COMP7/8118 M50

Data Mining

Support Vector Machine

Xiaofei Zhang

Slides compiled from Jiawei Han and Raymond C.W. Wong's work

THE UNIVERSITY OF MEMPHIS

SVM: History & Applications

- Vapnik and colleagues (1992)—groundwork from Vapnik & Chervonenkis' statistical learning theory in 1960s
- Features: training can be slow but accuracy is high owing to their ability to model complex nonlinear decision boundaries (margin maximization)
- Used for: classification and numeric prediction
- Applications:
  - handwritten digit recognition, object recognition, speaker identification, benchmarking time-series prediction tests

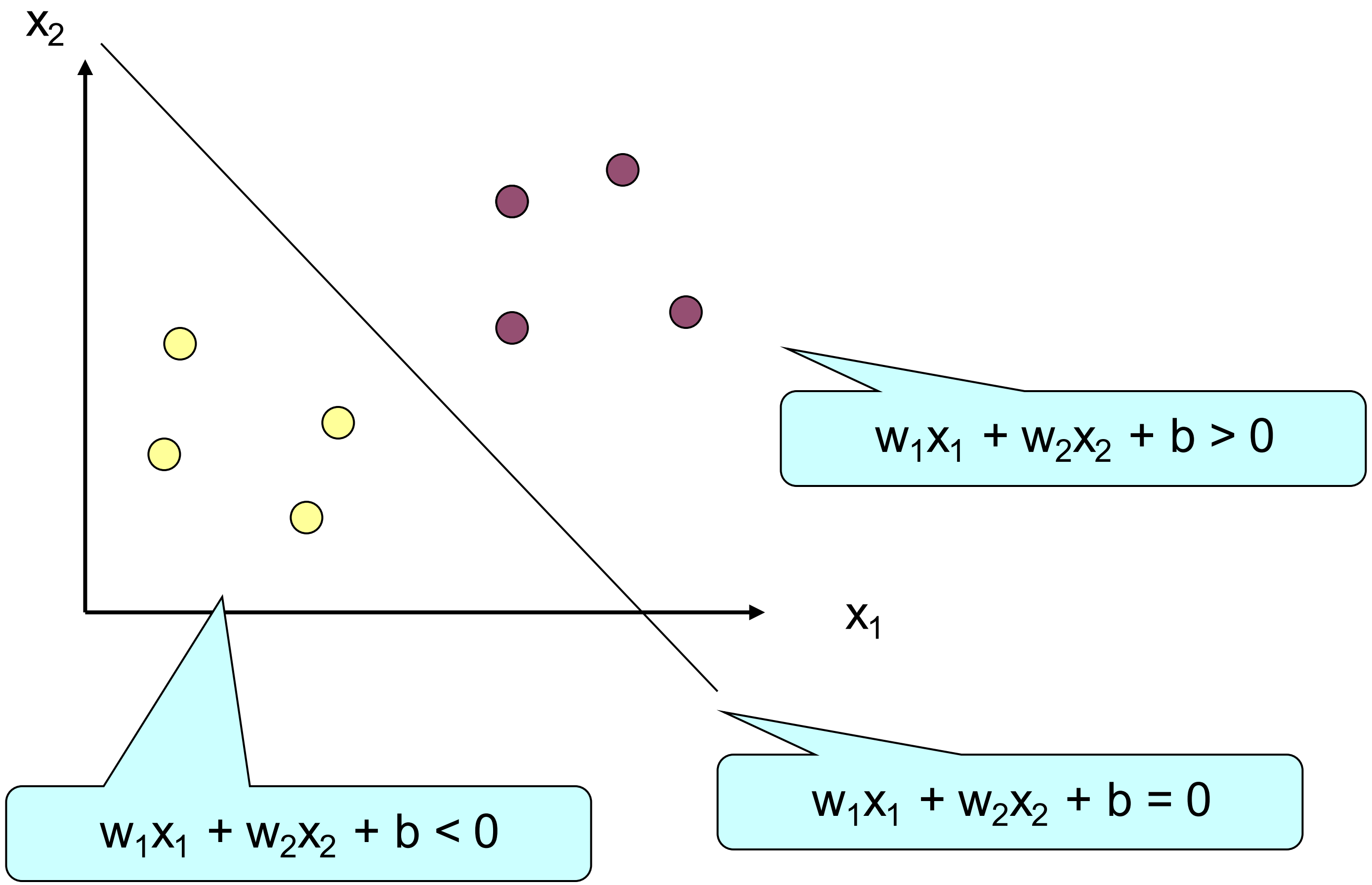
SVM

- Support Vector Machine (SVM)
    - Linear Support Vector Machine
    - Non-linear Support Vector Machine

SVM

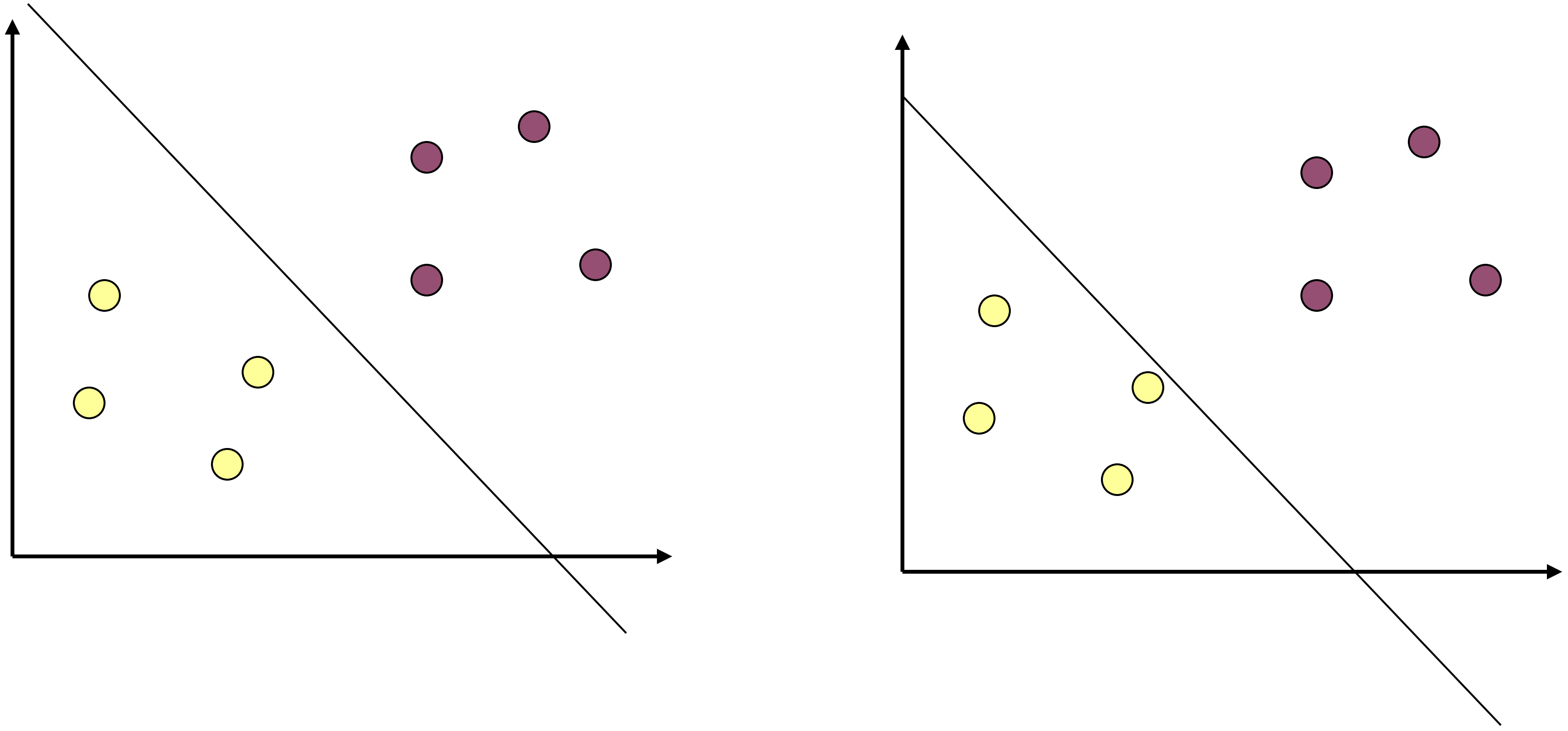
- Advantages:
  - Can be visualized
  - Accurate when the data is well partitioned
  - Fast evaluation of the learned target function
    - Bayesian networks are normally slow

