



# **The Introduction To Artificial Intelligence**

**Yuni Zeng [yunizeng@zstu.edu.cn](mailto:yunizeng@zstu.edu.cn)  
2022-2023-1**

# The Introduction to Artificial Intelligence

- Part I Brief Introduction to AI & Different AI tribes
- Part II Knowledge Representation & Reasoning
- Part III AI GAMES and Searching
- Part IV Model Evaluation and Selection
- ✚ Part V Machine Learning

# Machine Learning



Supervised  
learning

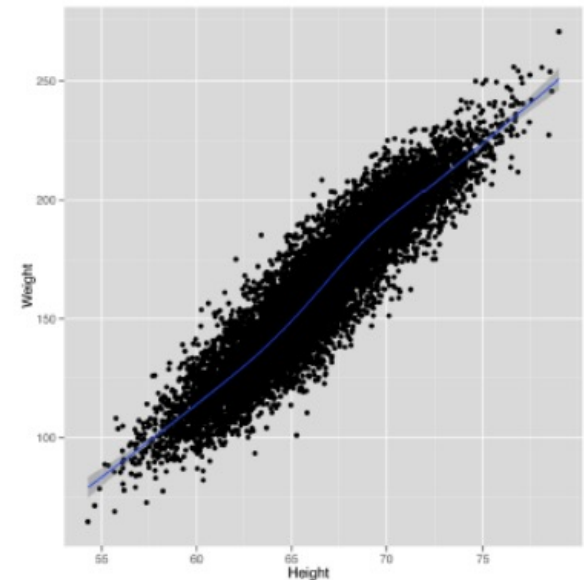
Unsupervised  
learning

Reinforcement  
learning

# Linear Regression

## □ What is regression?

Regression is to relate **input variables** to the **output variable**, to either **predict** outputs for new inputs and/or to **interpret** the effect of the input on the output.



Height is correlated with weight.

# Supervised learning

- *Linear Regression*
- Logistic Regression
- Classification
  - Distance-based algorithms
  - Linear classifiers
  - Other classifiers
- .....

# Linear Regression

---

## □ Linear Regression Model

- Only **one independent variable**,  $x$
- Relationship between  $x$  and  $y$  is described by a **linear function**
- Changes in  $y$  are assumed to be related to changes in  $x$

# Linear Regression

## □ Linear Regression Model

The diagram illustrates the Linear Regression Model equation,  $y_i = b_0 + b_1 x_i + \epsilon_i$ , with labels and components:

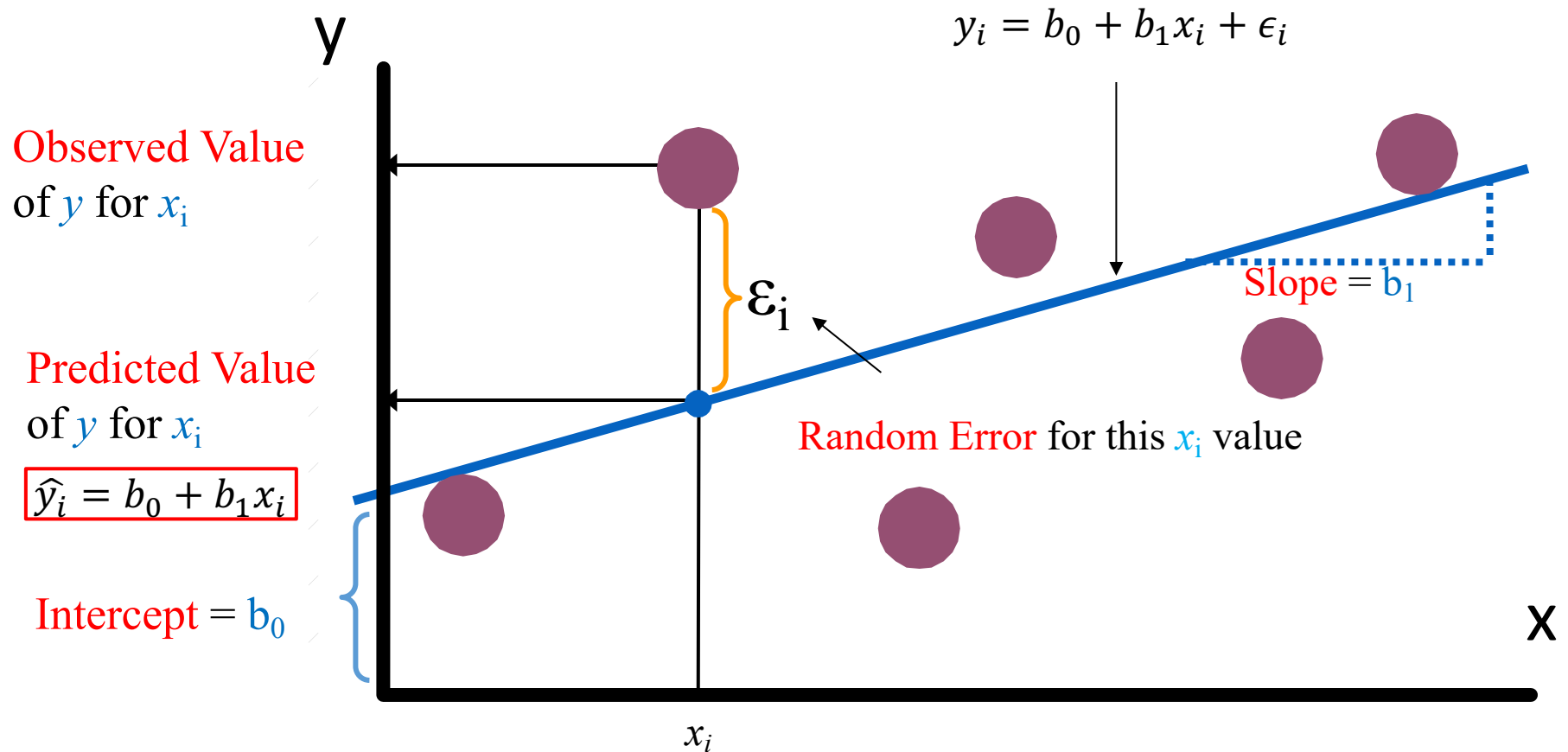
- Dependent Variable**: Points to  $y_i$ .
- intercept**: Points to  $b_0$ .
- Slope Coefficient**: Points to  $b_1$ .
- Independent Variable**: Points to  $x_i$ .
- Random Error term**: Points to  $\epsilon_i$ .

The equation is displayed in a light orange box. Below the equation, two curly braces identify the components:

- Linear component**: Brackets the terms  $b_0 + b_1 x_i$ .
- Random Error component**: Brackets the term  $\epsilon_i$ .

# Linear Regression

## □ Linear Regression Model



Question: How to obtain the best line?



# Linear Regression

## □ The Least Squares Method

$b_0$  and  $b_1$  are obtained by finding the values of that minimize the **sum** of the squared **differences** between  $y_i$  and  $\hat{y}_i$  **for all  $i$**  :

$$\min \sum (y_i - \hat{y}_i)^2$$



$$\hat{y}_i = b_0 + b_1 x_i$$

$$\min \sum (y_i - (b_0 + b_1 x_i))^2 \longrightarrow \text{Objective function}$$

**Question: How to calculate  $b_0$  and  $b_1$ ?**

$$\text{derivative}[\sum (y_i - (b_0 + b_1 x_i))^2] = 0 \quad \rightarrow \quad \text{solve for } b_0, b_1$$

# Linear Regression

## □ The Least Squares Method

- Considering the objective function:

$$J = \sum (y_i - (b_0 + b_1 x_i))^2$$

- Rewrite it in matrix form as:

$$J = \|Y - \theta^T X\|_2^2$$

where  $Y = [y_1, \dots, y_n]$ ,  $X = \begin{bmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{bmatrix}$ , and  $\theta = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$

$$\frac{\partial J}{\partial \theta} = -2(Y - \theta^T X)X^T = 0$$

$$\theta^* = (XX^T)^{-1}XY^T$$

# Supervised learning

- Linear Regression
- *Logistic Regression*
- Classification
  - Distance-based algorithms
  - Linear classifiers
  - Other classifiers
- .....



# Logistic Regression

## □ Logistic Regression Model

Define logistic model as

$$\ln \frac{p}{1-p} = b_0 + b_1 X$$

We obtained that,

$$p = \frac{1}{1 + e^{-(b_0 + b_1 X)}}$$

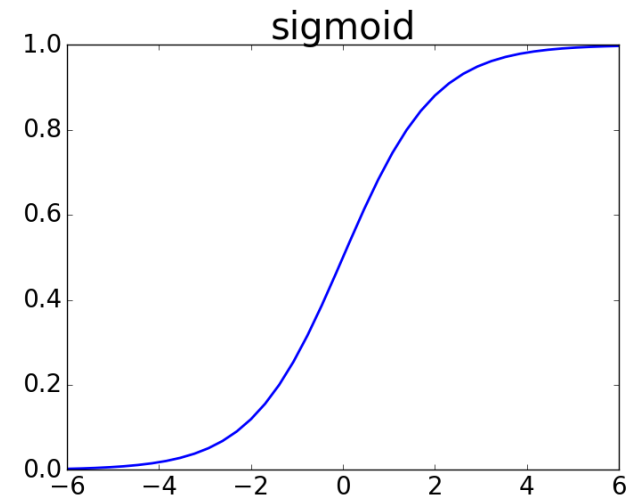
$$h_{\theta}(X) = \frac{1}{1 + e^{-\theta^T X}}$$

Therefore,

$$P(\text{class} = 1|x; \theta) = h_{\theta}(X)$$

$$P(\text{class} = 0|x; \theta) = 1 - h_{\theta}(X)$$

$$g(z) = \frac{1}{1 + e^{-z}}$$



The output of sigmoid function could be used to indicate the probability.

# Logistic Regression

## □ Logistic Regression Model

$$P(\text{class} = 1|x; \theta) = h_{\theta}(X)$$

$$P(\text{class} = 0|x; \theta) = 1 - h_{\theta}(X)$$



$$P(\text{class} = y|x; \theta) = h_{\theta}(X)^y (1 - h_{\theta}(X))^{1-y}$$

Considering all the given data (training set):

$$X = [x_1, \dots, x_n], \quad Y = [y_1, \dots, y_n],$$

$$L(\theta) = \prod_i^n h_{\theta}(x_i)^{y_i} (1 - h_{\theta}(x_i))^{1-y_i}$$

$$\text{The cost function : } J = -\frac{1}{n} \log (L(\theta))$$

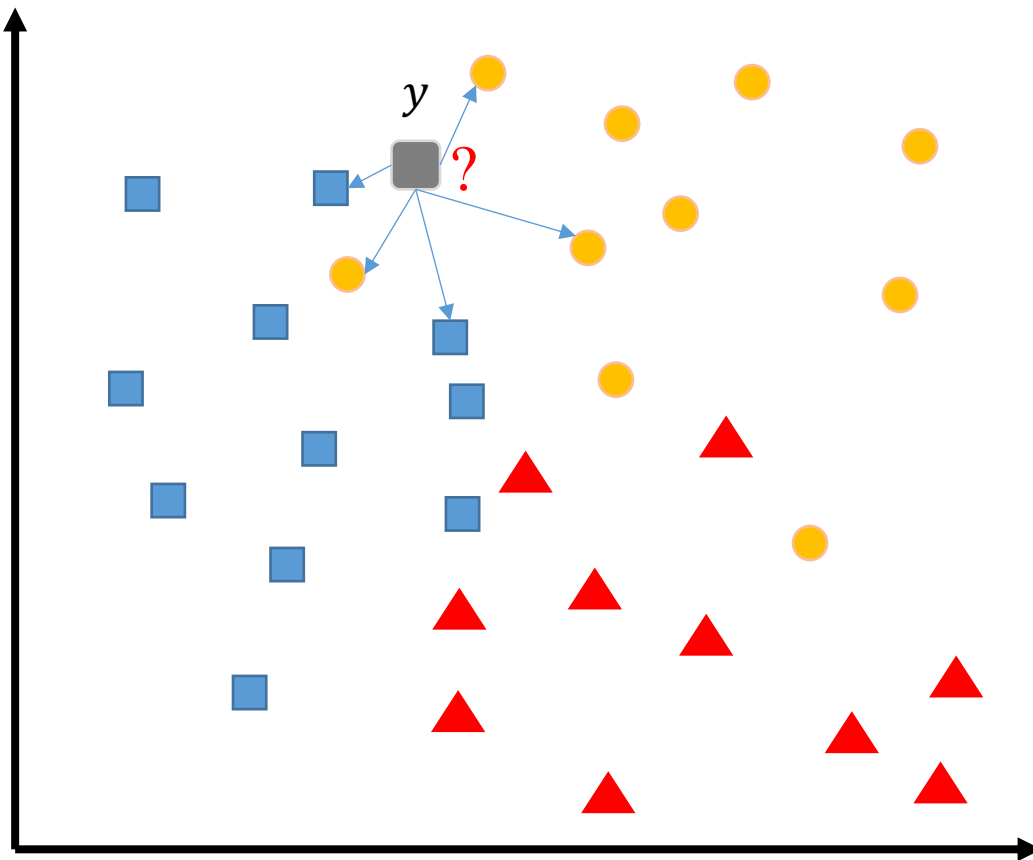
# Supervised learning

- Linear Regression
- Logistic Regression
- *Classification*
  - *Distance-based algorithms*
  - Linear classifiers
  - Other classifiers
- .....



