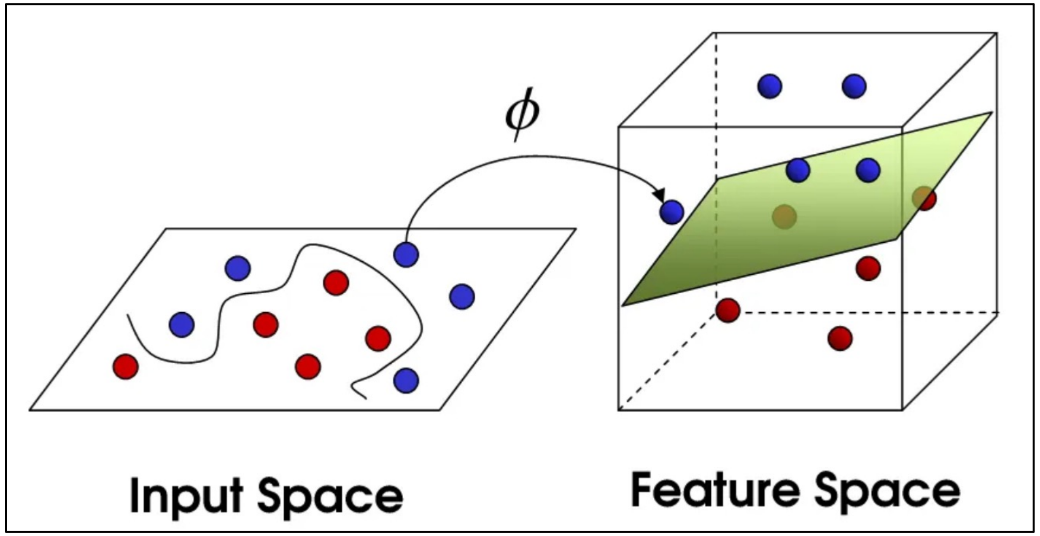
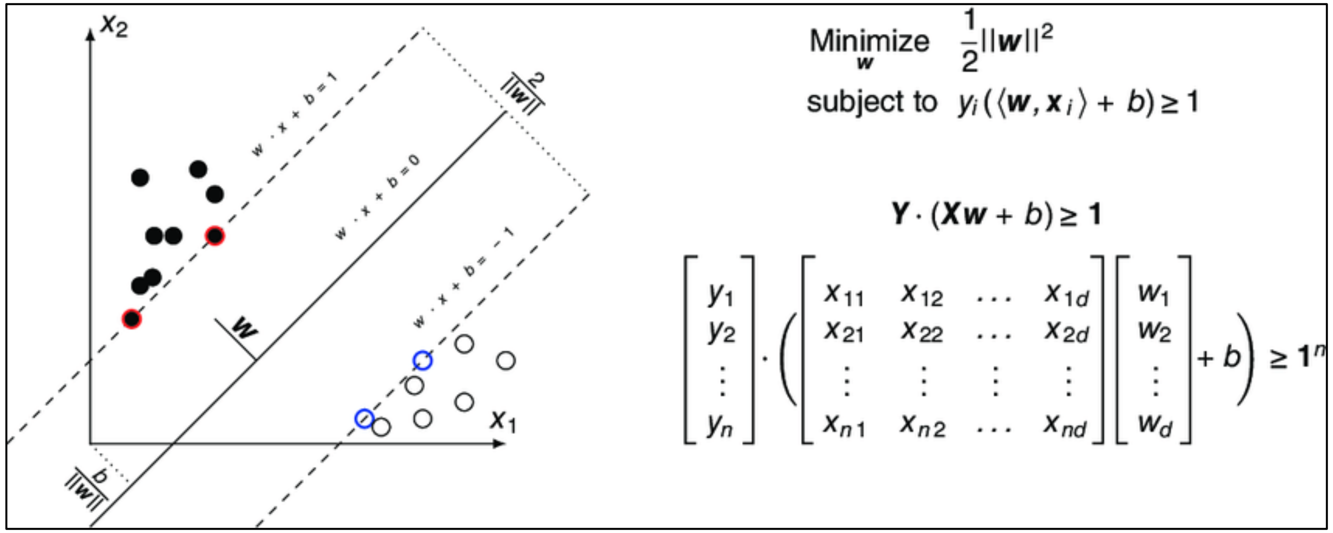
Support Vector Machine (First Look)