Linear Support Vector Machine

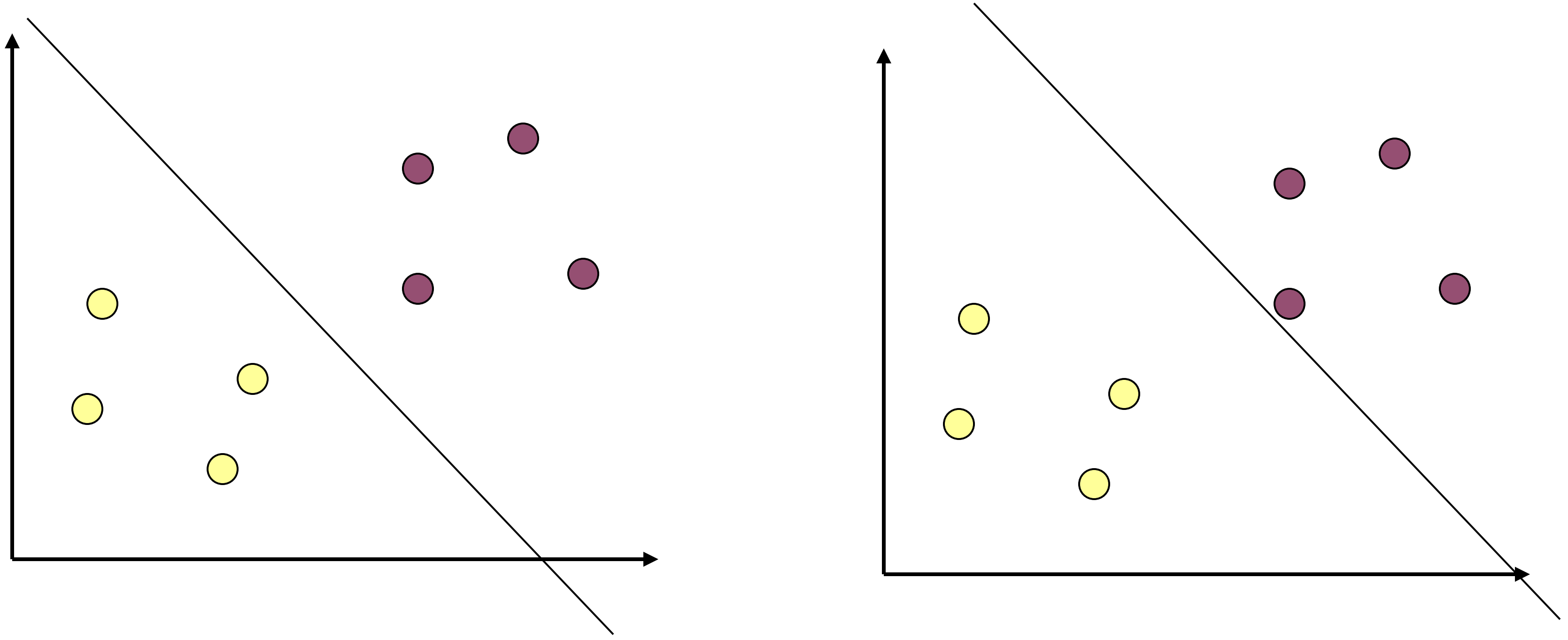


x_2
x_1
$w_1x_1 + w_2x_2 + b > 0$
$w_1x_1 + w_2x_2 + b = 0$
$w_1x_1 + w_2x_2 + b < 0$