# Classification

## □ Nearest neighbor



How to decide which is the nearest

$$d^j(x^{(y)}, y) = \sqrt{\sum_{i=1}^n (x_i^{(j)} - y)^2}$$

Calculate all the distances from the training data to the test data  $y$ , and we obtain:

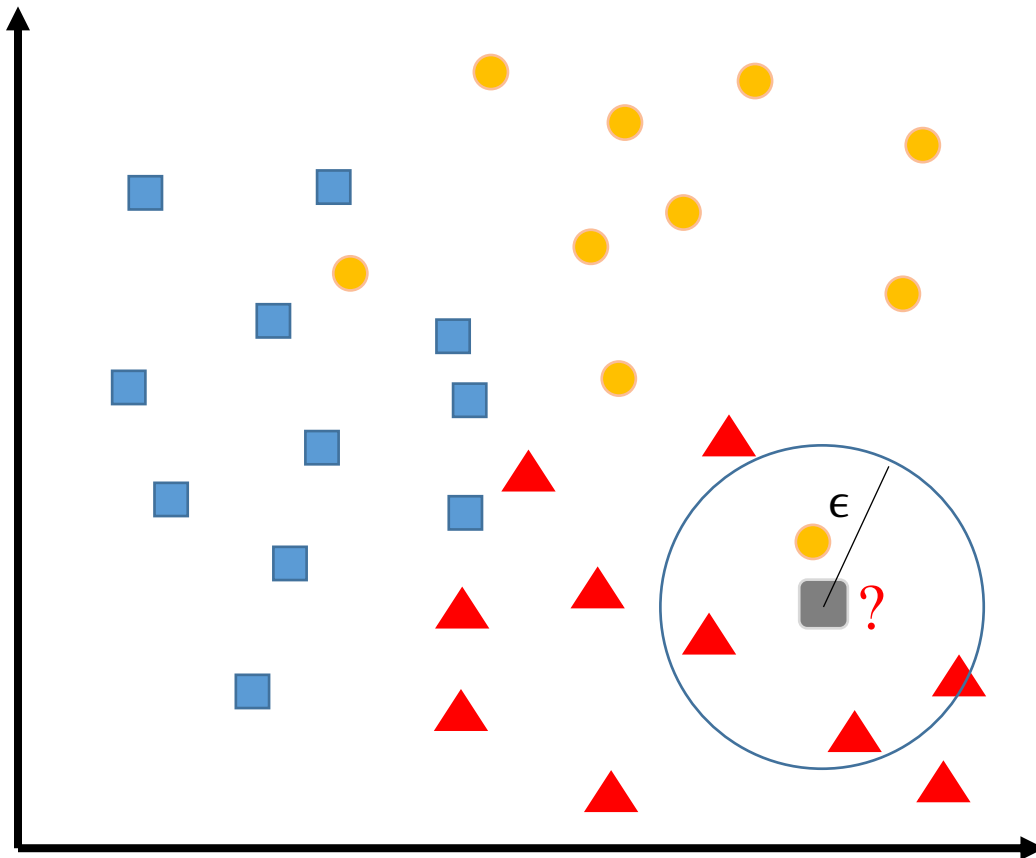
$$D = [d^{(1)}, d^{(2)}, \dots, d^{(N)}]$$

$$s = \operatorname{argmin}_i d^{(i)}$$

$$\operatorname{label}(y) = \operatorname{label}(x^{(s)}) = \text{blue square}$$

# Classification

## □ $\epsilon$ -ball Nearest neighbor



Select a value  $\epsilon$ , then draw a ball in  $\mathbb{R}^n$  with  $y$  as the center and  $\epsilon$  as the radius.

The label of  $y$  is decided by majority labels of points in this ball.

In this ball:

▲ : 3

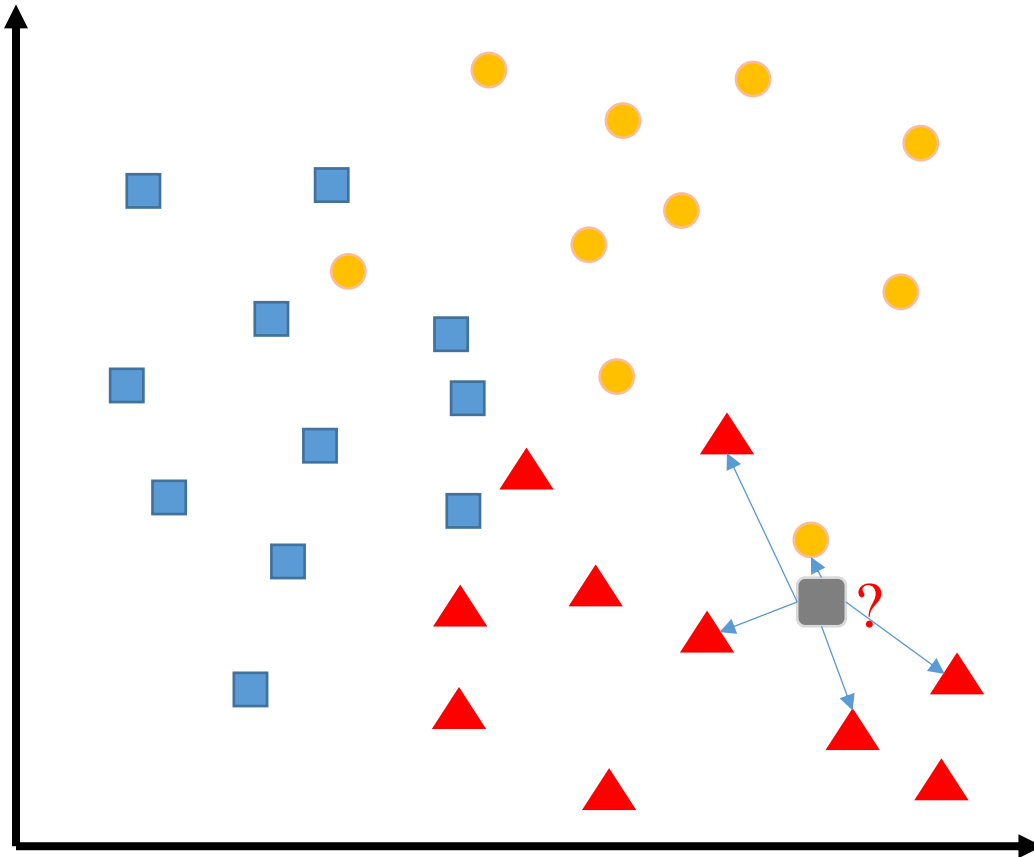
● : 1

■ belongs to ▲



# Classification

## □ K Nearest neighbor



Select a value  $k$ , then find  $y$ 's  $k$  nearest neighbor.

The label of  $y$  is decided by majority labels of  $y$ 's  $k$  neighbors.

Let  $k$  be 5,

▲ : 5      ● : 1

■ belongs to ▲

# Classification



## □ Nearest neighbor classifier

Problem:





















- Need to determine value of parameter  $K$
- Distance based learning is not clear which **type of distance** to use and which attribute to use to produce the best results.
- Computation cost is quite high because we need to compute distance of each query instance to all training samples.

# Classification

## □ Example

- Each image is represented by a vector of dimension 784.

The matrix indicates the pairwise distances.

										
	0	2.8735	2.1766	2.6559	2.2201	2.2500	2.0893	2.4795	2.8443	2.1202
	2.8735	0	2.5055	2.8681	2.9475	2.6062	2.8493	2.8330	2.9434	3.1619
	2.1766	2.5055	0	2.9024	2.3556	0.7858	2.3561	2.2060	2.5274	2.4331
	2.6559	2.8681	2.9024	0	2.7428	2.9531	3.0539	2.8362	2.8488	2.6425
	2.2201	2.9475	2.3556	2.7428	0	2.5284	2.1733	2.4262	2.3432	2.5895
	2.2500	2.6062	0.7858	2.9531	2.5284	0	2.4679	2.2906	2.5549	2.3900
	2.0893	2.8493	2.3561	3.0539	2.1733	2.4679	0	2.5580	2.7456	2.3759
	2.4795	2.8330	2.2060	2.8362	2.4262	2.2906	2.5580	0	2.8885	2.5823
	2.8443	2.9434	2.5274	2.8488	2.3432	2.5549	2.7456	2.8885	0	2.9773
	2.1202	3.1619	2.4331	2.6425	2.5895	2.3900	2.3759	2.5823	2.9773	0

The distance between the data is inconsistent with similarity of the content of the image .

# Supervised learning

- Linear Regression
- Logistic Regression
- Classification
  - Distance-based algorithms
  - *Linear classifiers*
  - Other classifiers
- .....



# Classification

## □ Linearly separable

Apple = [diameter, color, shape, spots, place of production]



$$A_1 = \begin{bmatrix} 7.8 \\ 0.2 \end{bmatrix}$$



$$A_2 = \begin{bmatrix} 7.4 \\ 0.2 \end{bmatrix}$$



$$A_3 = \begin{bmatrix} 7.1 \\ 0.1 \end{bmatrix}$$



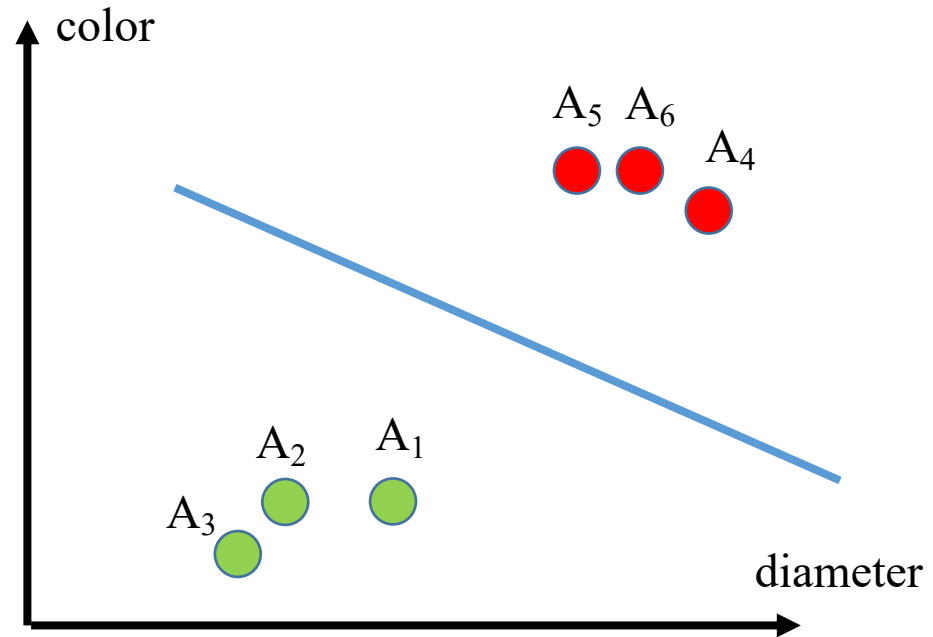
$$A_4 = \begin{bmatrix} 8.5 \\ 0.7 \end{bmatrix}$$



$$A_5 = \begin{bmatrix} 8.1 \\ 0.8 \end{bmatrix}$$



$$A_6 = \begin{bmatrix} 8.3 \\ 0.8 \end{bmatrix}$$



These training data are *linearly separable*

# Classification

## □ Linearly separable

Apple = [diameter, color, shape, spots, place of production]



