机器学习引论

彭玺
pengxi@scu.edu.cn
www.pengxi.me

提纲

一．Review

二．Kernel and Nonlinear SVM

提纲

一．Review

二．Kernel and Nonlinear SVM

Review - Two Limitations of KNN

- Prob: It do not learn knowledge from training data

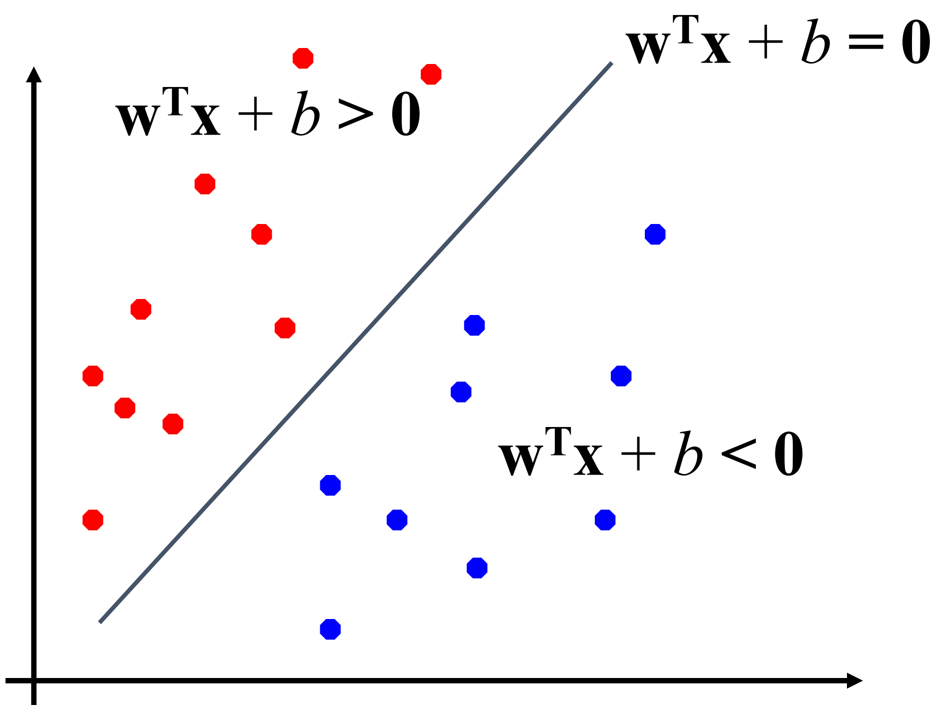
- Prob: It requires that the data come from the Euclidean space so that the obtained neighbors and the data point itself come from the same subject.

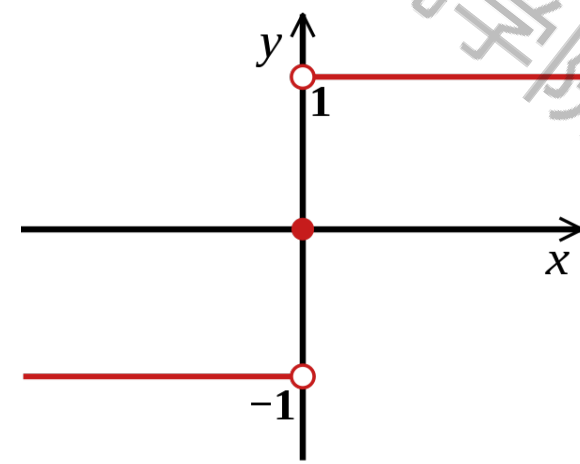
Review - Two Limitations of KNN

- Prob: It do not learn knowledge from training data
- Sol: Perceptron -> Linear SVM
- Prob: It requires that the data come from the Euclidean space so that the obtained neighbors and the data point itself come from the same subject.
- Sol: Kernel + SVM = Nonlinear SVM

Review

- Binary classification can be viewed as the task of separating classes in feature space:



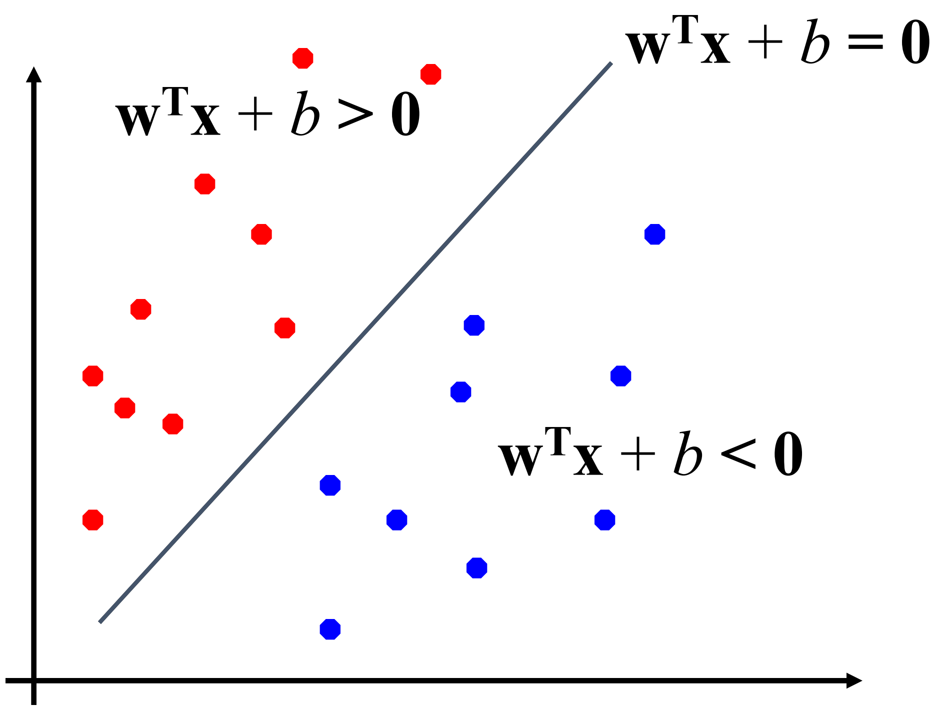


$$\operatorname{sgn}(x) := \begin{cases} -1 & \text{if } x < 0, \\ 0 & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases}$$

Activate function

Review

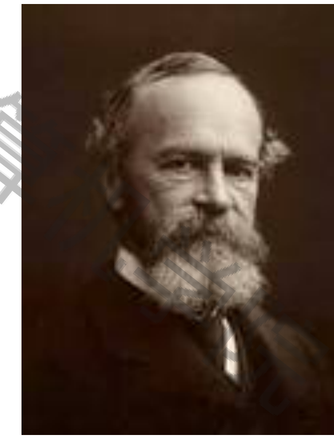
- Binary classification can be viewed as the task of separating classes in feature space:

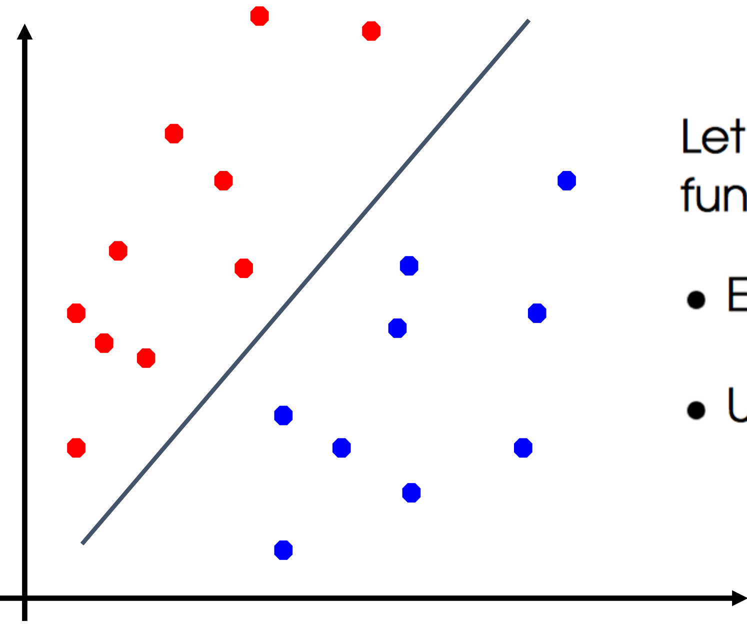


$$y = \mathrm{sign}(\mathbf{w}^{\mathrm{T}}\mathbf{x} + b)$$

Review

- 1890：美国心理学家和哲学家William James在其著作中指出——当两个事件同时发生时，涉及到的大脑过程间的连接将会增强，这是无监督的Hebb学习规则的灵感来源。此外，James还提出了加权（weighted）、可变（modifiable）、及并行连接（parallel connections）等神经网络至今采用的基本概念。