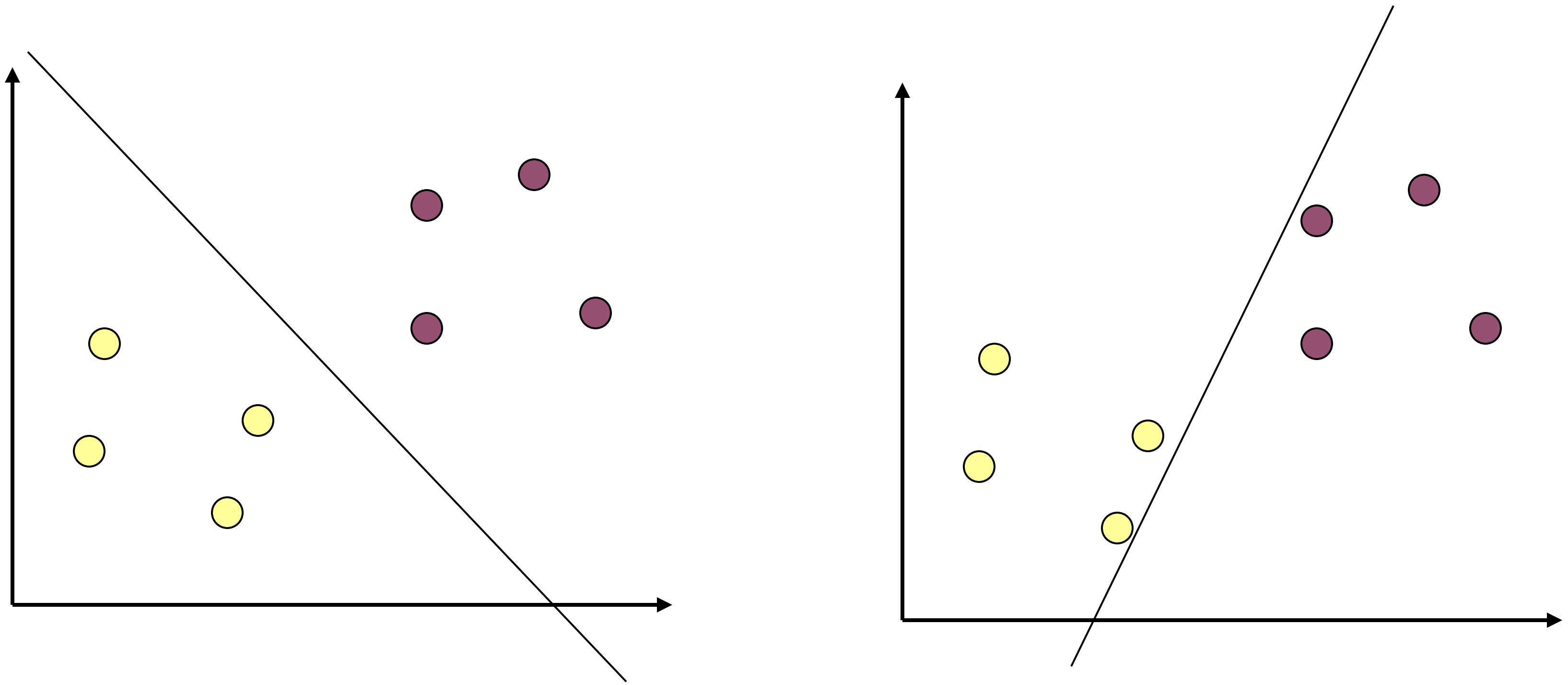
Linear Support Vector Machine



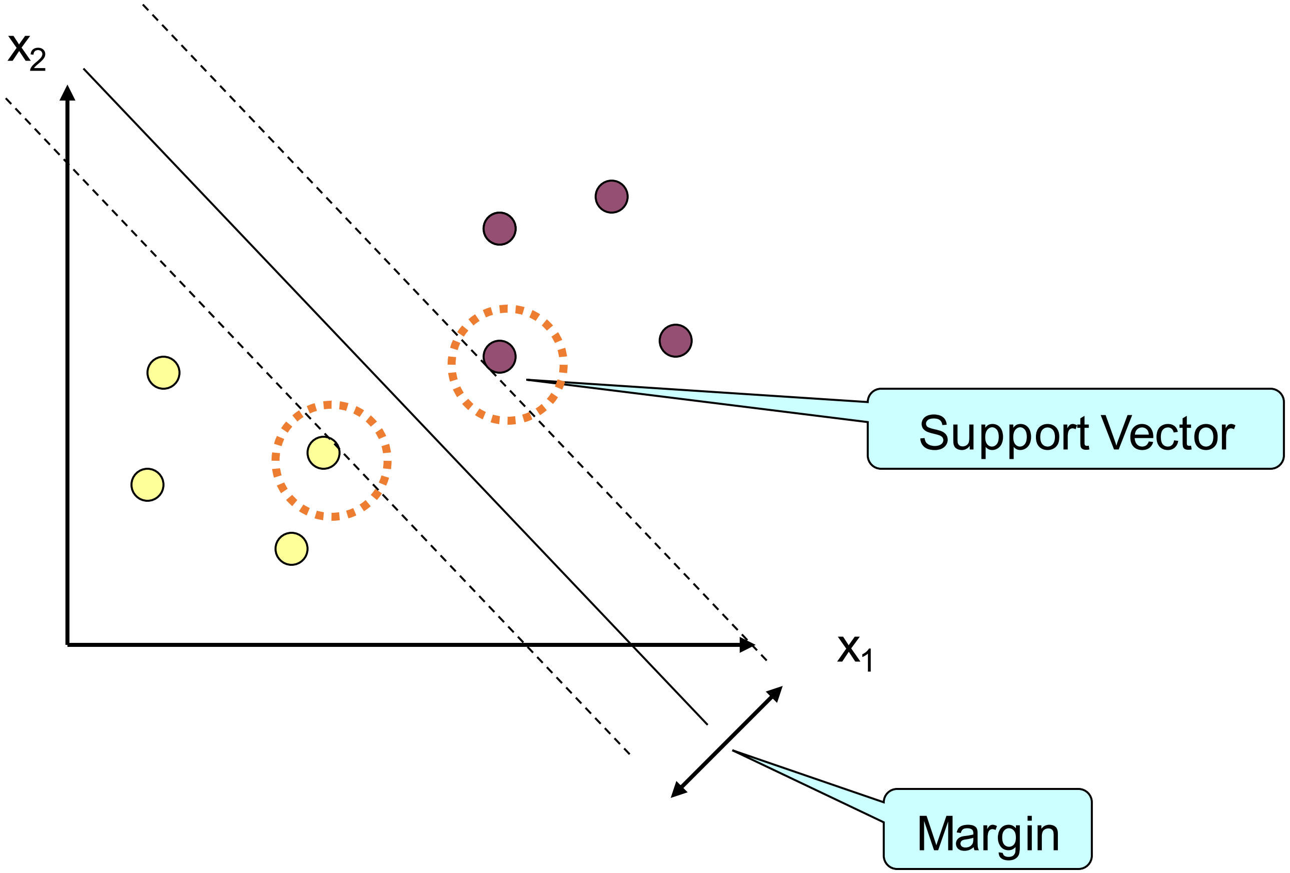
Linear Support Vector Machine



Linear Support Vector Machine



Linear Support Vector Machine

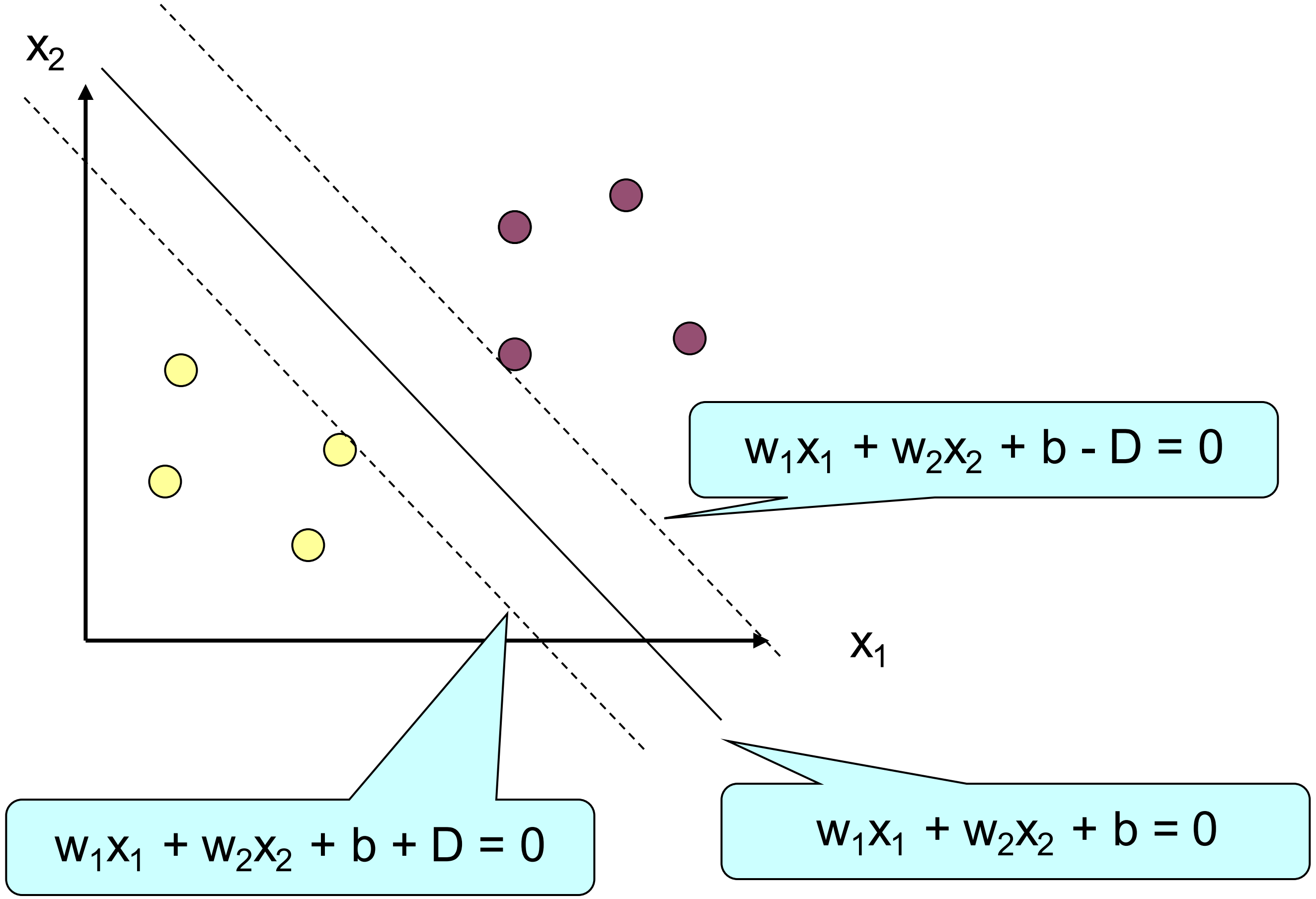