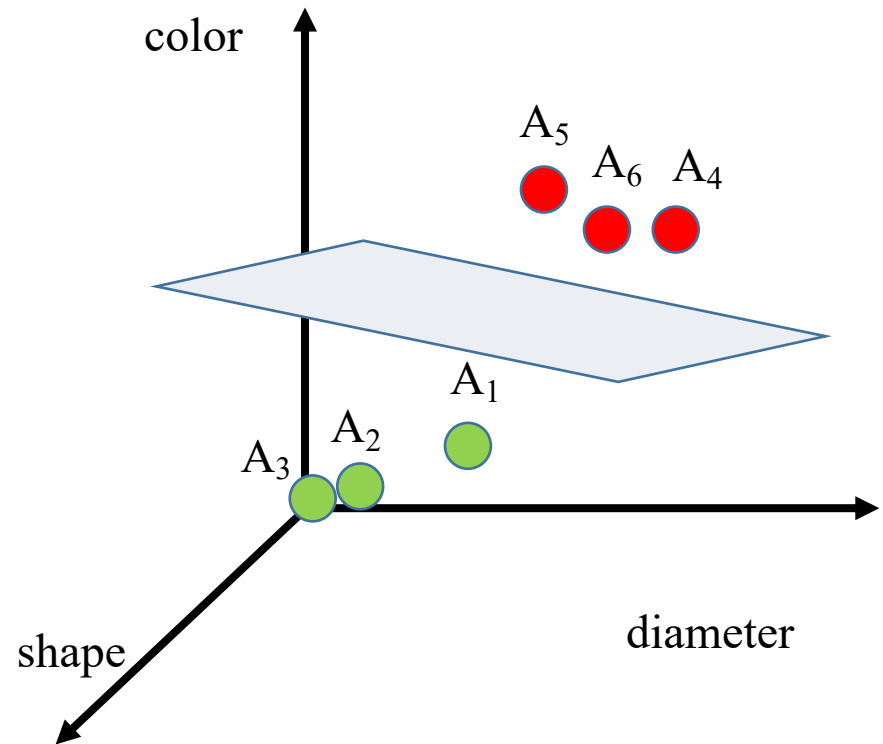
$$A_1 = \begin{bmatrix} 7.8 \\ 0.2 \\ 0.6 \end{bmatrix}$$



$$A_2 = \begin{bmatrix} 7.4 \\ 0.2 \\ 0.7 \end{bmatrix}$$



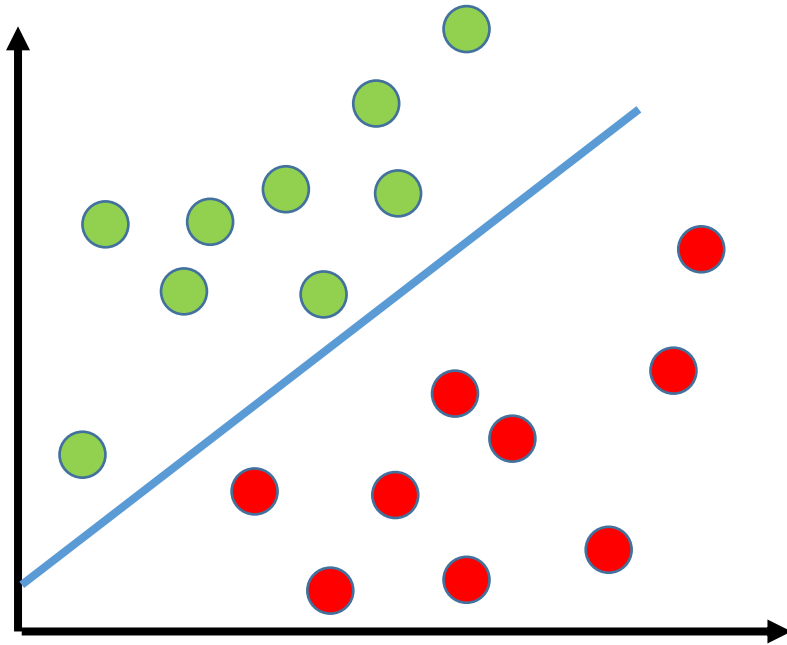
$$A_3 = \begin{bmatrix} 7.1 \\ 0.1 \\ 0.6 \end{bmatrix}$$



In  $n$  dimensions a hyperplane is needed for the separation.

# Classification

## □ Linearly separable

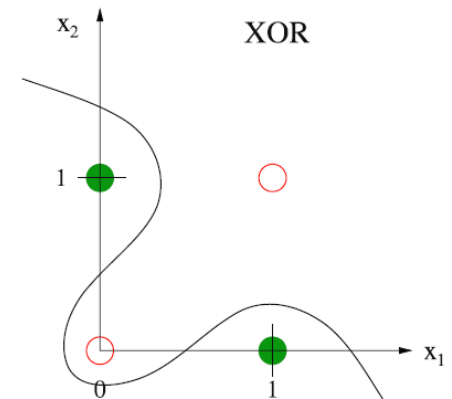
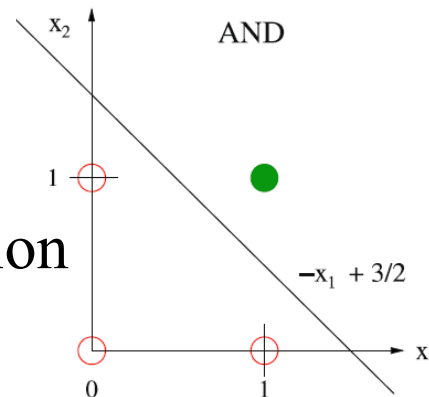


A linearly separable two dimensional data set. The equation for the dividing straight line is

$$w_1x_1 + w_2x_2 = 1$$

Every  $(n - 1)$ -dimensional hyperplane in  $R^n$  can be described by an equation

$$\sum_{i=1}^n w_i x_i + b = 0$$



The boolean function AND is linearly separable, but XOR is not (●  $\hat{=}$  true, ○  $\hat{=}$  false)

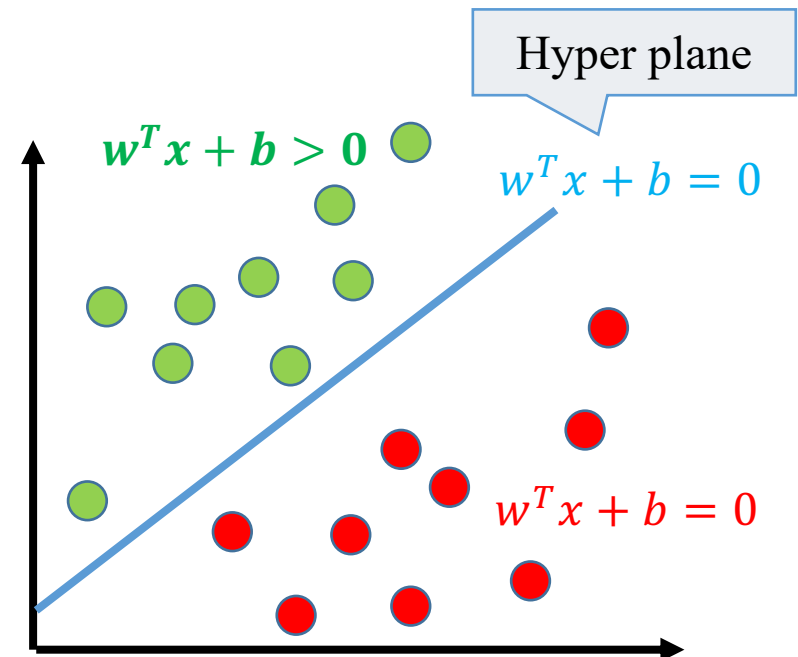
# Classification

## □ Linearly separable

- **Definition** Two sets  $M_1 \subset R^n$  and  $M_2 \subset R^n$  are called *linearly separable*.
- if real vector  $\mathbf{w}=[w_1, w_2, \dots, w_n]$ ,  $b$  exist with

$$\begin{aligned} \mathbf{w}^T \mathbf{x} + b &> 0 \text{ for all } \mathbf{x} \in M_1 \\ \text{and} \\ \mathbf{w}^T \mathbf{x} + b &\leq 0 \text{ for all } \mathbf{x} \in M_2 \end{aligned}$$

$$\text{classify}(\mathbf{x}; \mathbf{w}, b) = \text{sign}(\mathbf{w}^T \mathbf{x} + b)$$





# Classification

## □ Linearly separable

- Given a training set which is linearly separable:

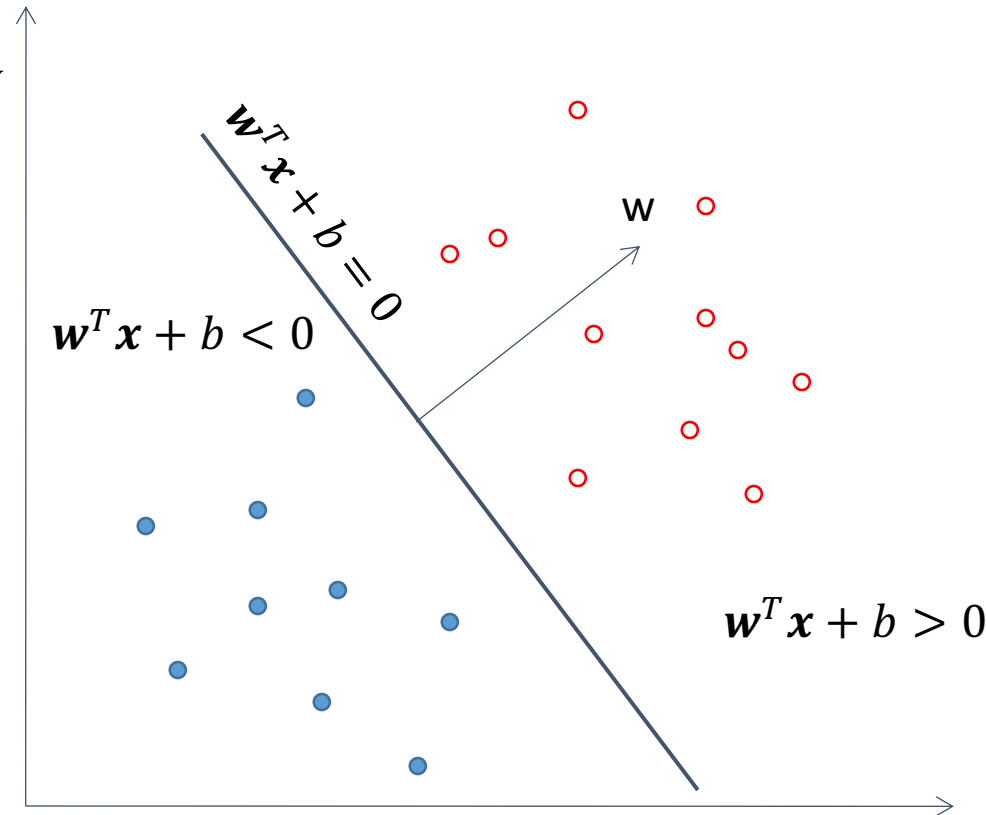
$$D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\},$$

