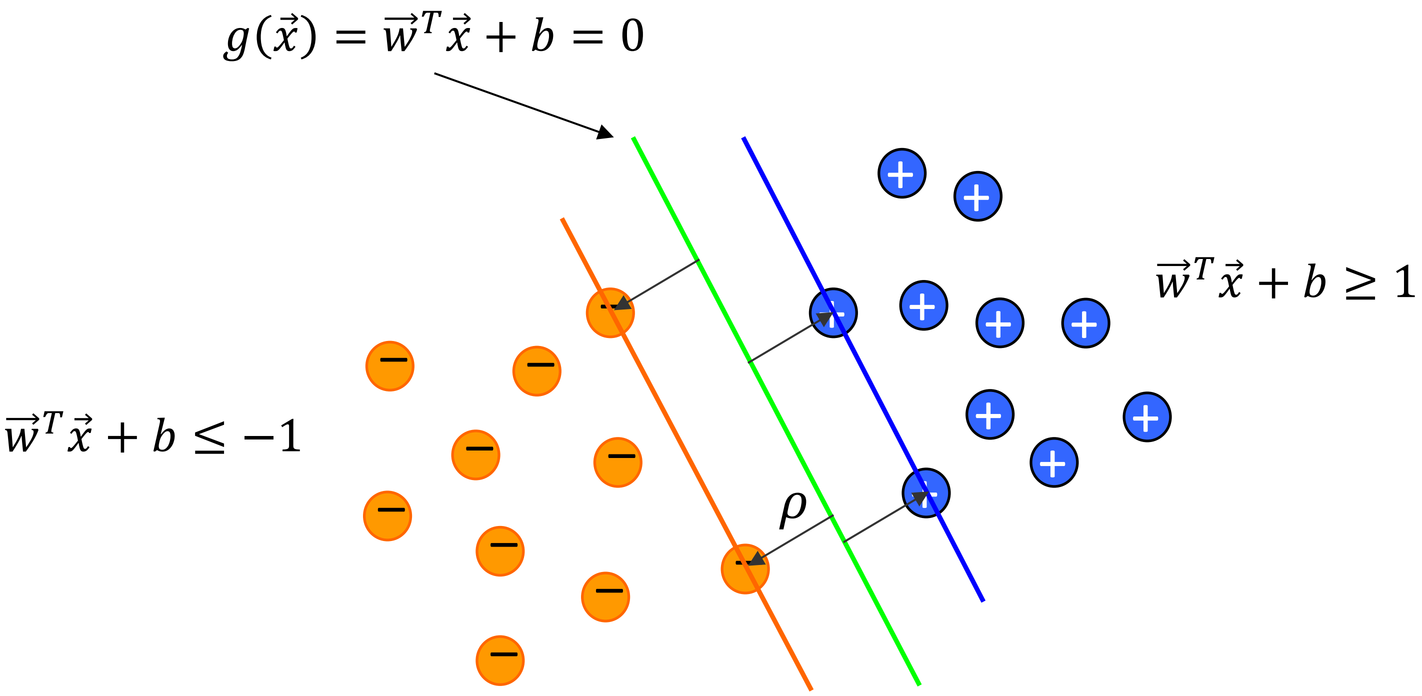
SVM – Geometric Interpretation

SVM Learning finds a separation with the maximal margin



$\boldsymbol{\rho}$: Margin of separation

SVM – Mathematical Formulation



A Support Vector $\vec{x}^{(s)}$ satisfies $r^{(s)} = \frac{g(\vec{x}^{(s)})}{\|\vec{w}\|} = \begin{cases} \frac{1}{\|\vec{w}\|} & \text{if } y_i = +1 \\ \frac{-1}{\|\vec{w}\|} & \text{if } y_i = -1 \end{cases}$