Support Vector

Margin

We want to maximize the margin | Why?

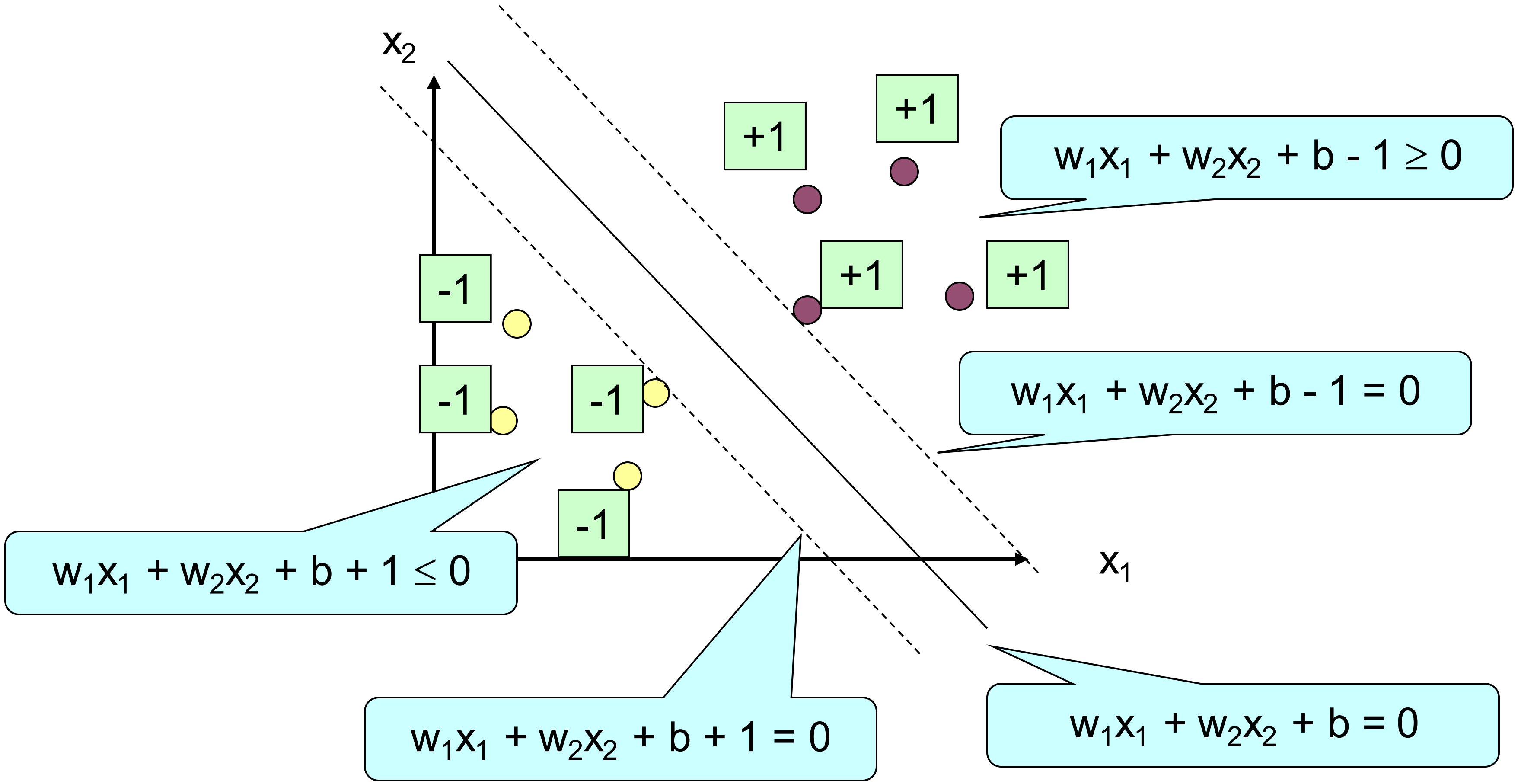
Linear Support Vector Machine



x_2
x_1
$w_1x_1 + w_2x_2 + b - D = 0$
$w_1x_1 + w_2x_2 + b + D = 0$
$w_1x_1 + w_2x_2 + b = 0$