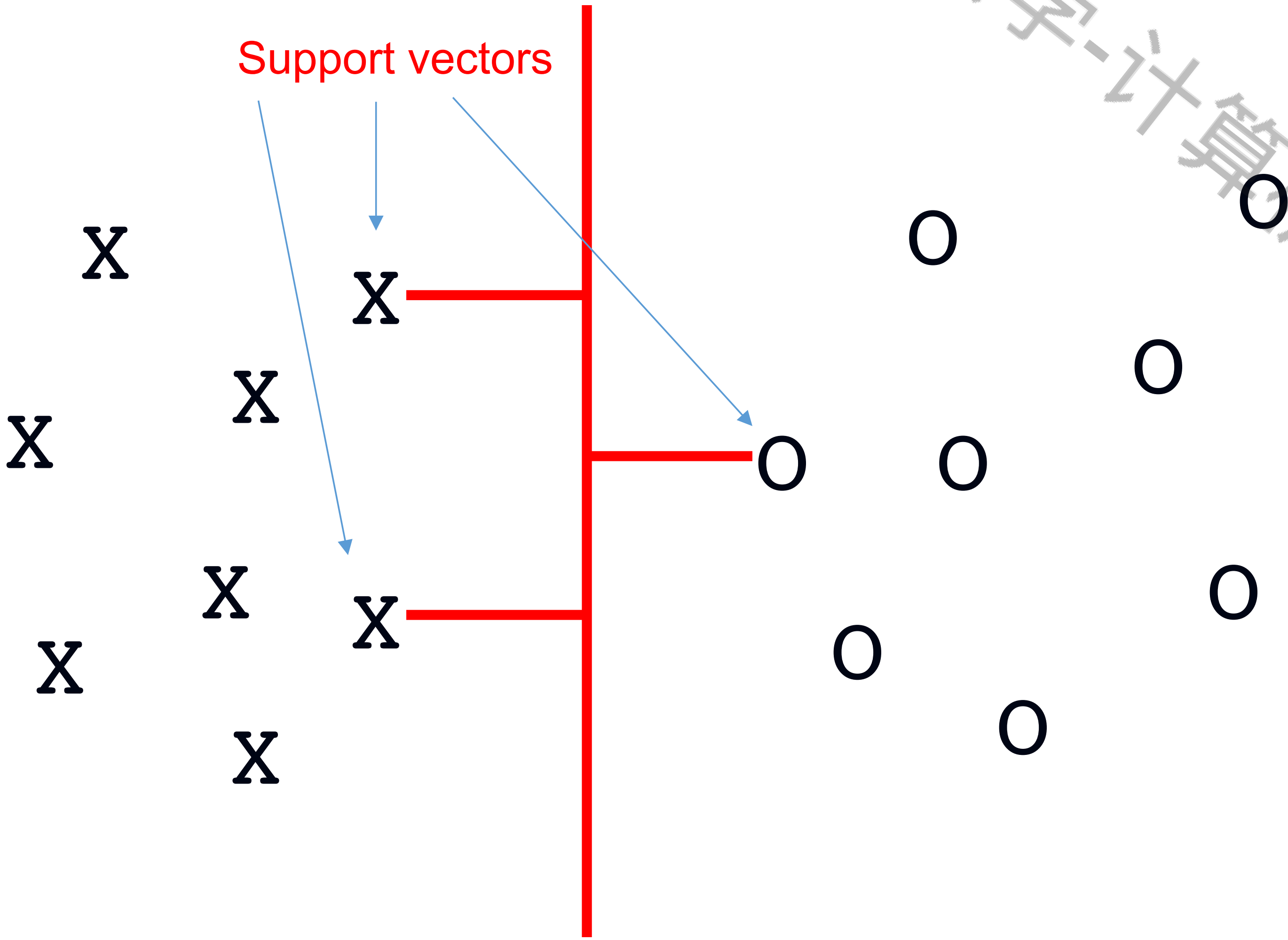




Let y be the correct output, and $f(x)$ the output function of the network.