$$\mathbf{x}_i \in X = \mathbb{R}^d,$$

$$y_i \in Y = \{-1, +1\}, \quad i = 1, 2, \dots, m$$

- Goal: solve a separating hyperplane

$$g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b = 0$$

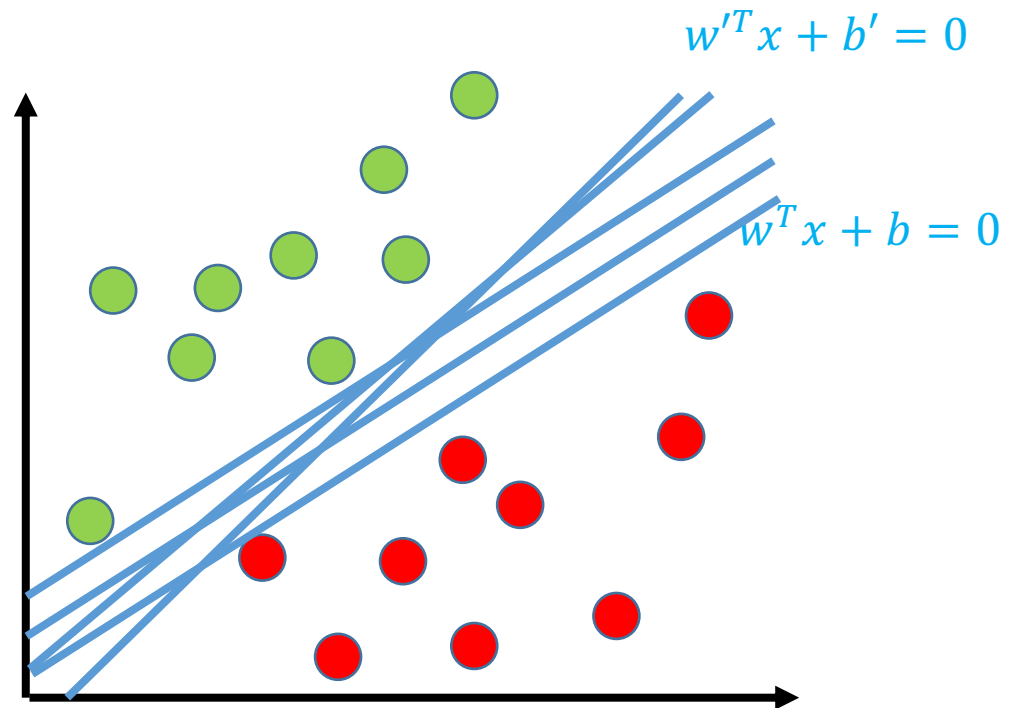


# Classification

## □ Linearly separable

Any of these would be fine.

But which is best?

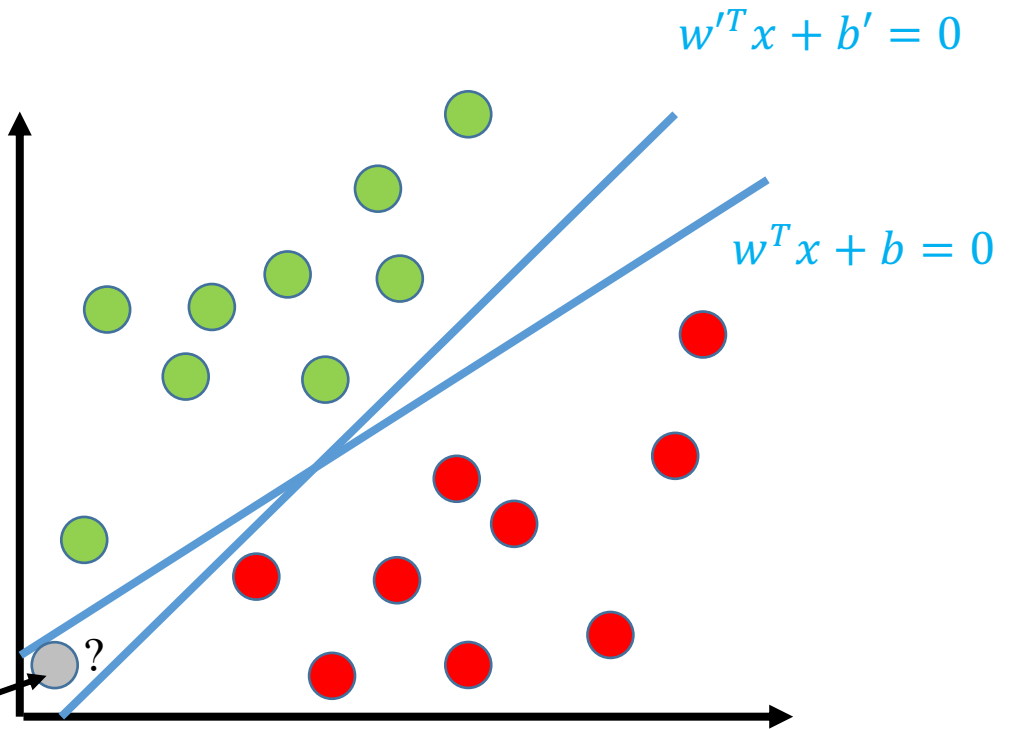


# Classification

## □ Linearly separable

These two lines can separate the two classes of points.

How would you classify this point?

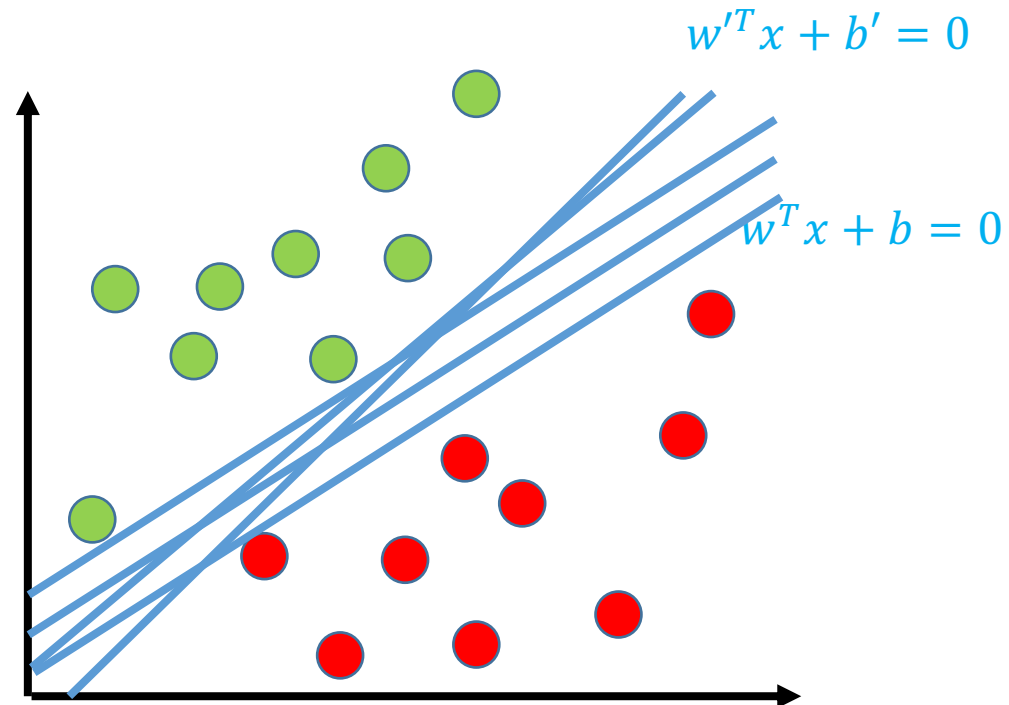


# Classification

## □ Linearly separable

The distance between a sample and hyperplane indicates the **classification confidence**:

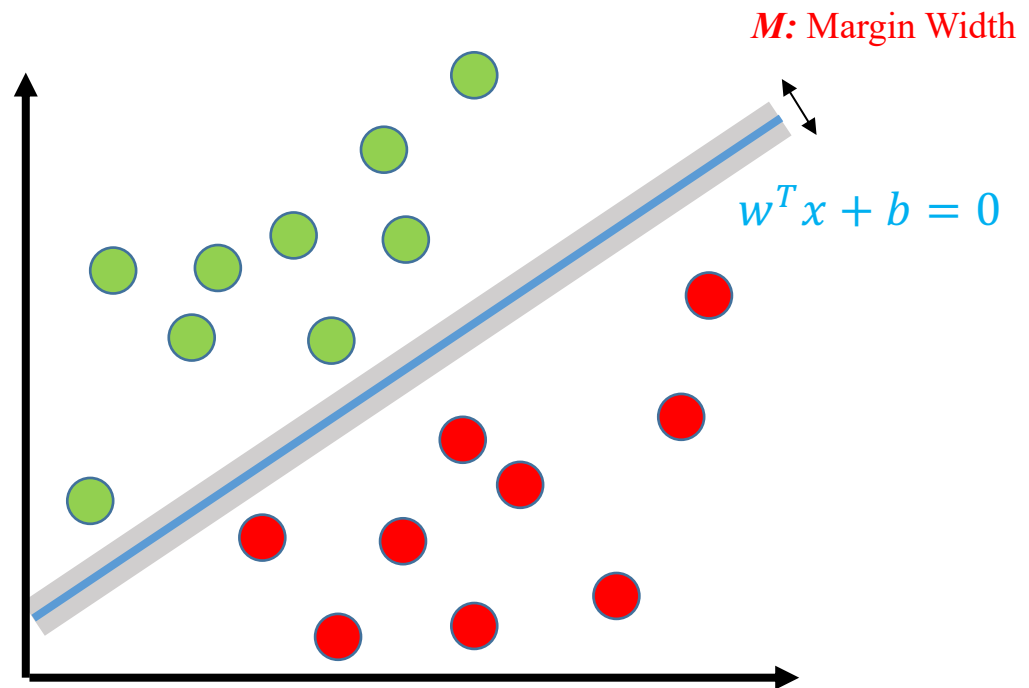
- The farther the sample is from the hyperplane, the higher the confidence that it will be correctly classified.
- The closer the sample is to the hyperplane, the lower the confidence that it will be correctly classified.



# Classification

## □ Linearly separable

Define the **margin** of a linear classifier as the width that the boundary could be increased by before hitting a data point.

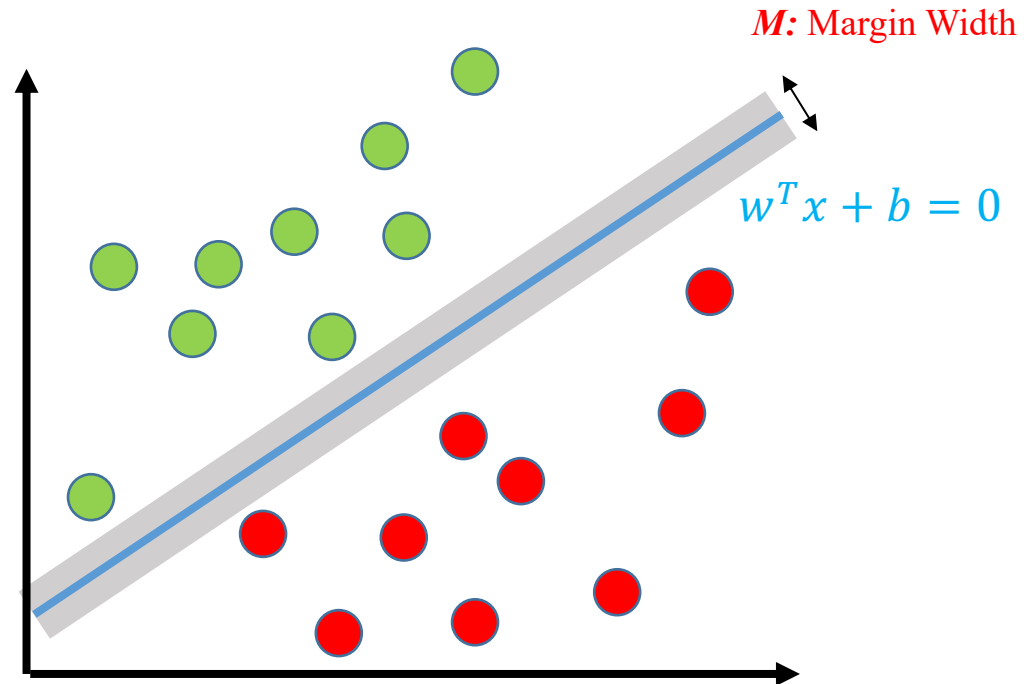


# Classification

## □ Linearly separable

**Hyperplane with maximum margin:** the hyperplane separating samples with maximum margin, which is more robust for classification.