Vinh Dinh Nguyen
PhD in Computer Science

Year 2023

Outline

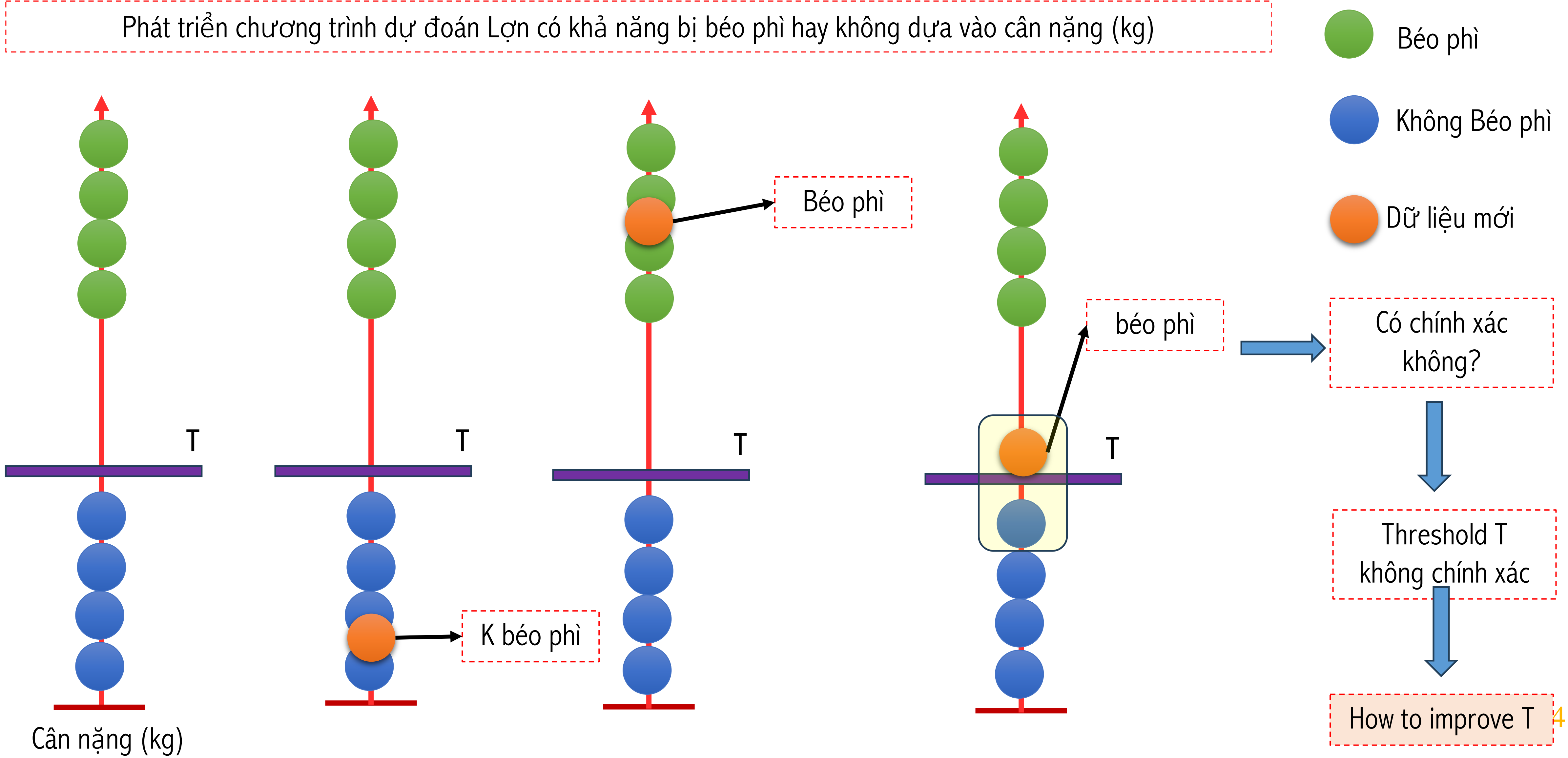
- Maximal Margin Classifier
- Support Vector Classifier
- Support Vector Machine
- Polynomial Kernel
- Radial Basic Function Kernel (RBF)
- Example

Outline

- Maximal Margin Classifier
- Support Vector Classifier
- Support Vector Machine
- Polynomial Kernel
- Radial Basic Function Kernel (RBF)
- Example

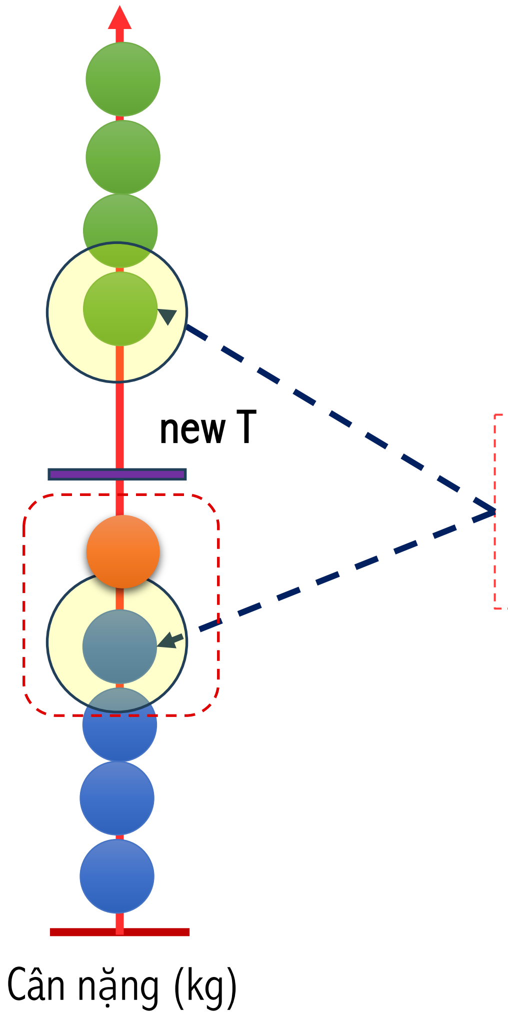
SVM Motivation

Phát triển chương trình dự đoán Lợn có khả năng bị béo phì hay không dựa vào cân nặng (kg)