- Error: $E = y - f(x)$
- Update weights: $W_j \leftarrow W_j + \alpha x_j E$

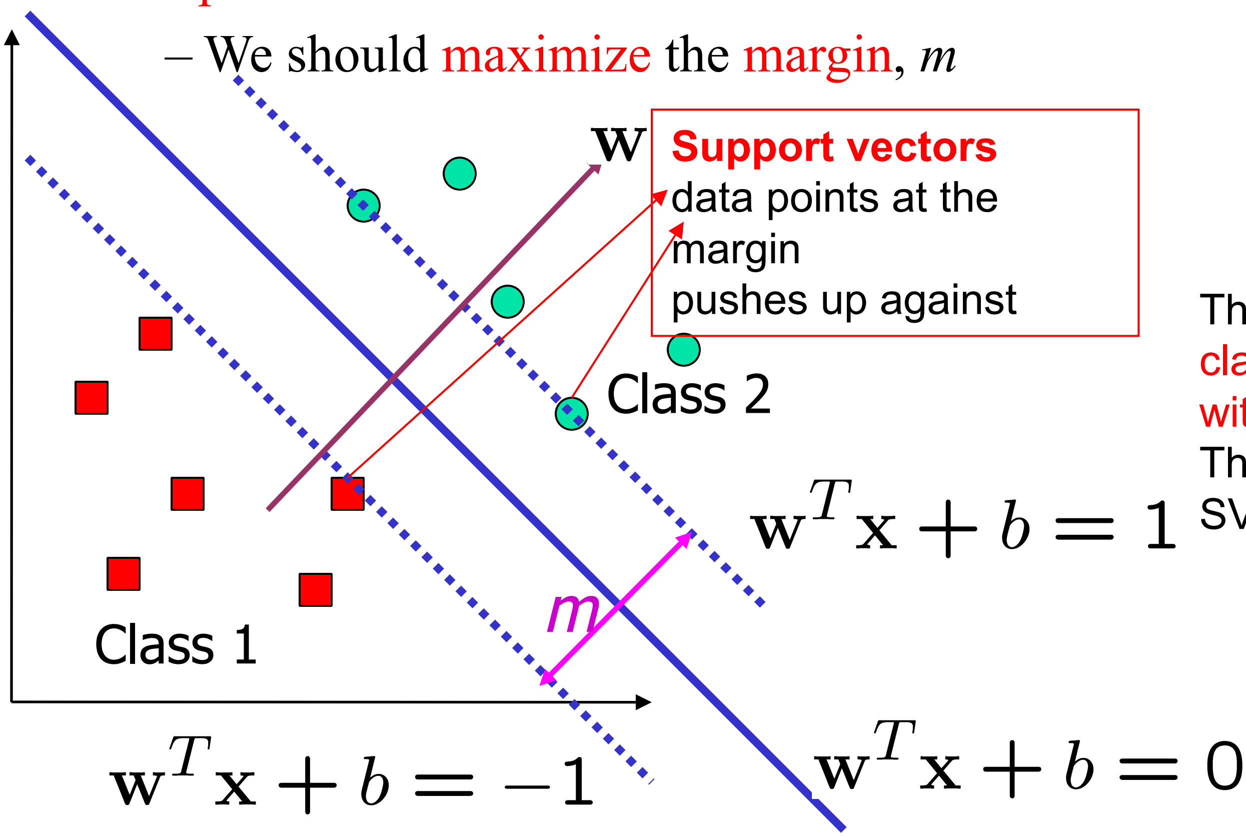
Review



Review

The decision boundary should be as far away from the data of both classes as possible

– We should maximize the margin, m



The maximum margin linear classifier is the linear classifier with the maximum margin. This is the simplest kind of SVM (Called an Linear SVM)

Review

- Let training set $\{(\mathbf{x}_i, y_i)\}_{i=1..n}$, $\mathbf{x}_i \in \mathrm{R}^d$, $y_i \in \{-1, 1\}$ be separated by a hyperplane with margin m. Then for each training example $(\mathbf{x}_i, y_i)$:

$$\begin{array}{ll} \mathbf{w}^{\mathbf{T}}\mathbf{x}_i + b \le -m/2 & \text{if } y_i = -1 \\ \mathbf{w}^{\mathbf{T}}\mathbf{x}_i + b \ge m/2 & \text{if } y_i = 1 \end{array} \quad \Leftrightarrow \quad y_i(\mathbf{w}^{\mathbf{T}}\mathbf{x}_i + b) \ge m/2$$

- For every support vector $\mathbf{x}_s$, the above inequality is an equality. After rescaling $\mathbf{w}$ and b by $m/2$ in the equality, we obtain that distance between each $\mathbf{x}_s$ and the hyperplane is $r = \frac{\mathrm{y}_s(\mathbf{w}^T\mathbf{x}_s + b)}{\|\mathbf{w}\|} = \frac{1}{\|\mathbf{w}\|}$

- Then the margin can be expressed through (rescaled) $\mathbf{w}$ and b as:

$$m = 2r = \frac{2}{\|\mathbf{w}\|}$$

Review

- Then we can formulate the *quadratic optimization problem:*

Find $\mathbf{w}$ and b such that

$$\rho = \frac{2}{\|\mathbf{w}\|} \quad \text{is maximized}$$

and for all $(\mathbf{x}_i, y_i)$, i=1..n : $\quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$

Which can be reformulated as:

Find $\mathbf{w}$ and b such that

$$\mathbf{\Phi}(\mathbf{w}) = ||\mathbf{w}||^2 = \mathbf{w}^T\mathbf{w} \quad \text{is minimized}$$

and for all $(\mathbf{x}_i, y_i)$, i=1..n : $\quad y_i\,(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$