SVM Learning – Linearly Separable Patterns

Training data: $\{\langle \vec{x}_i, y_i \rangle\}_{i=1}^{N}$

Find the optimal hyperplane by minimizing the cost function

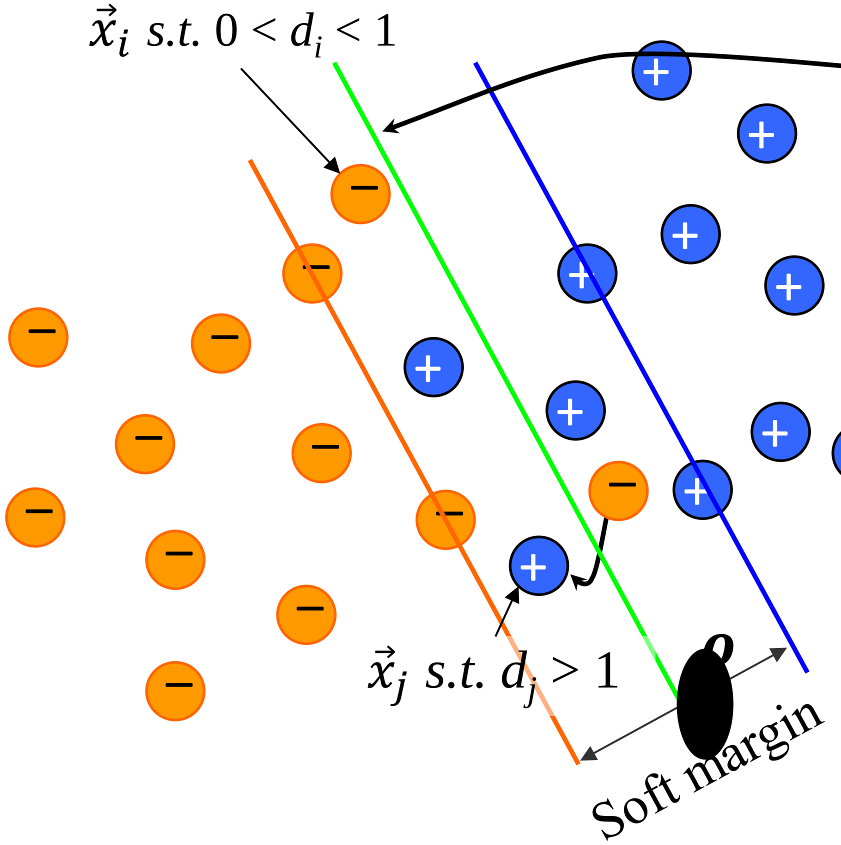
$$C(\vec{w}) = \frac{1}{2}\|\vec{w}\|^2 = \frac{1}{2}\vec{w}^T\vec{w}$$

subject to the constraints

$$g(\vec{x}_i)y_i = (\vec{w}^T\vec{x}_i + b)y_i \geq 1$$ for i = 1, …, N

Objective function: $\frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{N} \alpha_i [y_i(\vec{w}^T\vec{x}_i + b) - 1]$