- Two-class samples are separating on corresponding side of the hyperplane;
- The distance from the closest sample point to the hyperplane on both sides to the hyperplane is maximized.

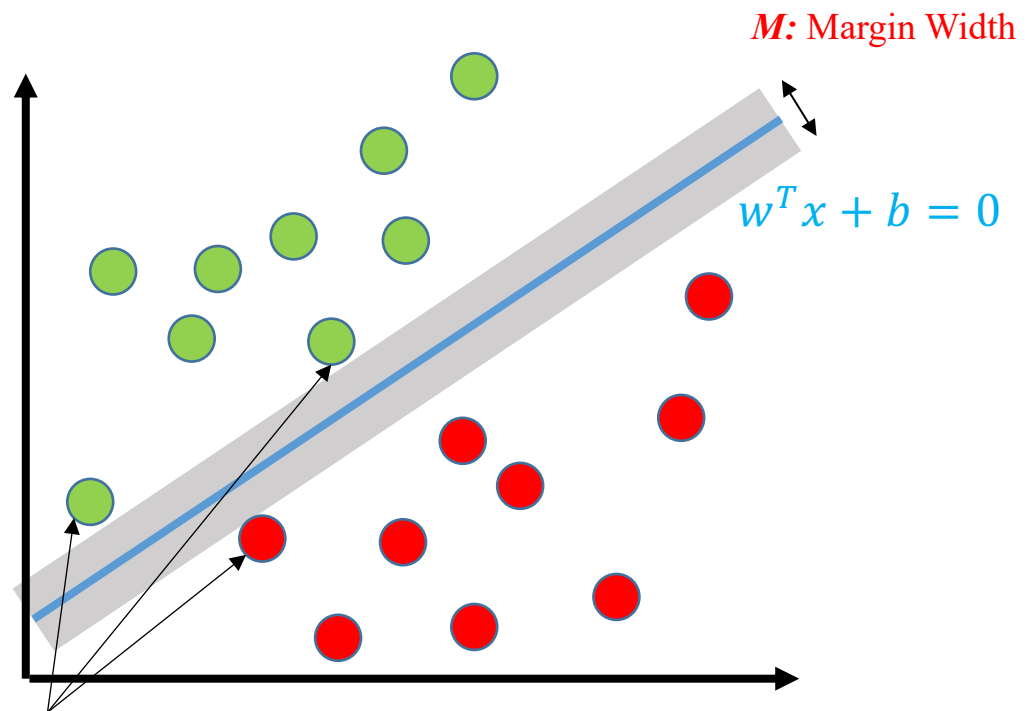


**Support Vector Machine (SVM):** the optimal separating hyperplane when samples are linearly separable.

# Classification

## □ Support vector machine (SVM)

1. Maximizing the margin is good.
2. Implies that only support vectors are important; other training examples are ignorable.
3. Empirically it works very well.



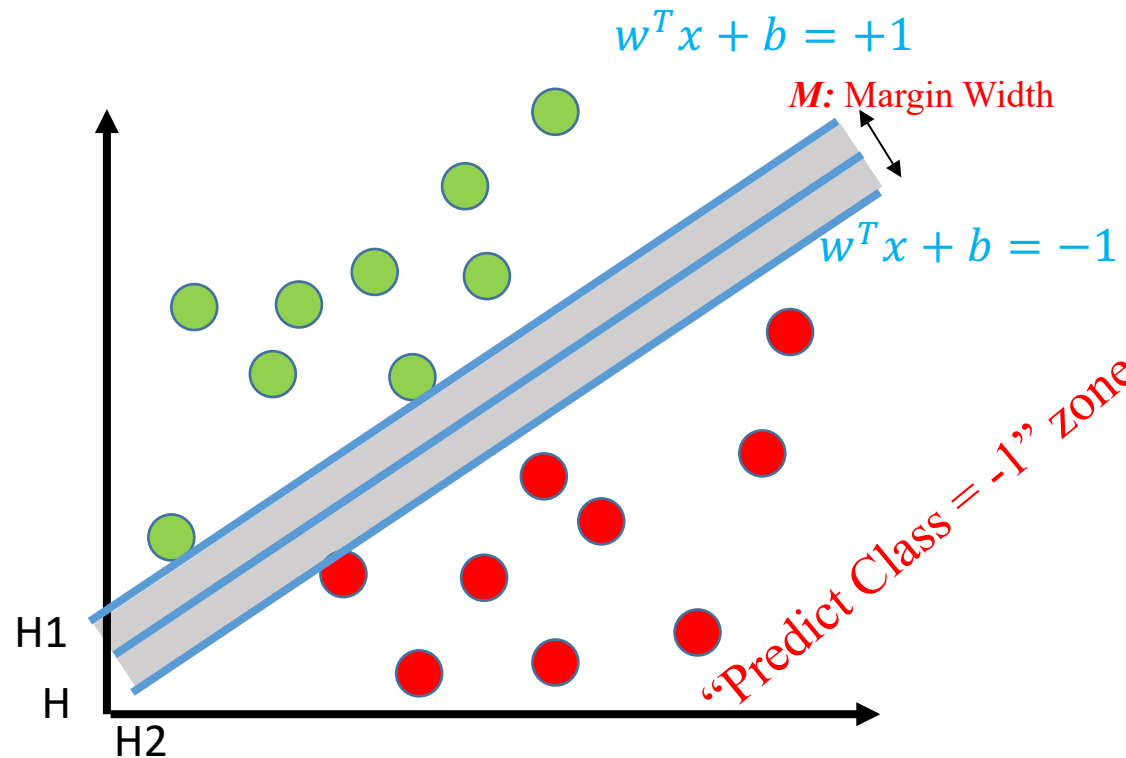
**Support Vectors:** are those data points that the margin pushes up against.

This is the simplest kind of SVM (Called an LSVM)

# Classification

## □ Linearly separable

- **Margin:**  $H_1$  and  $H_2$  are boundary hyperplanes that pass through the samples closest to the  $H$  and parallel to  $H$ . The distance between  $H_1$  and  $H_2$  is called separating margin.
- **Optimal separating hyperplane:** A hyperplane separating samples correctly, and samples (two-class) closest to the hyperplane also has the maximum distance from the hyperplane.





# Classification

## □ Linearly separable

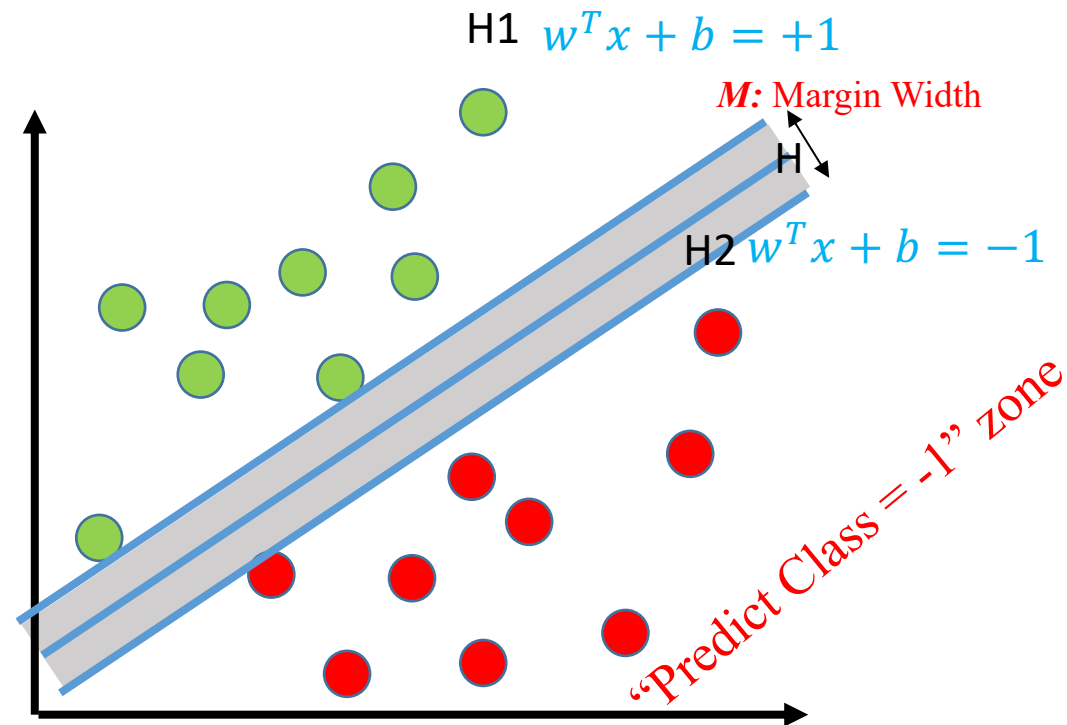
For any line for classification, we can calculate the margin.

Then, which line is the best?

Goal:

- 1) Correctly classify all training data
- 2) Maximize the Margin

How to achieve this goal?



# Classification

## □ SVM

- Given a training set which is linearly separable:  $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$ ,  $\mathbf{x}_i \in X = R^d$ ,  $y_i \in Y = \{-1, +1\}$
- Hyperplane H:  $\mathbf{w}^T \mathbf{x} + b = 0$
- The distance between any sample  $\mathbf{x}$  in feature space to H:  
$$r = \frac{|\mathbf{w}^T \mathbf{x} + b|}{\|\mathbf{w}\|}$$

# Classification

## □ SVM-- Goal 1

- Linearly separable
- $\begin{cases} \mathbf{w}^T \mathbf{x}_i + b > 0, y_i = +1 \\ \mathbf{w}^T \mathbf{x}_i + b < 0, y_i = -1 \end{cases}$
- Linearly separable sample with high confidence and accuracy

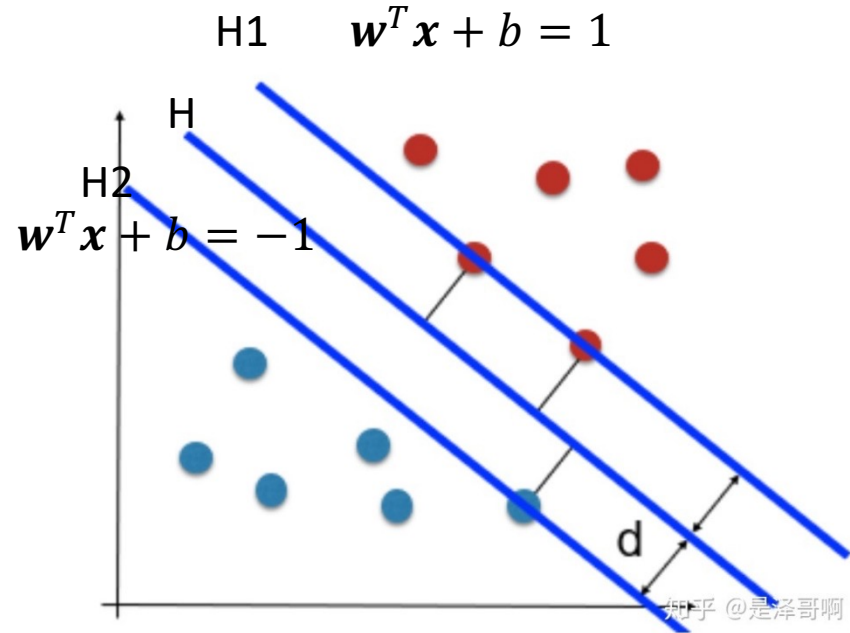
$$\begin{cases} \frac{\mathbf{w}^T \mathbf{x}_i + b}{\|\mathbf{w}\|} \geq d, y_i = +1 \\ \frac{\mathbf{w}^T \mathbf{x}_i + b}{\|\mathbf{w}\|} \leq -d, y_i = -1 \end{cases}$$

$$\Rightarrow \begin{cases} \frac{\mathbf{w}^T \mathbf{x}_i + b}{\|\mathbf{w}\|d} \geq 1, y_i = +1 \\ \frac{\mathbf{w}^T \mathbf{x}_i + b}{\|\mathbf{w}\|d} \leq -1, y_i = -1 \end{cases}$$