Review

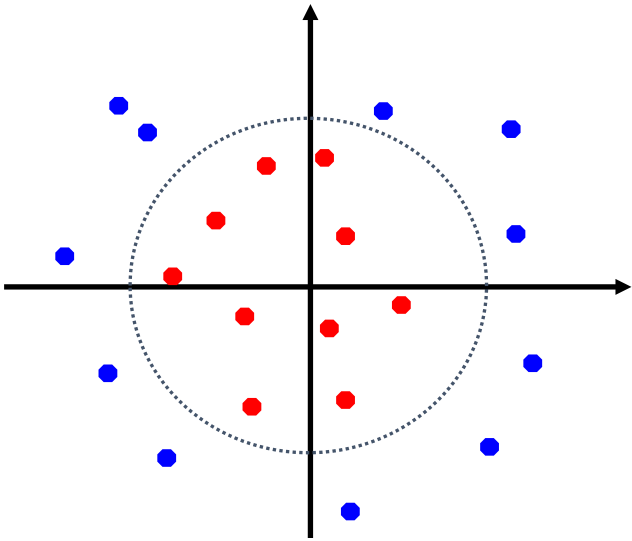
The objective function:

Minimize $\frac{1}{2}||\mathbf{w}||^2$

subject to $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \qquad \forall i$

Review - Two Limitations of KNN

- Prob: It do not learn knowledge from training data
- Sol: Perceptron -> Linear SVM
- Prob: It requires that the data come from the Euclidean space so that the obtained neighbors and the data point itself come from the same subject.
- Sol: Kernel + SVM = Nonlinear SVM