SVM – Linearly Non-separable Patterns



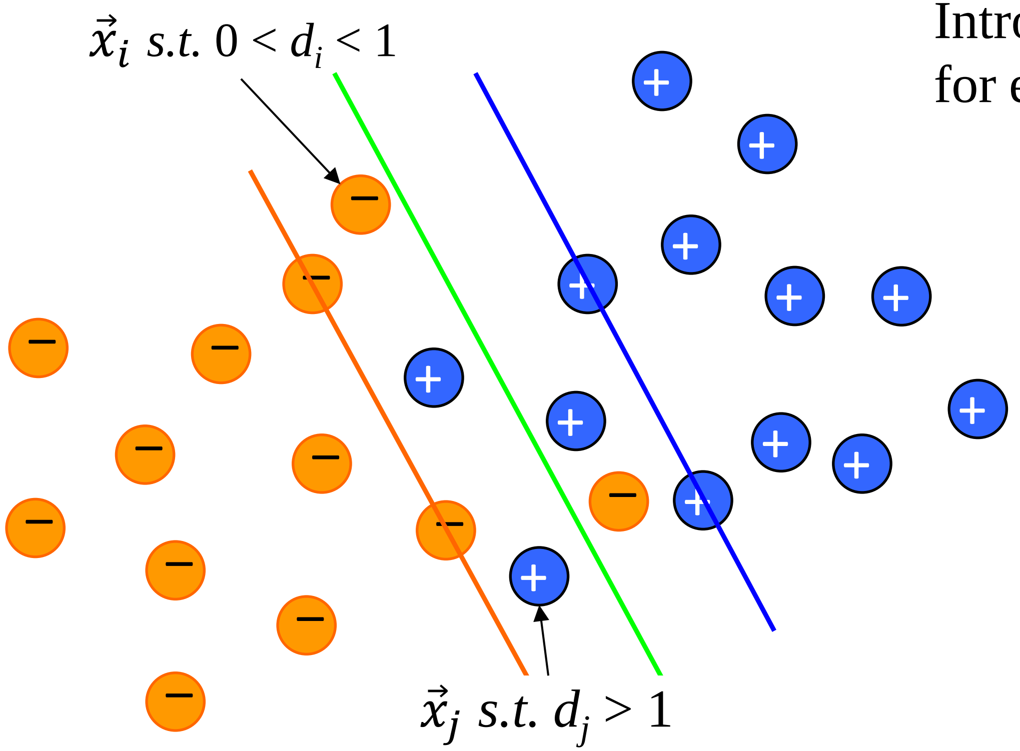
(a) A data point falls inside the region of separation but on the right side of the decision surface.

(b) A data point falls on the wrong side of the decision surface.

In such a case, it is impossible to construct a separating hyperplane without misclassifications

Construct an optimal hyperplane that minimizes the probability of classification error.

SVM – Linearly Non-separable Patterns



Introduce a non-negative slack variable d_i for each sample $\langle \vec{x}_i, y_i \rangle$

$$y_i(\vec{w}^T \vec{x}_i + b) \geq 1 - d_i$$

d_i measures the deviation of $\vec{x}_i$ from the ideal separation

Find a separating hyperplane ($\vec{w}$ & b) that minimizes

$$\sum_{i=1}^{N} \mathbf{I}(d_i - 1)$$

$$\mathbf{I}(d_i - 1) = \begin{cases} 0 & d_i \leq 1 \\ 1 & \text{otherwise} \end{cases}$$

Subject to the constraints $y_i(\vec{w}^T \vec{x}_i + b) \geq 1 - d_i$ for $i = 1, \ldots, N$

NP-complete problem !

SVM – Linearly Non-separable Patterns

Simplify the computation by formulating the minimization problem as

$$\Phi(\vec{w}, b, \{d_i\}) = \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{N} d_i$$

Subject to $y_i(\vec{w}^T\vec{x}_i + b) \geq 1 - d_i$ for $i = 1, \ldots, N$

$\sum_{i=1}^{N} d_i$ is the upper bound on the number of misclassifications.