Linear Support Vector Machine

Let y be the label of a point



$w_1x_1 + w_2x_2 + b - 1 \geq 0$

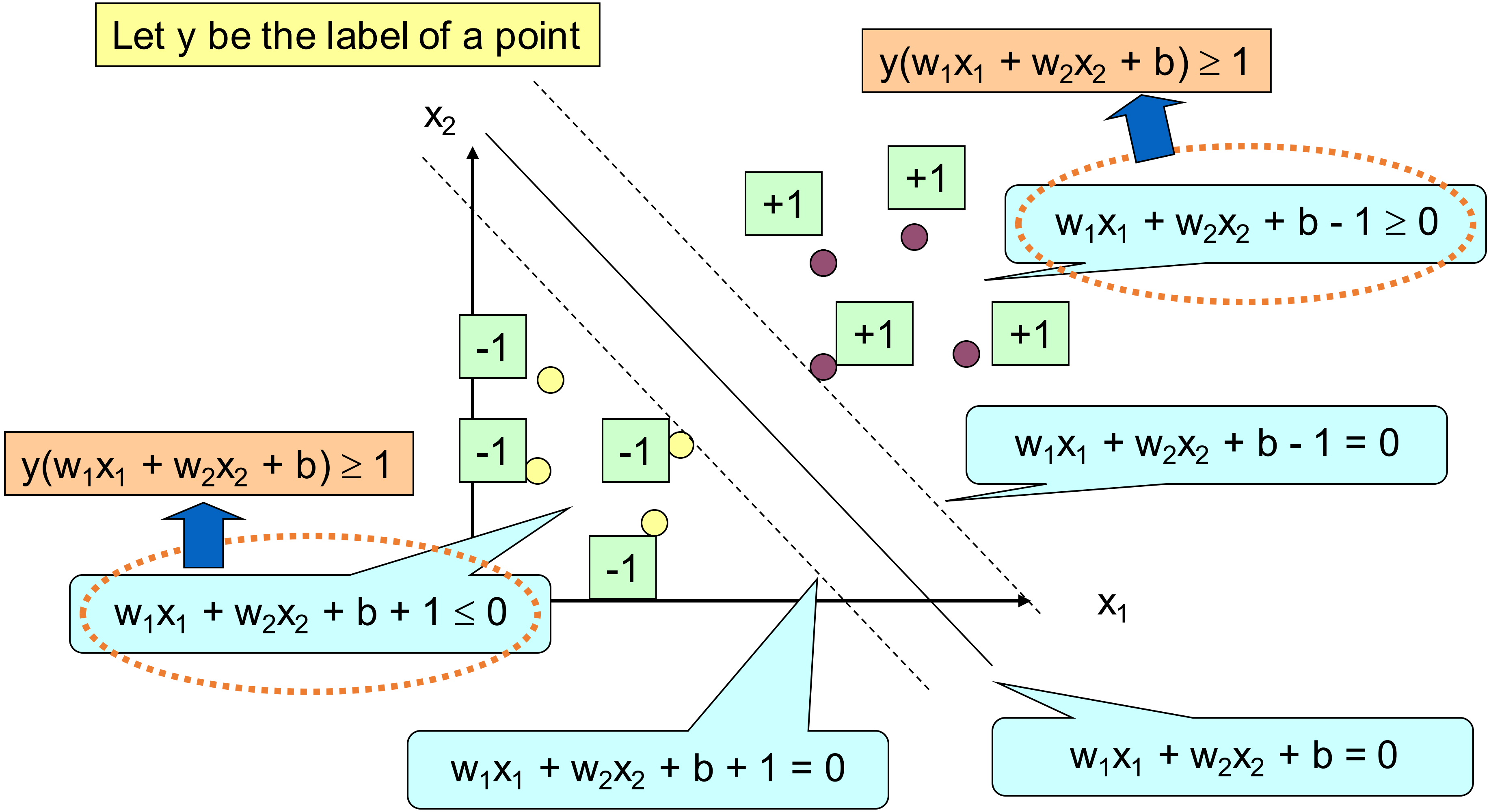
$w_1x_1 + w_2x_2 + b - 1 = 0$

$w_1x_1 + w_2x_2 + b + 1 \leq 0$

$w_1x_1 + w_2x_2 + b + 1 = 0$

$w_1x_1 + w_2x_2 + b = 0$

Linear Support Vector Machine



Let y be the label of a point

$y(w_1x_1 + w_2x_2 + b) \geq 1$

$w_1x_1 + w_2x_2 + b - 1 \geq 0$

$w_1x_1 + w_2x_2 + b - 1 = 0$

$y(w_1x_1 + w_2x_2 + b) \geq 1$

$w_1x_1 + w_2x_2 + b + 1 \leq 0$

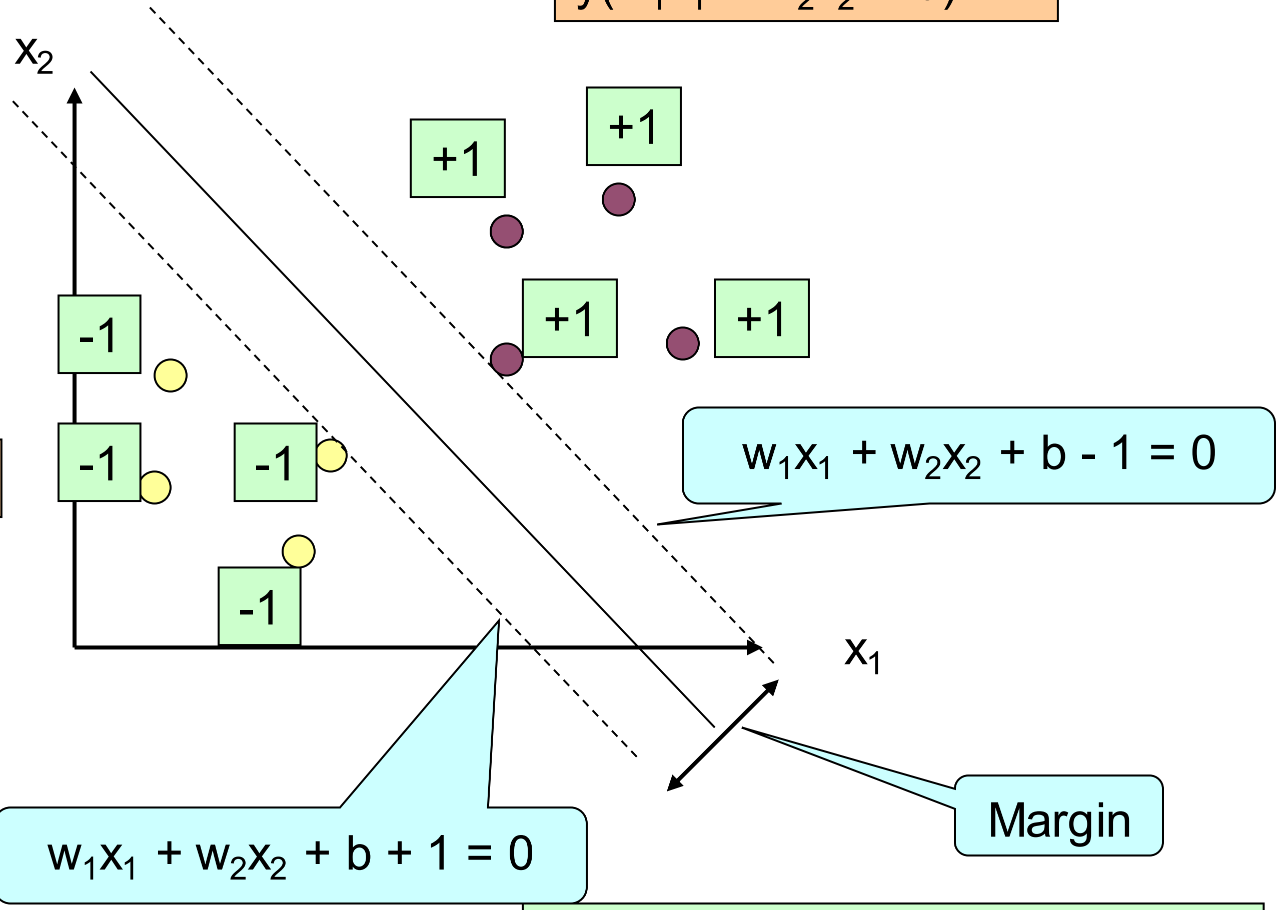
$w_1x_1 + w_2x_2 + b + 1 = 0$

$w_1x_1 + w_2x_2 + b = 0$

Linear Support Vector Machine

Let y be the label of a point

$y(w_1x_1 + w_2x_2 + b) \geq 1$



$w_1x_1 + w_2x_2 + b - 1 = 0$

$w_1x_1 + w_2x_2 + b + 1 = 0$

Margin

$y(w_1x_1 + w_2x_2 + b) \geq 1$

Margin

$$= \frac{|(b+1) - (b-1)|}{\sqrt{w_1^2 + w_2^2}}$$

$$= \frac{2}{\sqrt{w_1^2 + w_2^2}}$$

We want to maximize the margin

Linear Support Vector Machine

- Maximize $\text{Margin} = \frac{2}{\sqrt{w_1^2 + w_2^2}}$

- Subject to $\boxed{y(w_1x_1 + w_2x_2 + b) \geq 1}$

- for each data point (x_1, x_2, y) where y is the label of the point (+1/-1)