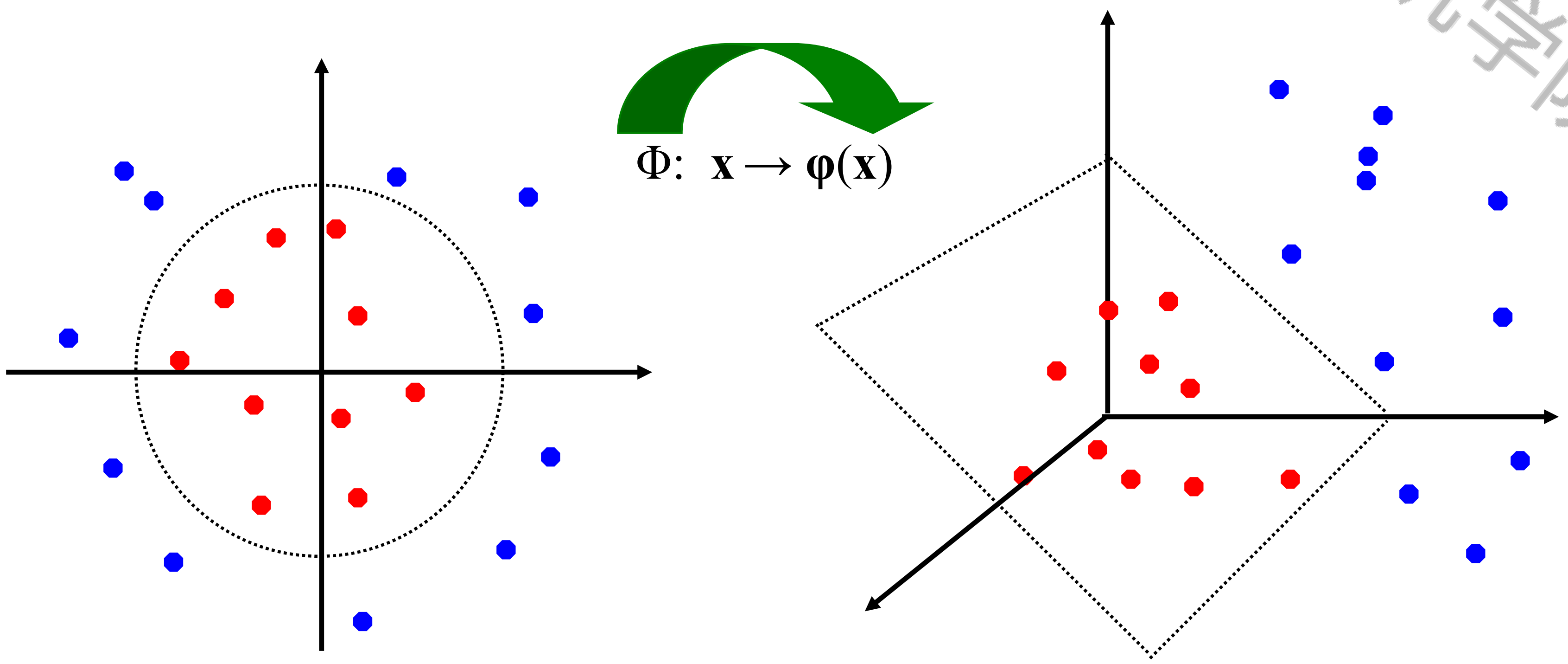
提纲

一．Review

二．Kernel and Nonlinear SVM

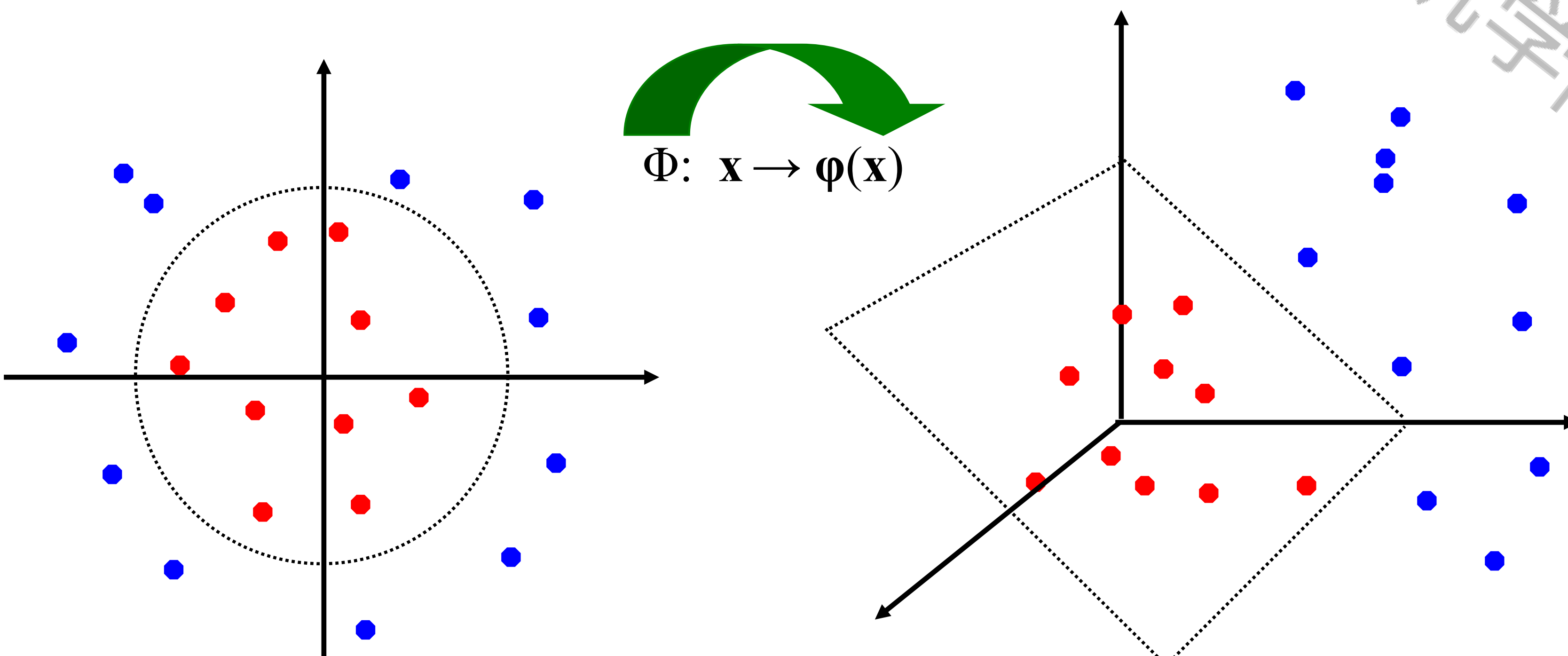
二、 Kernel and Nonlinear SVM

- General idea: the original feature space can always be mapped to some higher-dimensional feature space where the training set is separable:



二、 Kernel and Nonlinear SVM

- General idea: the original feature space can always be mapped to some higher-dimensional feature space where the training set is separable:



Data points are linearly separable

in the space $(x_1^2, x_2^2, \sqrt{2}x_1x_2)$

二、 Kernel and Nonlinear SVM

Lagrangian of Original Problem

$$\text{Minimize } \frac{1}{2}||\mathbf{w}||^2$$

$$\text{subject to } 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b) \le 0 \qquad \text{for } i = 1, \ldots, n$$

The Lagrangian is

Lagrangian multipliers

$$\mathcal{L} = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{i=1}^{n} \alpha_i \left(1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)\right)$$