$\|\mathbf{w}\|d=1$   
Normalization

$$\begin{cases} \mathbf{w}^T \mathbf{x}_i + w_0 \geq 1, y_i = +1 \\ \mathbf{w}^T \mathbf{x}_i + w_0 \leq -1, y_i = -1 \end{cases}$$

$$y_i(\mathbf{w}^T \mathbf{x}_i + w_0) \geq 1, \\ \text{Then } |\mathbf{w}^T \mathbf{x}_i + w_0| = y_i(\mathbf{w}^T \mathbf{x}_i + w_0)$$



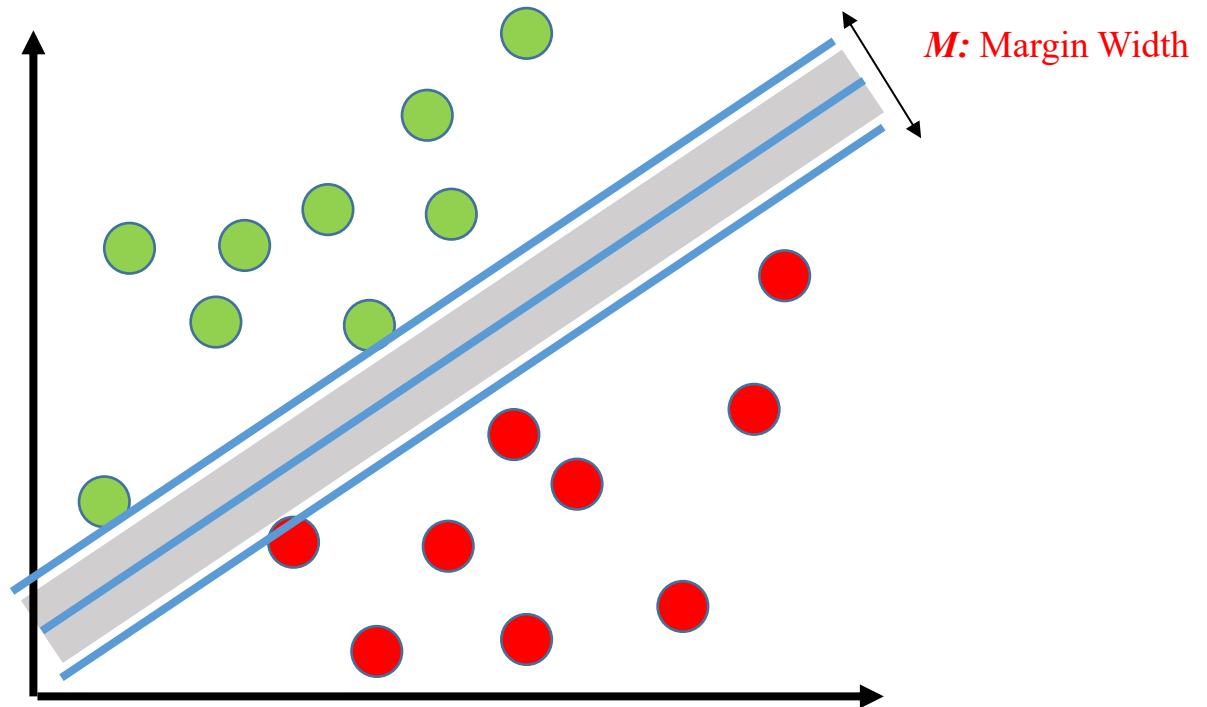
# Classification

## □ SVM

Goal:

1) Correctly classify all training data:

$$y (w^T x + b) \geq 1 \text{ for all } y$$



# Classification

## □ SVM

- Margin:

$$r = 2d = \frac{2}{\|w\|}$$

$$w^T(x_i^+ - x_i^-) = 2$$

$$w^T(x_i^+ - x_i^-) = \|w\| \cdot \|x_i^+ - x_i^-\| \cdot \cos\theta$$

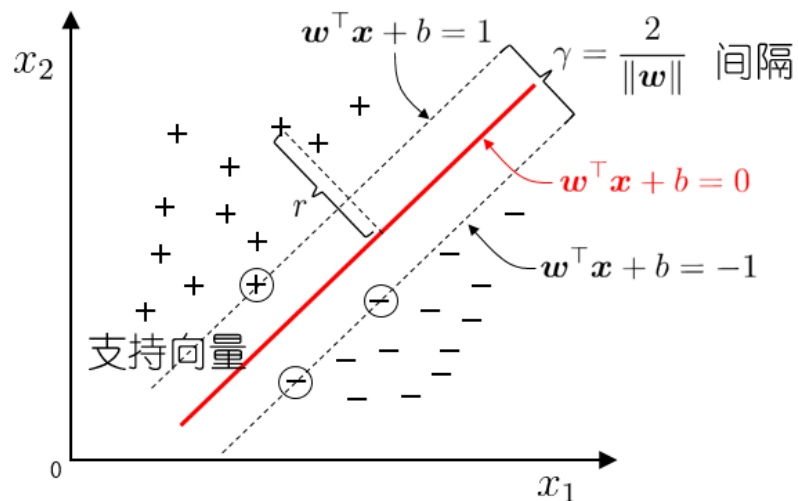
$$= w^T(x_i^+ - x_i^-) = \|w\| \cdot \|x_i^+ - x_i^-\| \cdot \frac{r}{\|x_i^+ - x_i^-\|} = \|w\| \cdot r$$

- **Maximum margin:** solve  $w$  and  $b$  to get maximum margin  $\gamma$

$$\max_{w,b} \frac{2}{\|w\|} \quad s.t. \quad y_i(w^T x_i + b) \geq 1, i = 1, 2, \dots, m$$



$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2} \|w\|^2 \\ s.t. \quad & y_i(w^T x_i + b) \geq 1, i = 1, 2, \dots, m \end{aligned}$$



# Classification

## □ SVM

Goal:

- 1) Correctly classify all training data:

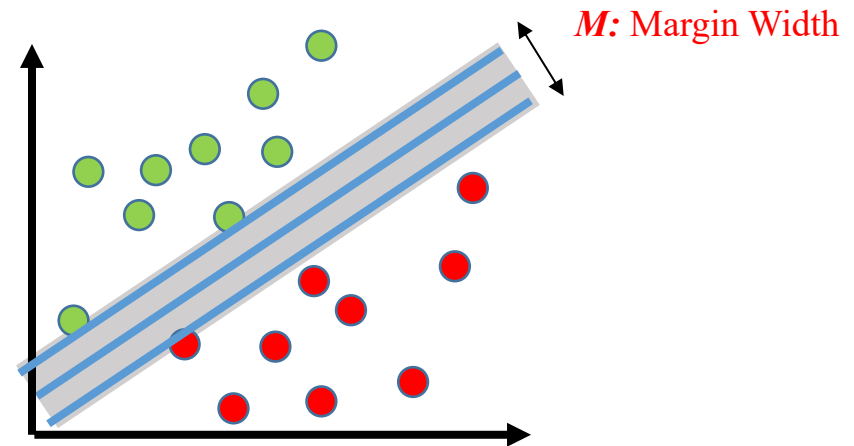
$$y_i (w^T x_i + b) \geq 1 \text{ for all } i$$

- 2) Maximize the margin

$$\min_{w,b} \frac{1}{2} \|w\|^2$$

We can formulate a **Quadratic Optimization Problem** and solve for **w** and **b**

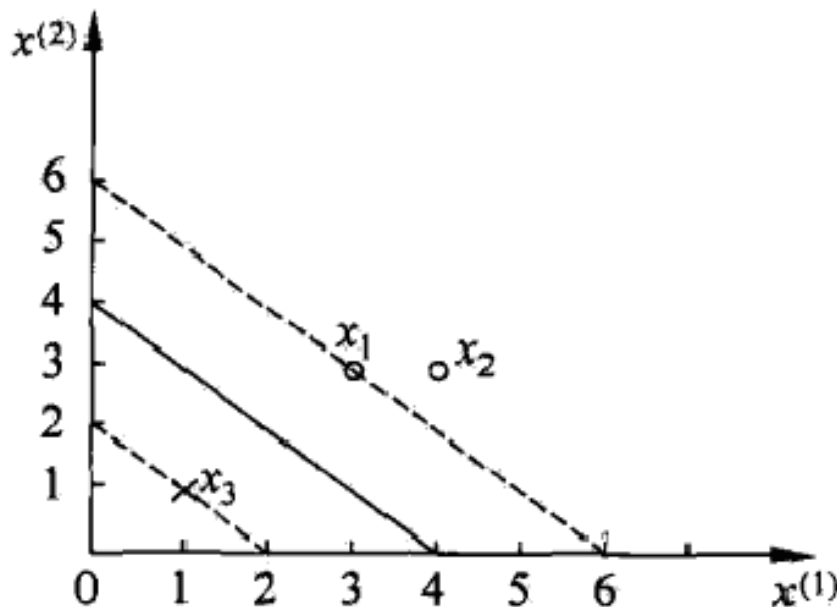
$$\begin{aligned} &\text{Minimize } \Phi(w) = \frac{1}{2} w^T w \\ &\text{subject to } y_i (w x_i + b) \geq 1 \text{ for all } i \end{aligned}$$



The primary problem of SVM: Solve  $d + 1$  variables  $(w, b)$  with  $m$  inequality constraints, which is suitable for low dimensions.

# Example 1——Solve the primary problem of SVM

- Problem: Given a training set size of 3, in which  $(x_1, y_1)$  and  $(x_2, y_2)$  are positive samples, and  $(x_3, y_3)$  is negative sample.  $x_1 = (3; 3)$ ,  $y_1 = +1$ ,  $x_2 = (4; 3)$ ,  $y_2 = +1$ ;  $x_3 = (1; 1)$ ,  $y_3 = -1$ . Solve the optimal separating hyperplane  $H$  with maximum margin.



Solve  $\mathbf{w} = (w_1; w_2)$  and  $w_0$  ( $w_0$  is  $b$ )

# Example 1——Solve the primary problem of SVM

Answer:

- Step 1: Build the primary problem of SVM upon the sample set

$$\begin{aligned} & \min_{\mathbf{w}, b} \frac{1}{2} (w_1^2 + w_2^2) \\ & s. t. \begin{cases} 3w_1 + 3w_2 + w_0 \geq 1 \\ 4w_1 + 3w_2 + w_0 \geq 1 \\ -w_1 - w_2 + w_0 \geq 1 \end{cases} \end{aligned}$$

$$\begin{aligned} & \min_{\mathbf{w}, w_0} \frac{1}{2} \|\mathbf{w}\|^2 \\ & s. t. y_i(\mathbf{w}^T \mathbf{x}_i + w_0) \geq 1, i = 1, 2, \dots, m \end{aligned}$$



# Example 1——Solve the primary problem of SVM

- Step 2: Build Lagrange function by setting Lagrange multiplier  $\alpha_i \geq 0$  for each inequality constraint

$$L(w_1, w_2, w_0, \alpha_1, \alpha_2, \alpha_3) \\ = \frac{1}{2}(w_1^2 + w_2^2) - \alpha_1(3w_1 + 3w_2 + w_0 - 1) - \alpha_2(4w_1 + 3w_2 + w_0 - 1) - \alpha_3(-w_1 - w_2 + w_0 - 1)$$