Linear Support Vector Machine

- Minimize $w_1^2 + w_2^2$

Quadratic objective

Linear constraints

- Subject to $y(w_1x_1 + w_2x_2 + b) \geq 1$

- for each data point (x_1, x_2, y), where y is the label of the point (+1/-1)

Quadratic programming

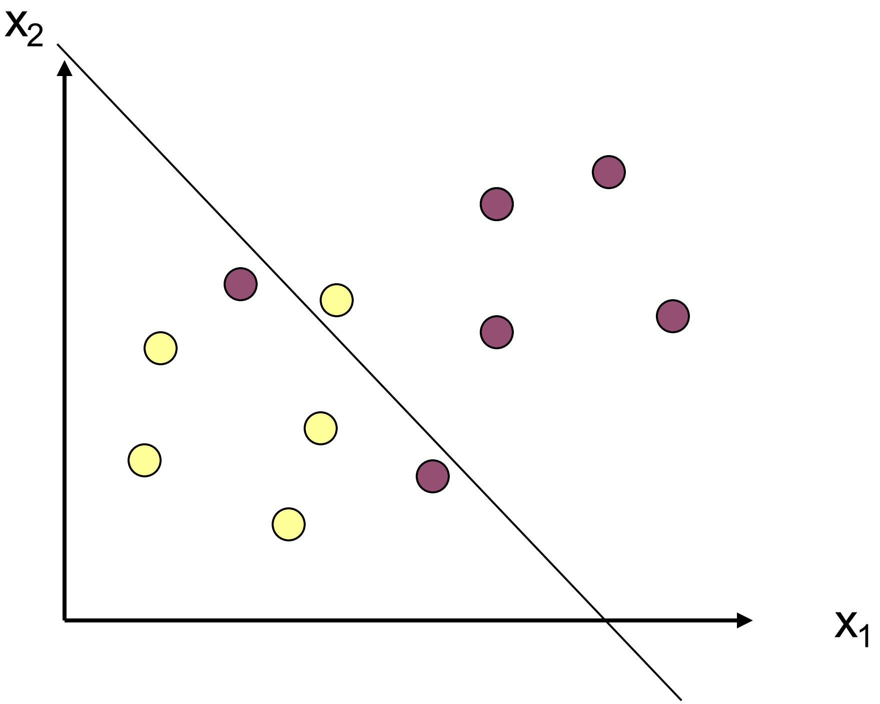
Linear Support Vector Machine

- We have just described 2-dimensional space
- We can divide the space into two parts by a line
- For n-dimensional space where n >=2,
  - We use a **hyperplane** to divide the space into two parts

Support Vector Machine

- Support Vector Machine (SVM)
  - Linear Support Vector Machine
  - Non-linear Support Vector Machine

Non-linear Support Vector Machine



Non-linear Support Vector Machine

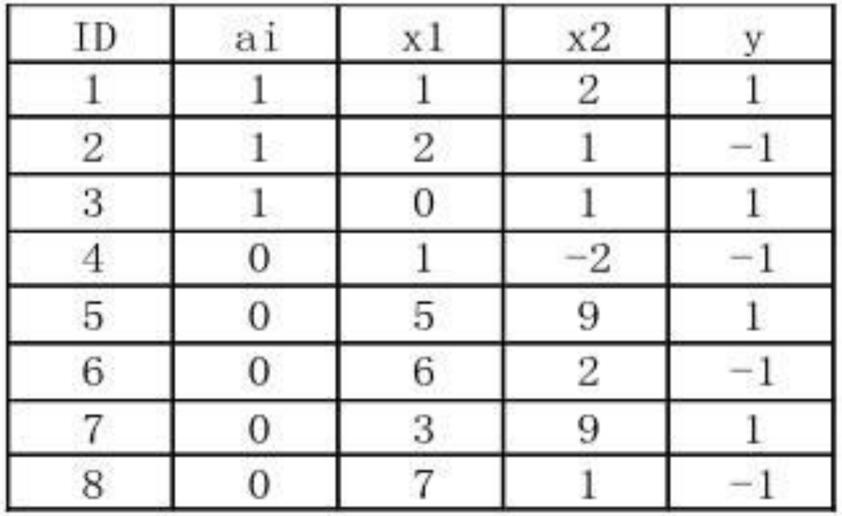
- Two Steps
  - **Step 1:** Transform the data into a higher dimensional space using a "nonlinear" mapping
  - **Step 2:** Use the Linear Support Vector Machine in this high-dimensional space

Non-linear Support Vector Machine

- Rationale
  - With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane
  - SVM finds this hyperplane using support vectors ("essential" training tuples) and margins (defined by the support vectors)

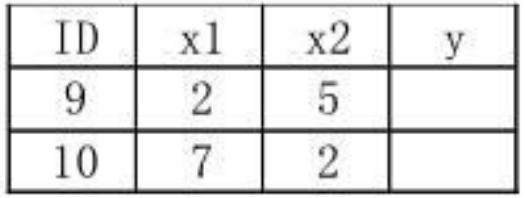
SVM

- Consider the following data points. Please use SVM to train a classifier, and then classify these data points. Points with a_i=1 means this point is **support vector**. For example, point 1 (1,2) is the support vector, but point 5 (5,9) is not the support vector.
- Training data:



ID	ai	x1	x2	y
1	1	1	2	1
2	1	2	1	−1
3	1	0	1	1
4	0	1	−2	−1
5	0	5	9	1
6	0	6	2	−1
7	0	3	9	1
8	0	7	1	−1