Note that $||\mathbf{w}||^2 = \mathbf{w}^T\mathbf{w}$

Setting the gradient of $\mathcal{L}$ w.r.t. $\mathbf{w}$ and b to zero, we have

$$\mathbf{w} + \sum_{i=1}^{n} \alpha_i(-y_i)\mathbf{x}_i = \mathbf{0} \quad \Rightarrow \quad \boxed{\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i}$$

$$\boxed{\sum_{i=1}^{n} \alpha_i y_i = 0}$$

$\alpha_i \ge 0$

二、 Kernel and Nonlinear SVM

The Dual Optimization Problem

We can transform the problem to its dual

Dot product of X

$$\max.\ W(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{subject to } \alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i y_i = 0$$

α's → New variables
(Lagrangian multipliers)

This is a convex quadratic programming (QP) problem

– Global maximum of α_i can always be found

→well established tools for solving this optimization problem (e.g. cplex)

Note:

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

General Idea: Lagrange Optimization

$$
\begin{aligned}
&\min_{w} f(w) \\
&\text{s.t.} \quad g_i(w) \leq 0, \quad i = 1, \dots, k \\
&h_i(w) = 0, \quad i = 1, \dots, l
\end{aligned}
$$

1) Formulate Lagrangian function (primal problem)

$L_p(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$

2) Minimize Lagrangian wrt primal variable w: $\frac{\vartheta L_p(w, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\vartheta w} = 0$

3) Substitute the primal variable w and express Lagrangian wrt dual variables α_i, β_i: $L_d(\boldsymbol{\alpha}, \boldsymbol{\beta})$

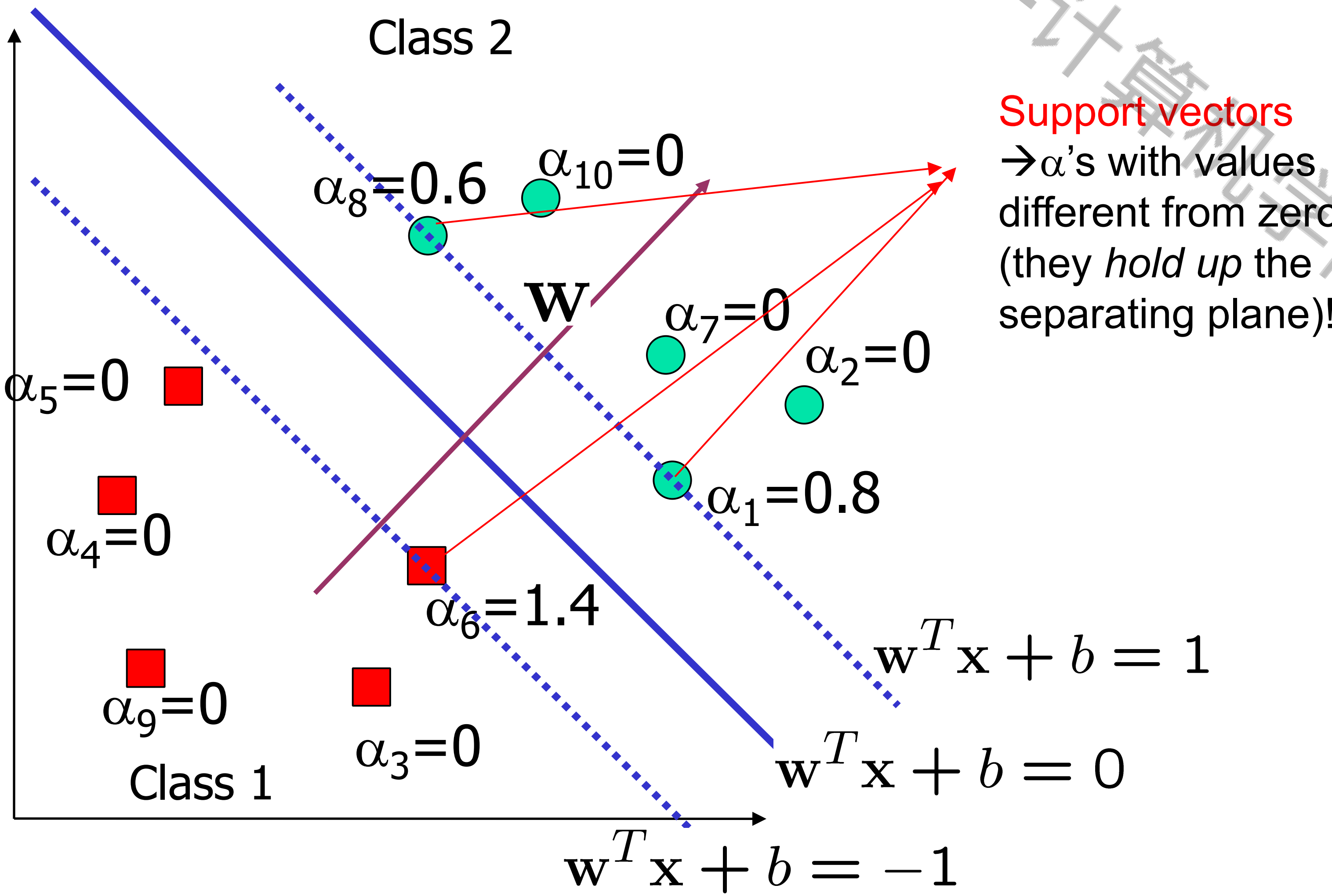
4) Maximize the Lagrangian with respect to dual variables and solve for dual variables (dual problem)