Set the partial as 0

$$\left\{ \begin{array}{l} \frac{\partial L}{\partial w_1} = w_1 - 3\alpha_1 - 4\alpha_2 + \alpha_3 = 0 \\ \frac{\partial L}{\partial w_2} = w_2 - 3\alpha_1 - 3\alpha_2 + \alpha_3 = 0 \\ \frac{\partial L}{\partial w_0} = -\alpha_1 - \alpha_2 + \alpha_3 = 0 \\ \alpha_1(3w_1 + 3w_2 + w_0 - 1) = 0 \\ \alpha_2(4w_1 + 3w_2 + w_0 - 1) = 0 \\ \alpha_3(-w_1 - w_2 + w_0 - 1) = 0 \\ 3w_1 + 3w_2 + w_0 - 1 \geq 0 \\ 4w_1 + 3w_2 + w_0 - 1 \geq 0 \\ -w_1 - w_2 + w_0 - 1 \geq 0 \\ \alpha_1 \geq 0, \alpha_2 \geq 0, \alpha_3 \geq 0 \end{array} \right. \Rightarrow \left\{ \begin{array}{l} w_1 = 3\alpha_1 + 4\alpha_2 - \alpha_3 \\ w_2 = 3\alpha_1 + 3\alpha_2 - \alpha_3 \\ \alpha_3 = \alpha_1 + \alpha_2 \\ \alpha_1(3w_1 + 3w_2 + w_0 - 1) = 0 \\ \alpha_2(4w_1 + 3w_2 + w_0 - 1) = 0 \\ \alpha_3(-w_1 - w_2 + w_0 - 1) = 0 \\ 3w_1 + 3w_2 + w_0 - 1 \geq 0 \\ 4w_1 + 3w_2 + w_0 - 1 \geq 0 \\ -w_1 - w_2 + w_0 - 1 \geq 0 \\ \alpha_1 \geq 0, \alpha_2 \geq 0, \alpha_3 \geq 0 \end{array} \right.$$

# Example 1——Solve the primary problem of SVM

- Step 3: only keep  $\alpha_1, \alpha_2, w_0$

$$\left\{ \begin{array}{l} w_1 = 3\alpha_1 + 4\alpha_2 - \alpha_3 = 2\alpha_1 + 3\alpha_2 \\ w_2 = 3\alpha_1 + 3\alpha_2 - \alpha_3 = 2\alpha_1 + 2\alpha_2 \\ \alpha_3 = \alpha_1 + \alpha_2 \\ \alpha_1(12\alpha_1 + 15\alpha_2 + w_0 - 1) = 0 \\ \alpha_2(14\alpha_1 + 18\alpha_2 + w_0 - 1) = 0 \\ (\alpha_1 + \alpha_2)(-4\alpha_1 - 5\alpha_2 - w_0 - 1) = 0 \\ 12\alpha_1 + 15\alpha_2 + w_0 - 1 \geq 0 \\ 14\alpha_1 + 18\alpha_2 + w_0 - 1 \geq 0 \\ -4\alpha_1 - 5\alpha_2 - w_0 - 1 \geq 0 \\ \alpha_1 \geq 0, \alpha_2 \geq 0, \alpha_3 \geq 0 \end{array} \right.$$

(1) If  $\alpha_1 = 0, \alpha_2 = 0$ , then

$$w_1 = w_2 = 0$$

However,  $w_0 \geq 1$  and  $w_0 \leq -1$  are contradictory!!!

(2) If  $\alpha_1 > 0, \alpha_2 = 0$ , then

$$\left\{ \begin{array}{l} 12\alpha_1 + w_0 - 1 = 0 \\ -4\alpha_1 - w_0 - 1 = 0 \end{array} \right.$$

We will have

$$\left\{ \begin{array}{l} \alpha_1 = \alpha_3 = \frac{1}{4} \\ \alpha_2 = 0 \\ w_0 = -2 \\ w_1 = w_2 = \frac{1}{2} \end{array} \right.$$

Which satisfy all inequality constraints!

# Example 1——Solve the primary problem of SVM

- Step 3: only keep  $\alpha_1, \alpha_2, w_0$

$$\left\{ \begin{array}{l} w_1 = 3\alpha_1 + 4\alpha_2 - \alpha_3 = 2\alpha_1 + 3\alpha_2 \\ w_2 = 3\alpha_1 + 3\alpha_2 - \alpha_3 = 2\alpha_1 + 2\alpha_2 \\ \alpha_3 = \alpha_1 + \alpha_2 \\ \alpha_1(12\alpha_1 + 15\alpha_2 + w_0 - 1) = 0 \\ \alpha_2(14\alpha_1 + 18\alpha_2 + w_0 - 1) = 0 \\ (\alpha_1 + \alpha_2)(-4\alpha_1 - 5\alpha_2 - w_0 - 1) = 0 \\ 12\alpha_1 + 15\alpha_2 + w_0 - 1 \geq 0 \\ 14\alpha_1 + 18\alpha_2 + w_0 - 1 \geq 0 \\ -4\alpha_1 - 5\alpha_2 - w_0 - 1 \geq 0 \\ \alpha_1 \geq 0, \alpha_2 \geq 0, \alpha_3 \geq 0 \end{array} \right.$$

(3) If  $\alpha_1 = 0, \alpha_2 > 0$ , then

$$\begin{cases} 18\alpha_2 + w_0 - 1 = 0 \\ -5\alpha_2 - w_0 - 1 = 0 \end{cases}$$

We will have

$$\left\{ \begin{array}{l} \alpha_1 = 0 \\ \alpha_2 = \alpha_3 = \frac{2}{13} \\ w_0 = -\frac{23}{13} \\ w_1 = \frac{6}{13} \\ w_2 = \frac{4}{13} \end{array} \right.$$

Substitute in the inequality constraints to get the following expression

$$12\alpha_1 + 15\alpha_2 + w_0 - 1 = -\frac{6}{13} < 0$$

Which violates the constraint!

# Example 1——Solve the primary problem of SVM

- Step 3: only keep  $\alpha_1, \alpha_2, w_0$

$$\left\{ \begin{array}{l} w_1 = 3\alpha_1 + 4\alpha_2 - \alpha_3 = 2\alpha_1 + 3\alpha_2 \\ w_2 = 3\alpha_1 + 3\alpha_2 - \alpha_3 = 2\alpha_1 + 2\alpha_2 \\ \alpha_3 = \alpha_1 + \alpha_2 \\ \alpha_1(12\alpha_1 + 15\alpha_2 + w_0 - 1) = 0 \\ \alpha_2(14\alpha_1 + 18\alpha_2 + w_0 - 1) = 0 \\ (\alpha_1 + \alpha_2)(-4\alpha_1 - 5\alpha_2 - w_0 - 1) = 0 \\ 12\alpha_1 + 15\alpha_2 + w_0 - 1 \geq 0 \\ 14\alpha_1 + 18\alpha_2 + w_0 - 1 \geq 0 \\ -4\alpha_1 - 5\alpha_2 - w_0 - 1 \geq 0 \\ \alpha_1 \geq 0, \alpha_2 \geq 0, \alpha_3 \geq 0 \end{array} \right.$$

(4) If  $\alpha_1 > 0, \alpha_2 > 0$ , then

$$\left\{ \begin{array}{l} 12\alpha_1 + 15\alpha_2 + w_0 - 1 = 0 \\ 14\alpha_1 + 18\alpha_2 + w_0 - 1 = 0 \\ -4\alpha_1 - 5\alpha_2 - w_0 - 1 = 0 \end{array} \right.$$

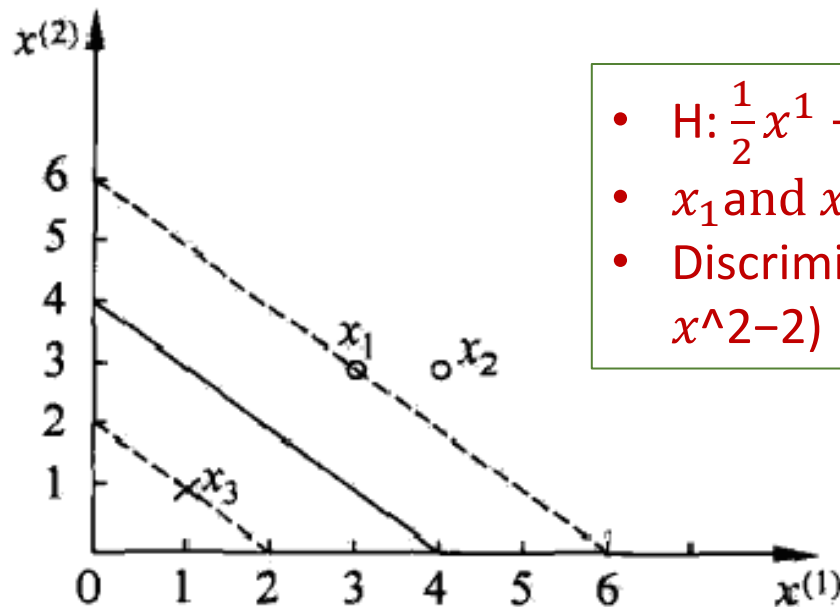
We will have

$$\left\{ \begin{array}{l} \alpha_1 = \frac{3}{2} \\ \alpha_2 = -1 < 0 \\ \alpha_3 = \frac{1}{2} \\ w_0 = -2 \\ w_1 = 0 \\ w_2 = 1 \end{array} \right.$$

Which violates the constraint!

# Example 1——Solve the primary problem of SVM

- Problem: Given a training set size of 3, in which  $(x_1, y_1)$  and  $(x_2, y_2)$  are positive samples, and  $(x_3, y_3)$  is negative sample.  $x_1 = (3; 3)$ ,  $y_1 = +1$ ,  $x_2 = (4; 3)$ ,  $y_2 = +1$ ;  $x_3 = (1; 1)$ ,  $y_3 = -1$ . Solve the optimal separating hyperplane  $H$  with maximum margin.



- $H: \frac{1}{2}x^1 + \frac{1}{2}x^2 - 2 = 0$
- $x_1$  and  $x_3$  are support vectors:  $y_i g(x_i) = 1$
- Discriminant Function:  $g(x) = \text{sign}(\frac{1}{2}x^1 + \frac{1}{2}x^2 - 2)$

- 
- $m$  inequality constraints means  $2^m$  cases!!

Primary Problem  $\rightarrow$  Duel Problem

# SVM-Duel Problem

- The primary problem of SVM

$$\min_{\mathbf{w}, w_0} \frac{1}{2} \|\mathbf{w}\|^2$$

$$s. t. y_i(\mathbf{w}^T \mathbf{x}_i + w_0) \geq 1, i = 1, 2, \dots, m$$

- Lagrange function: Lagrange multiplier  $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_m)^T$ :

$$L(\mathbf{w}, w_0, \boldsymbol{\alpha}) = \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^m \alpha_i (1 - y_i(\mathbf{w}^T \mathbf{x}_i + w_0))$$

- **Primary problem → Duel problem** (maxi-mini problem)

$$\max_{\boldsymbol{\alpha}} \min_{\mathbf{w}, w_0} L(\mathbf{w}, w_0, \boldsymbol{\alpha})$$

- To get the solution of SVM duel problem:

➤ Solve the minimum of  $L(\mathbf{w}, b, \boldsymbol{\alpha})$  on  $\mathbf{w}, w_0$ :  $\min_{\mathbf{w}, w_0} L(\mathbf{w}, w_0, \boldsymbol{\alpha})$

➤ Solve the maximum of  $\min_{\mathbf{w}, w_0} L(\mathbf{w}, w_0, \boldsymbol{\alpha})$  on  $\boldsymbol{\alpha}$ :  $\max_{\boldsymbol{\alpha}} \min_{\mathbf{w}, w_0} L(\mathbf{w}, w_0, \boldsymbol{\alpha})$

# SVM-Duel Problem

$$\begin{aligned} L(\mathbf{w}, w_0, \boldsymbol{\alpha}) &= \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^m \alpha_i (1 - y_i (\mathbf{w}^T \mathbf{x}_i + w_0)) \\ &= \frac{1}{2} \mathbf{w}^T \mathbf{w} + \sum_{i=1}^m \alpha_i - \sum_{i=1}^m \alpha_i y_i \mathbf{w}^T \mathbf{x}_i - \sum_{i=1}^m \alpha_i y_i w_0 \end{aligned}$$

- (1) Solve  $\min_{\mathbf{w}, w_0} L(\mathbf{w}, w_0, \boldsymbol{\alpha})$ , set the partial of  $L(\mathbf{w}, w_0, \boldsymbol{\alpha})$  on  $\mathbf{w}$  and  $w_0$  as 0

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i = 0, \text{ then } \mathbf{w} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial L}{\partial w_0} = -\sum_{i=1}^m \alpha_i y_i = 0, \text{ then } \sum_{i=1}^m \alpha_i y_i = 0$$