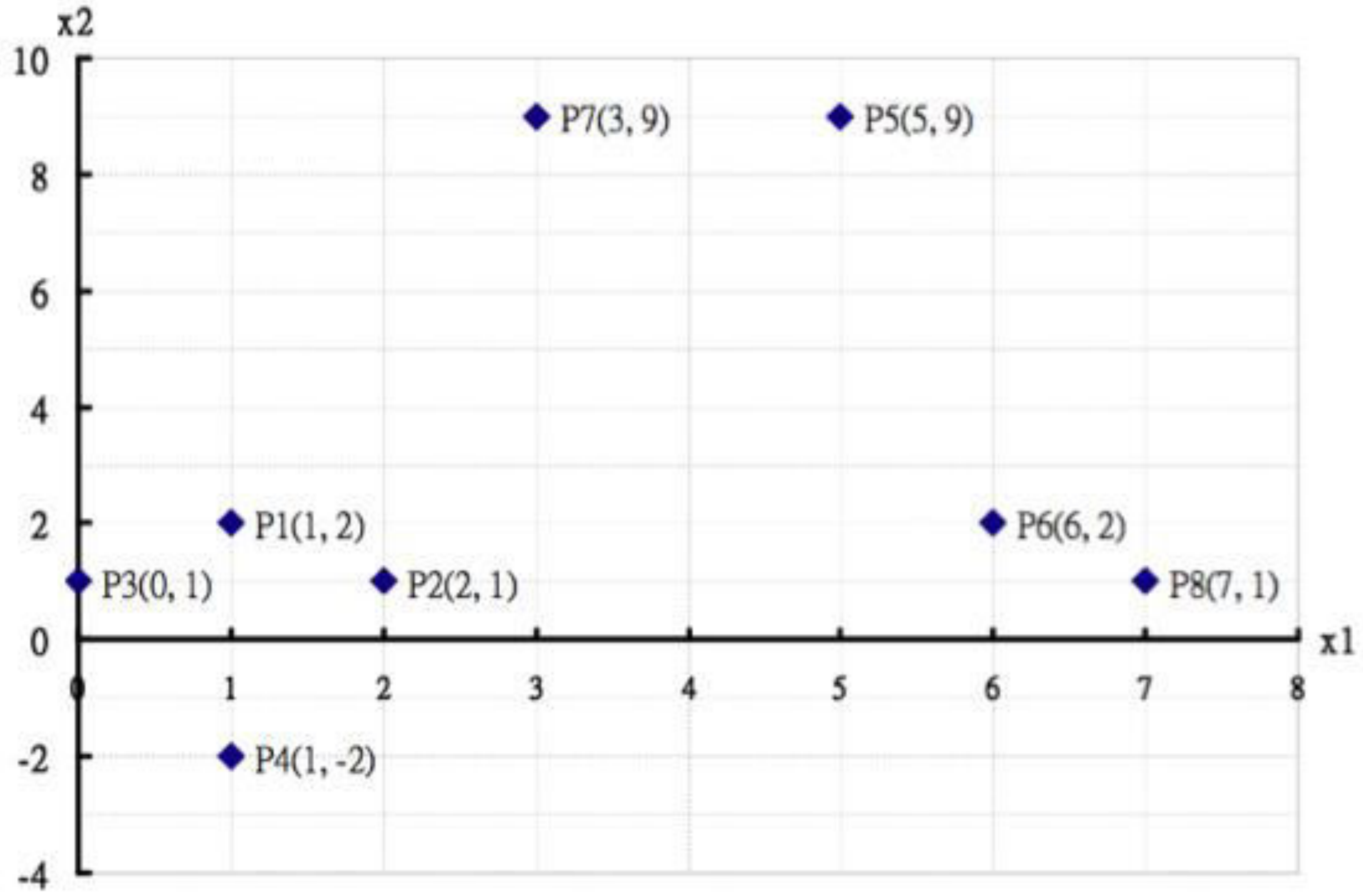
- Testing data:



ID	x1	x2	y
9	2	5	
10	7	2	

SVM

- Answer:
- a) As the picture shows, P1, P2, P3 are support vectors.



SVM

- Suppose w is (w_1,w_2). Since both P1(1,2) and P3(0,1) have y = 1, while P2(2,1) has y =-1:
  - $w_1*1+w_2*2+b = 1$
  - $w_1*0+w_2*1+b = 1$
  - $w_1*2+w_2*1+b =-1$

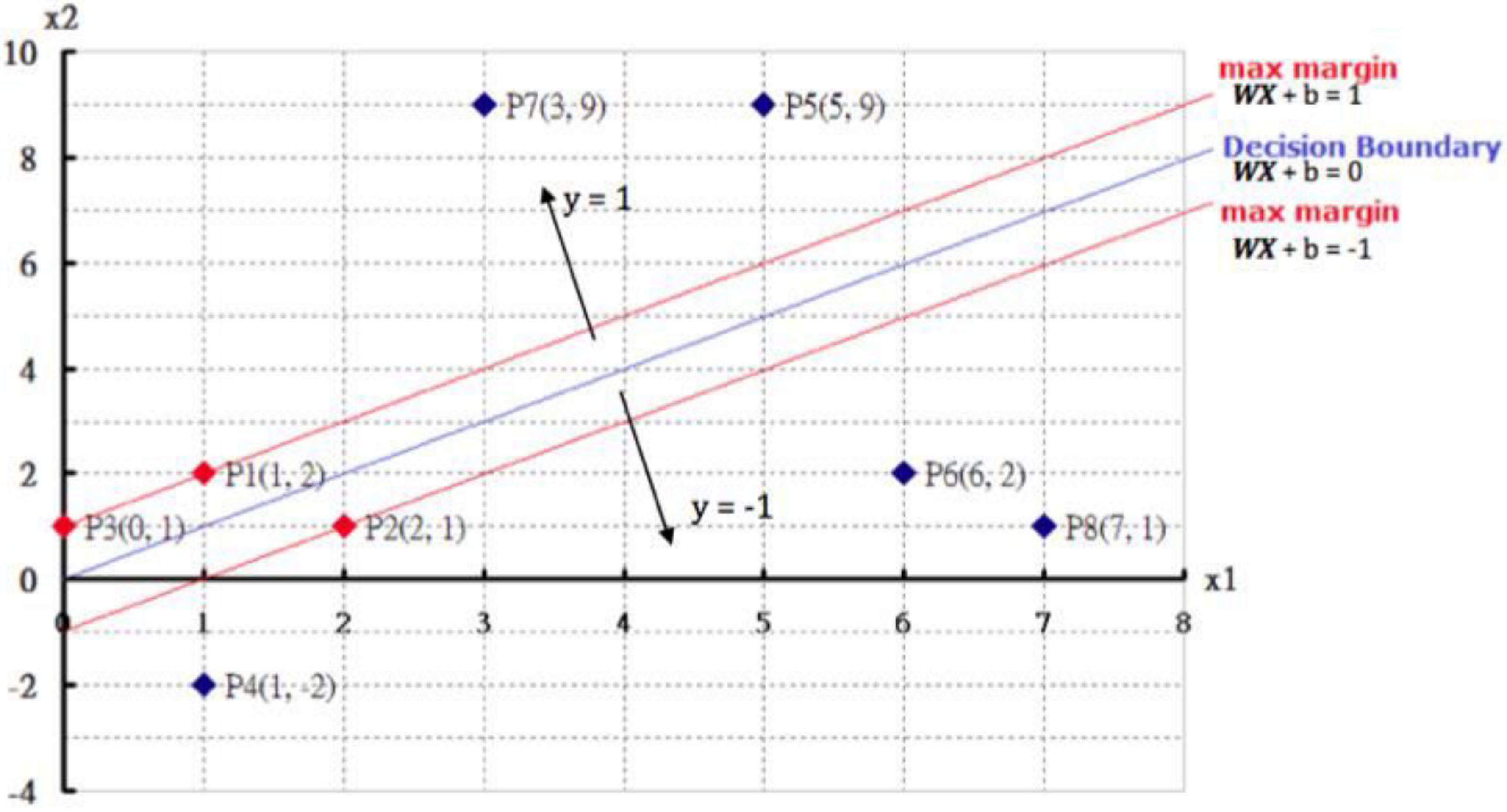
  ⇨$w_1= -1,\ w_2 = 1,\ b = 0$

  then, the decision boundary is:
  - $w_1 * x_1+w_2 * x_2 + b =0$

  ⇨**-x1+x2 = 0**

- Showed in the picture next page.