More general $\sum_{i=1}^{N} h(d_i)$

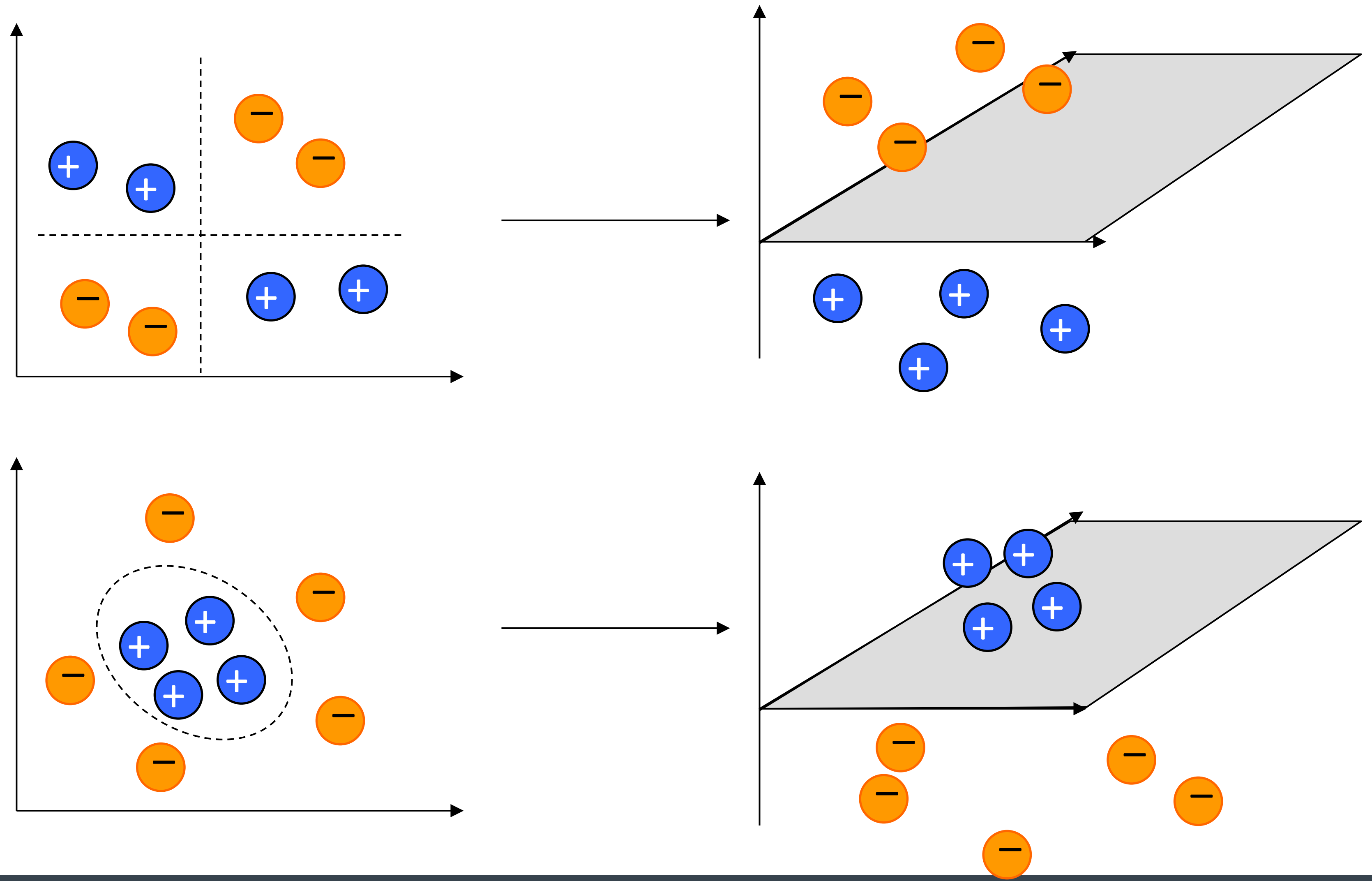
C to be decided by the user. It controls the trade-off between minimizing training errors and controlling model complexity.

The above problem can be converted into its dual problem that maximizes

$$Q(\{\alpha_i\}) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i=1}^{N} \alpha_i\alpha_j y_i y_j \vec{x}_i{}^T\vec{x}_j$$

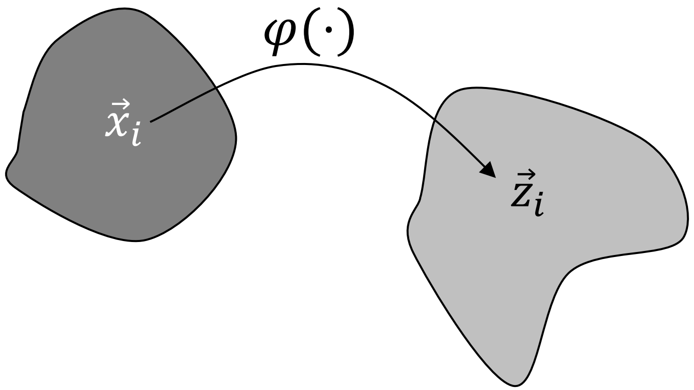
Subject to $\sum_{i=1}^{N}\alpha_i y_i = 0 \quad \& \quad 0 \leq \alpha_i \leq C$ for $i = 1, 2, \ldots, N$