ϑw

5) Recover the solution (for the primal variables) from the dual variables

二、 Kernel and Nonlinear SVM

A Geometrical Interpretation



二、 Kernel and Nonlinear SVM

Recall:

Note that data only appears as dot products

$$\begin{aligned} \max \quad & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=j=1}^{N} \alpha_i \alpha_j y_i y_j x_i x_j \\ \text{s.t.} \quad & C \geq \alpha_i \geq 0, \sum_{i=1}^{N} \alpha_i y_i = 0 \end{aligned}$$

Since data is only represented as dot products, we need not do the mapping explicitly.

Introduce a Kernel Function (*) *K* such that:

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$$

(*)Kernel function – a function that can be applied to pairs of input data to evaluate dot products in some corresponding feature space

二、 Kernel and Nonlinear SVM

Consider the following transformation

$$\phi(\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right]) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2)$$
$$\phi(\left[\begin{smallmatrix} y_1 \\ y_2 \end{smallmatrix}\right]) = (1, \sqrt{2}y_1, \sqrt{2}y_2, y_1^2, y_2^2, \sqrt{2}y_1y_2)$$

Define the kernel function *K* (**x**,**y**) as

$$\begin{aligned}\langle \phi(\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right]), \phi(\left[\begin{smallmatrix} y_1 \\ y_2 \end{smallmatrix}\right]) \rangle &= (1 + x_1y_1 + x_2y_2)^2 \\ &= K(\mathrm{x}, \mathrm{y})\end{aligned}$$

$$K(\mathbf{x}, \mathbf{y}) = (1 + x_1y_1 + x_2y_2)^2$$

The inner product ϕ(.)ϕ(.) can be computed by *K* without going through the map ϕ(.) explicitly!!!

二、 Kernel and Nonlinear SVM

Kernel SVM

Change all inner products to kernel functions

For training,

Original

$$\text{max.} \quad W(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{subject to } C \geq \alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i y_i = 0$$

With kernel function

$$\text{max.} \quad W(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{subject to } C \geq \alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i y_i = 0$$

二、 Kernel and Nonlinear SVM

- Linear: $K(\mathbf{x}_i,\mathbf{x}_j)= \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j$
  - Mapping Φ: $\mathbf{x} \rightarrow \boldsymbol{\varphi}(\mathbf{x})$, where $\boldsymbol{\varphi}(\mathbf{x})$ is $\mathbf{x}$ itself

- Polynomial of power p: $K(\mathbf{x}_i,\mathbf{x}_j)= (1+ \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j)^p$
  - Mapping Φ: $\mathbf{x} \rightarrow \boldsymbol{\varphi}(\mathbf{x})$, where $\boldsymbol{\varphi}(\mathbf{x})$ has $\binom{d+p}{p}$ dimensions

- Gaussian (radial-basis function): $K(\mathbf{x}_i,\mathbf{x}_j) = e^{-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{2\sigma^2}}$
  - Mapping Φ: $\mathbf{x} \rightarrow \boldsymbol{\varphi}(\mathbf{x})$, where $\boldsymbol{\varphi}(\mathbf{x})$ is *infinite-dimensional*: every point is mapped to *a function* (a Gaussian); combination of functions for support vectors is the separator.

- Higher-dimensional space still has *intrinsic* dimensionality d (the mapping is not *onto*), but linear separators in it correspond to *non-linear* separators in original space.

Test Questions

- Q1: The evolution of neuron from biology to mathematics?
- Q2: The key concepts of Perceptron and its limitations.
- Q3: Who is Vladimir N. Vapnik and what is his major contribution?
- Q4: Maximum Margin Principle and why support vector is important?
- Q5: How to compute the distance between a given data point to the decision boundary?
- Q6: What limitations the linear SVM suffered from?
- Q7: Why dual form of SVM is important?
- Q8: How to derive the dual form from the prime form of SVM?
- Q9: What the limitations the kernel method suffers from?
- Q10: the relation between Perception and SVM.
- Q11: are there other methods to address linear inseparable issue besides kernel?

Others

Further Reading

- What is hard margin in SVM?
- What is soft margin in SVM? Why soft-margin SVM is regarded better than hard one?

Q&A

THANKS!