SVM


x2
10
8
6
4
2
0
-2
-4
x1
0
1
2
3
4
5
6
7
8
P7(3, 9)
P5(5, 9)
P1(1, 2)
P3(0, 1)
P2(2, 1)
P6(6, 2)
P8(7, 1)
P4(1, -2)
y = 1
y = -1
max margin
WX + b = 1
Decision Boundary
WX + b = 0
max margin
WX + b = -1

SVM

- b) Use the decision boundary to classify the testing data:
  - For the point P9 (2,5)

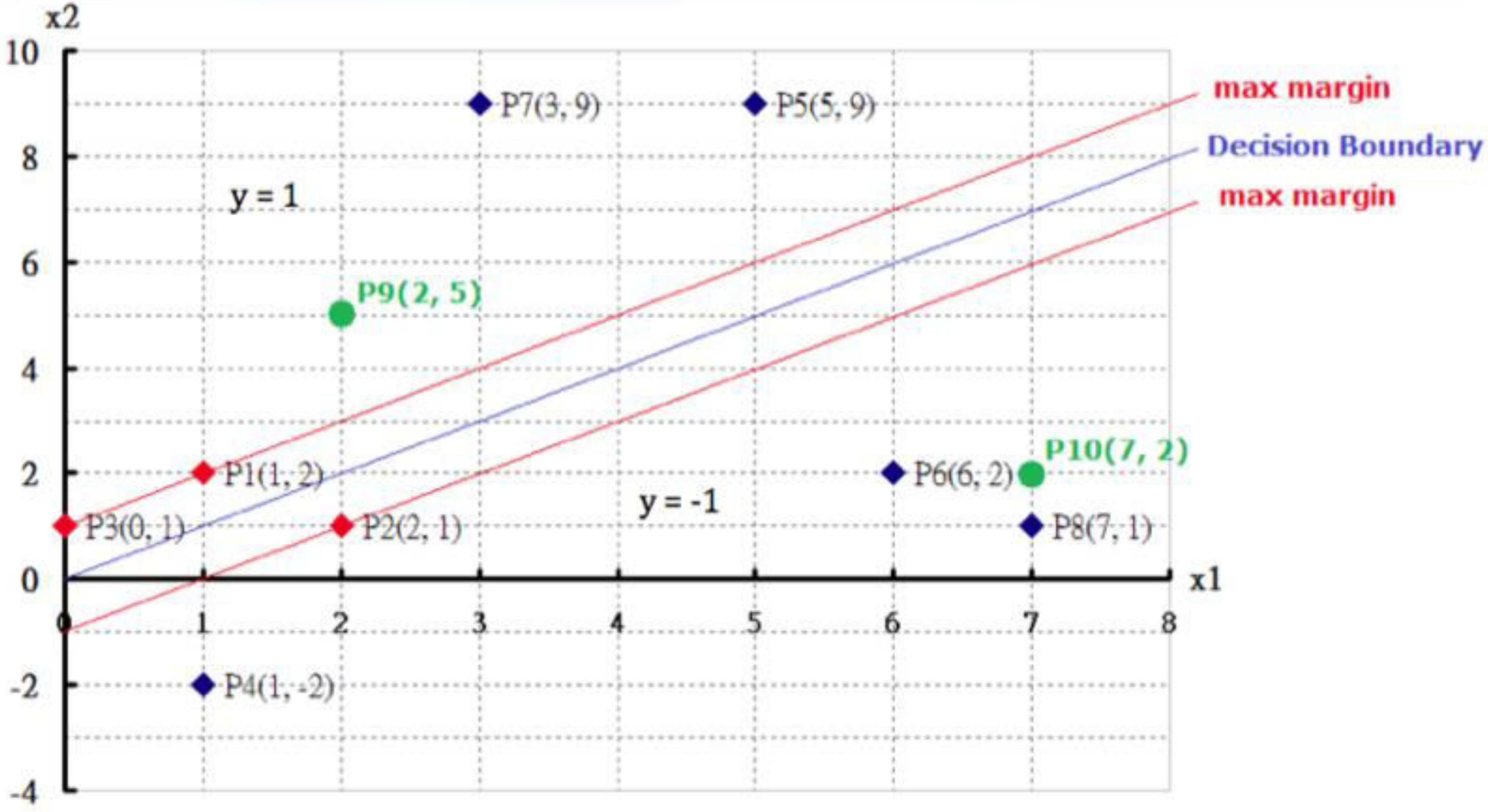
    $-x_1+x_2 = -2+5 = 3 >= 1$

    So we choose y = 1
  - For the point P10 (7,2)

    $-x_1+x_2 = -7+2 = -5 <= -1$

    So we choose y = -1
- Showed in the picture next page.

SVM


x2
10
8
6
4
2
0
-2
-4
x1
0
1
2
3
4
5
6
7
8
max margin
Decision Boundary
max margin
y = 1
y = -1
P7(3, 9)
P5(5, 9)
P9(2, 5)
P1(1, 2)
P3(0, 1)
P2(2, 1)
P6(6, 2)
P10(7, 2)
P8(7, 1)
P4(1, -2)

SVM

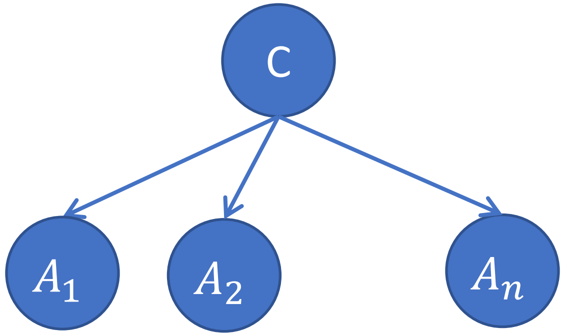
- Advantages:
  - Can be visualized
  - Accurate when the data is well partitioned
  - Fast evaluation of the learned target function
- Disadvantages:
  - Long training time
  - Difficult to understand the learned function (weight)
  - Not easy to incorporate domain knowledge

Effectiveness of SVM on High Dimensional Data

- The complexity of trained classifier is characterized by the # of support vectors rather than the dimensionality of the data
- If all other training examples are removed and the training is repeated, the same separating hyperplane would be found
- The number of support vectors found can be used to compute an (upper) bound on the expected error rate of the SVM classifier, which is independent of the data dimensionality
- Thus, an SVM with a small number of support vectors can have good generalization, even when the dimensionality of the data is high

Generative vs Discriminative models

- Naïve Bayes is a type of a generative model
  - Generative process:
    - First pick the category of the record
    - Then given the category, generate the attribute values from the distribution of the category
    - Conditional independence given C



- We use the training data to learn the distribution of the values in a class

Generative vs Discriminative models

- Logistic Regression and SVM are discriminative models
  - The goal is to find the boundary that discriminates between the two classes from the training data

- In order to classify the language of a document, you can
  - Either learn the two languages and find which is more likely to have generated the words you see
  - Or learn what differentiates the two languages.