Feature Spaces and Kernels



Linearly Separable in Higher Dimension

Inner-Product Kernel



Inner-product kernel

$$K\left(\vec{x}_i, \vec{x}_j\right) = \vec{z}_i^T \vec{z}_j = \varphi(\vec{x}_i)^T \varphi(\vec{x}_j)$$

Kernel SVM

$$\text{Maximize } Q(\{\alpha_i\}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(\vec{x}_i, \vec{x}_j)$$

Subject to $\sum_{i=1}^{N} \alpha_i y_i = 0$

$0 \leq \alpha_i \leq C$ for $i = 1, 2, \ldots, N$

Supporter Vectors are found in the space defined by $\varphi(\vec{x})$ without going into that space

Classify new sample $$y(\vec{x}) = \sum_{i=1}^{N} \alpha_i y_i K(\vec{x}, \vec{x}_i) + b$$

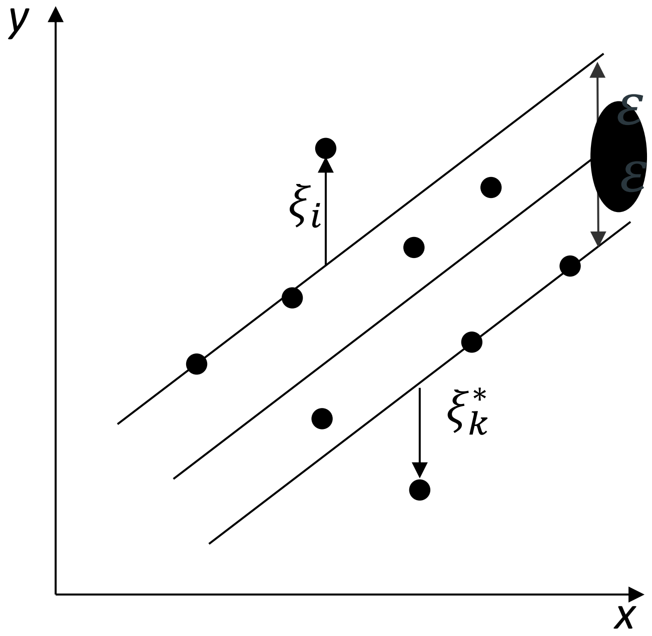
For comparison, the objective function without using kernels is $$Q(\{\alpha_i\}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i=1}^{N} \alpha_i \alpha_j y_i y_j \vec{x}_i{}^T \vec{x}_j$$

Examples of Inner-Product Kernels

Kernel	Inner-product Kernel
Polynomial function	$\left(s\vec{x}_i^T\vec{x}_j + t\right)^d$
Radial-basis function	$\exp\left(\frac{-1}{2\sigma^2}\left\Vert\vec{x}_i - \vec{x}_j\right\Vert^2\right)$
Hyperbolic tangent	$tanh\left(s\vec{x}_i^T\vec{x}_j + t\right)$

Support Vector Regression

The goal is to find the flattest ε-tube around the hyperplane that contains the most training samples. Define ε as the margin.



$$\text{Minimize} \quad \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$$

subject to the constraints

for $i = 1, \ldots, N$

$$y_i - (\vec{w}^T\vec{x}_i + b) \le \varepsilon + \xi_i$$

$$(\vec{w}^T\vec{x}_i + b) - y_i \le \varepsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \ge 0$$

Skicit Learn


Q
&
A