Then,

$$\min_{\mathbf{w}, w_0} L(\mathbf{w}, w_0, \boldsymbol{\alpha}) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$



# SVM-Duel Problem

- (2) Solve the maximum of  $\min_{\mathbf{w}, w_0} L(\mathbf{w}, w_0, \boldsymbol{\alpha})$  on  $\boldsymbol{\alpha}$

$$\max_{\boldsymbol{\alpha}} \min_{\mathbf{w}, w_0} L(\mathbf{w}, w_0, \boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}} \left( \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \right)$$
$$s. t. \sum_{i=1}^m \alpha_i y_i = 0, \alpha_i \geq 0, i = 1, 2, \dots, m$$

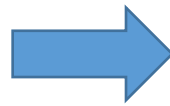
Substitute  $\alpha_i$  after solving to get  $\mathbf{w}$  and  $w_0$ ,  $w_0 = y_j - \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i^T \mathbf{x}_j$

$$g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0 = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + w_0$$

# SVM-Dual Problem——Solution Sparsity

- Final SVM:  $g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0 = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + w_0$
- SVM satisfies KKT(Karush-Kuhn-Tucker) conditions:

$$\begin{cases} \alpha_i \geq 0 \\ y_i g(\mathbf{x}_i) \geq 1 \\ \alpha_i (y_i g(\mathbf{x}_i) - 1) = 0 \end{cases}$$



For any sample  $(\mathbf{x}_i, y_i)$ , there must exist  $\alpha_i = 0$  or  $y_i g(\mathbf{x}_i) = 1$

- If  $\alpha_i = 0$ , then  $y_i g(\mathbf{x}_i) > 1$ ,  $(\mathbf{x}_i, y_i)$  does not affect SVM  $g(\mathbf{x})$ .
- If  $\alpha_i > 0$ , then  $y_i g(\mathbf{x}_i) = 1$ ,  $(\mathbf{x}_i, y_i)$  is on the boundary hyperplane H1 or H2, which is the support vector.

**Solution Sparsity of SVM:** After training, most of the training samples are not reserved. That is, the final SVM only concerns support vectors which is small amount.

For dual problem ,we only need to solve support vectors and corresponding multiplier  $\alpha$ .

## Example 2——Solve the dual problem of SVM

- Problem: Given a training set size of 3, in which  $(x_1, y_1)$  and  $(x_2, y_2)$  are positive samples, and  $(x_3, y_3)$  is negative sample.  $x_1 = (3; 3)$ ,  $y_1 = +1$ ,  $x_2 = (4; 3)$ ,  $y_2 = +1$ ;  $x_3 = (1; 1)$ ,  $y_3 = -1$ . Solve the linearly separable SVM.

# Example 2——Solve the dual problem of SVM

- SVM Primary Problem

$$\min_{\mathbf{w}, b} \frac{1}{2}(w_1^2 + w_2^2)$$
$$s. t. \begin{cases} 3w_1 + 3w_2 + w_0 \geq 1 \\ 4w_1 + 3w_2 + w_0 \geq 1 \\ -w_1 - w_2 + w_0 \geq 1 \end{cases}$$

$$\max_{\alpha} \min_{\mathbf{w}, w_0} L(\mathbf{w}, w_0, \alpha)$$
$$= \max_{\alpha} \left( \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \right)$$
$$s. t. \sum_{i=1}^m \alpha_i y_i = 0, \alpha_i \geq 0, i = 1, 2, \dots, m$$

- Step1: transform to SVM dual problem

$$\max_{\alpha_1, \alpha_2, \alpha_3} \left( \sum_{i=1}^3 \alpha_i - \frac{1}{2} \sum_{i=1}^3 \sum_{j=1}^3 \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \right)$$
$$s. t. \begin{cases} \alpha_1 + \alpha_2 - \alpha_3 = 0 \\ \alpha_1 \geq 0, \alpha_2 \geq 0, \alpha_3 \geq 0 \end{cases}$$



$$\min_{\alpha_1, \alpha_2, \alpha_3} \left( \frac{1}{2} (18\alpha_1^2 + 25\alpha_2^2 + 2\alpha_3^2 + 42\alpha_1\alpha_2 - 12\alpha_1\alpha_3 - 14\alpha_2\alpha_3) - \alpha_1 - \alpha_2 - \alpha_3 \right)$$
$$s. t. \begin{cases} \alpha_1 + \alpha_2 - \alpha_3 = 0 \\ \alpha_1 \geq 0, \alpha_2 \geq 0, \alpha_3 \geq 0 \end{cases}$$

# Example 2——Solve the dual problem of SVM

- Step 2: Substitute constraints  $\alpha_3 = \alpha_1 + \alpha_2$ , then the objective function is:

$$s(\alpha_1, \alpha_2) = 4\alpha_1^2 + \frac{13}{2}\alpha_2^2 + 10\alpha_1\alpha_2 - 2\alpha_1 - 2\alpha_2$$
$$s.t. \alpha_1 \geq 0, \alpha_2 \geq 0, \alpha_3 \geq 0$$

Solve the partial of  $s(\alpha_1, \alpha_2)$  on  $\alpha_1, \alpha_2$ , and set as 0:

$$\begin{cases} \frac{\partial s}{\partial \alpha_1} = 8\alpha_1 + 10\alpha_2 - 2 = 0 \\ \frac{\partial s}{\partial \alpha_2} = 13\alpha_2 + 10\alpha_1 - 2 = 0 \end{cases}$$

$$\text{then} \begin{cases} \alpha_1 = \frac{3}{2} \\ \alpha_2 = -1 < 0, \\ \alpha_3 = \frac{1}{2} \end{cases}$$

This violates constraints! We will find the minimum on boundary value of  $\alpha_i$ .

## Example 2——Solve the dual problem of SVM

- Step 3: Solve vector  $\mathbf{w}$  with KKT conditions

(1) When  $\alpha_1 = 0$ ,

$$s(0, \alpha_2) = \frac{13}{2} \alpha_2^2 - 2\alpha_2,$$

$$\text{Set } \frac{\partial s}{\partial \alpha_2} = 13\alpha_2 - 2 = 0, \text{ then } \alpha_2 = \frac{2}{13}, \quad s_{\min} = -\frac{2}{13}$$

(2) When  $\alpha_2 = 0$ ,

$$s(\alpha_1, 0) = 4\alpha_1^2 - 2\alpha_1,$$

$$\text{Set } \frac{\partial s}{\partial \alpha_1} = 8\alpha_1 - 2 = 0, \text{ then } \alpha_1 = \frac{1}{4}, \quad s_{\min} = -\frac{1}{4}$$

$$\text{Thus, when } \alpha_1 = \frac{1}{4}, \alpha_2 = 0, \alpha_3 = \frac{1}{4}, \quad s_{\min} = -\frac{1}{4}$$

$$\text{And } \mathbf{w} = \sum_{i=1}^3 \alpha_i y_i \mathbf{x}_i = \alpha_1 y_1 \mathbf{x}_1 + \alpha_3 y_3 \mathbf{x}_3 = \left(\frac{1}{2}, \frac{1}{2}\right)$$

# Example 2——Solve the dual problem of SVM

- Step 4: Solve  $w_0$  with KKT conditions

Since  $\alpha_1 = \frac{1}{4} > 0$ , the corresponding sample  $\mathbf{x}_1$  is the support vector, then  $y_1 g(\mathbf{x}_1) = 1$  and  $w_0 = -2$ .

- Hyperplane H (SVM)

$$\frac{1}{2}x^1 + \frac{1}{2}x^2 - 2 = 0$$

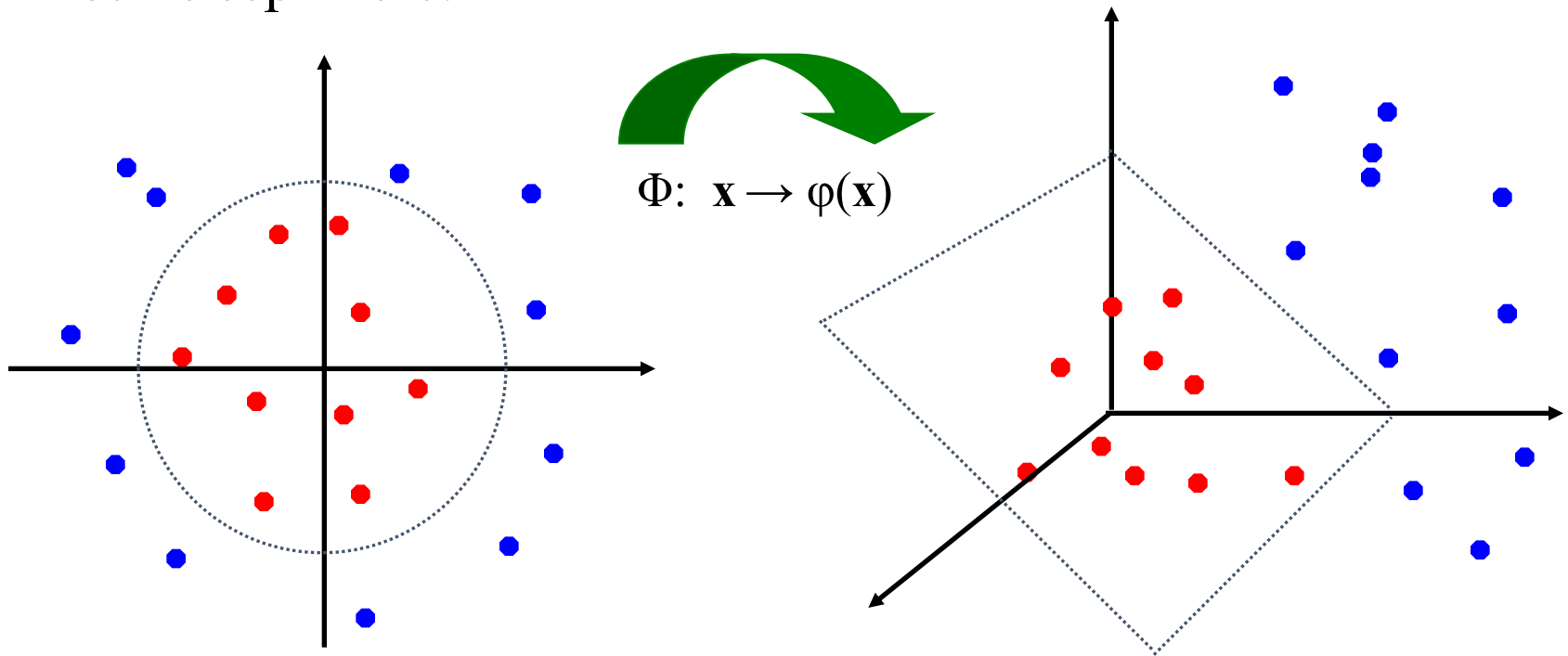
- For new sample, discriminant function is

$$g(\mathbf{x}) = \text{sign}\left(\frac{1}{2}x^1 + \frac{1}{2}x^2 - 2\right)$$

# Classification

## □ NON-linear SVMs

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



A **kernel function** is some function that corresponds to an inner product in some expanded feature space.



# Classification

## □ Weakness of SVM

- It is sensitive to noise
- It only considers two classes
  - how to do multi-class classification with SVM?

Suppose there are  $m$  different classes,

### 1) OVA-SVM : ( $m$ SVMs in total)

- $SVM_1$  learns “Output==1” vs “Output != 1”
- $SVM_2$  learns “Output==2” vs “Output != 2”
- :
- $SVM_m$  learns “Output== $m$ ” vs “Output !=  $m$ ”

### 2) OVO-SVM: ( $m(m-1)/2$ SVMs in total)

- $SVM_{12}$  learns “Output==1” vs “Output == 2”
- $SVM_{13}$  learns “Output==1” vs “Output == 3”
- :
- $SVM_{m(m-1)}$  learns “Output== $m$ ” vs “Output ==  $m-1$ ”

# Classification

---

## □ Other classifiers

- Decision trees
- Sparse representation classifier
- Neural networks
- .....

# Conclusion



- Linear Regression
- Logistic Regression
- Classification
  - Distance-based algorithms
  - Linear classifiers
  - Other classifiers
- .....