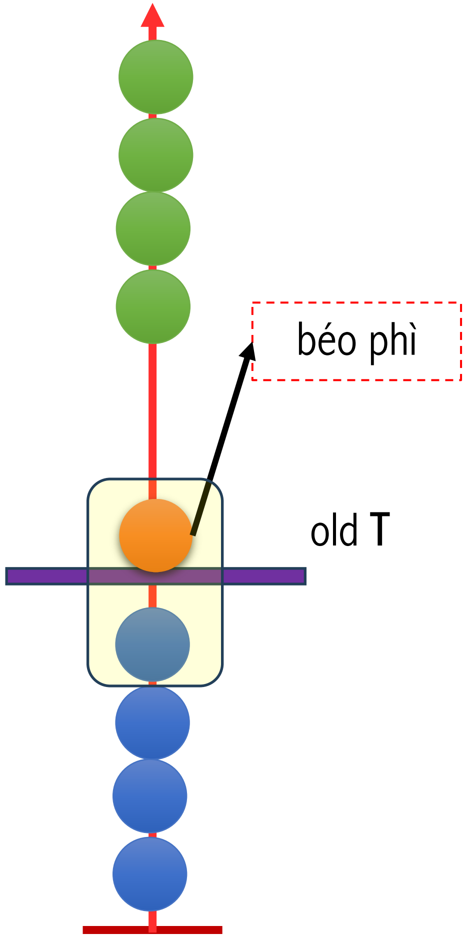


SVM Motivation

Phát triển chương trình dự đoán Lợn có khả năng bị béo phì hay không dựa vào cân nặng (kg)



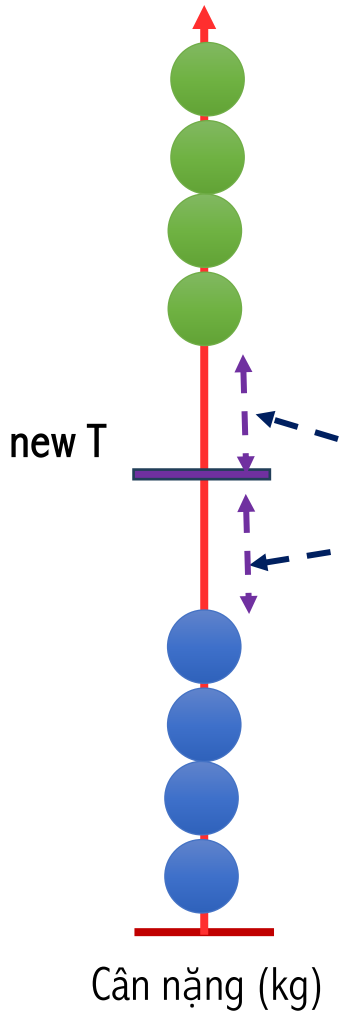
Quan sát 2 đối tượng biên của mỗi cluster. Xác định threshold T mới.
Kết quả dự đoán chính xác



Kết quả dự đoán không chính xác

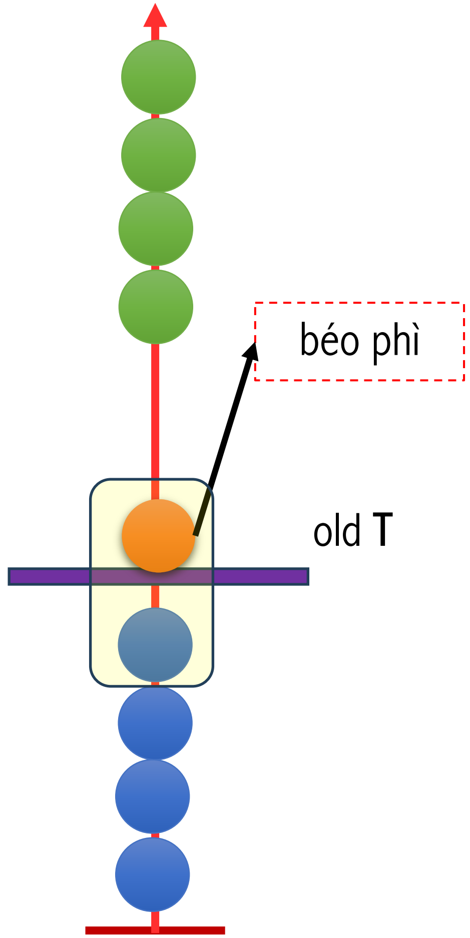
SVM Motivation

Phát triển chương trình dự đoán Lợn có khả năng bị béo phì hay không dựa vào cân nặng (kg)



The shortest distance between the obervations and T threshold is called the margin

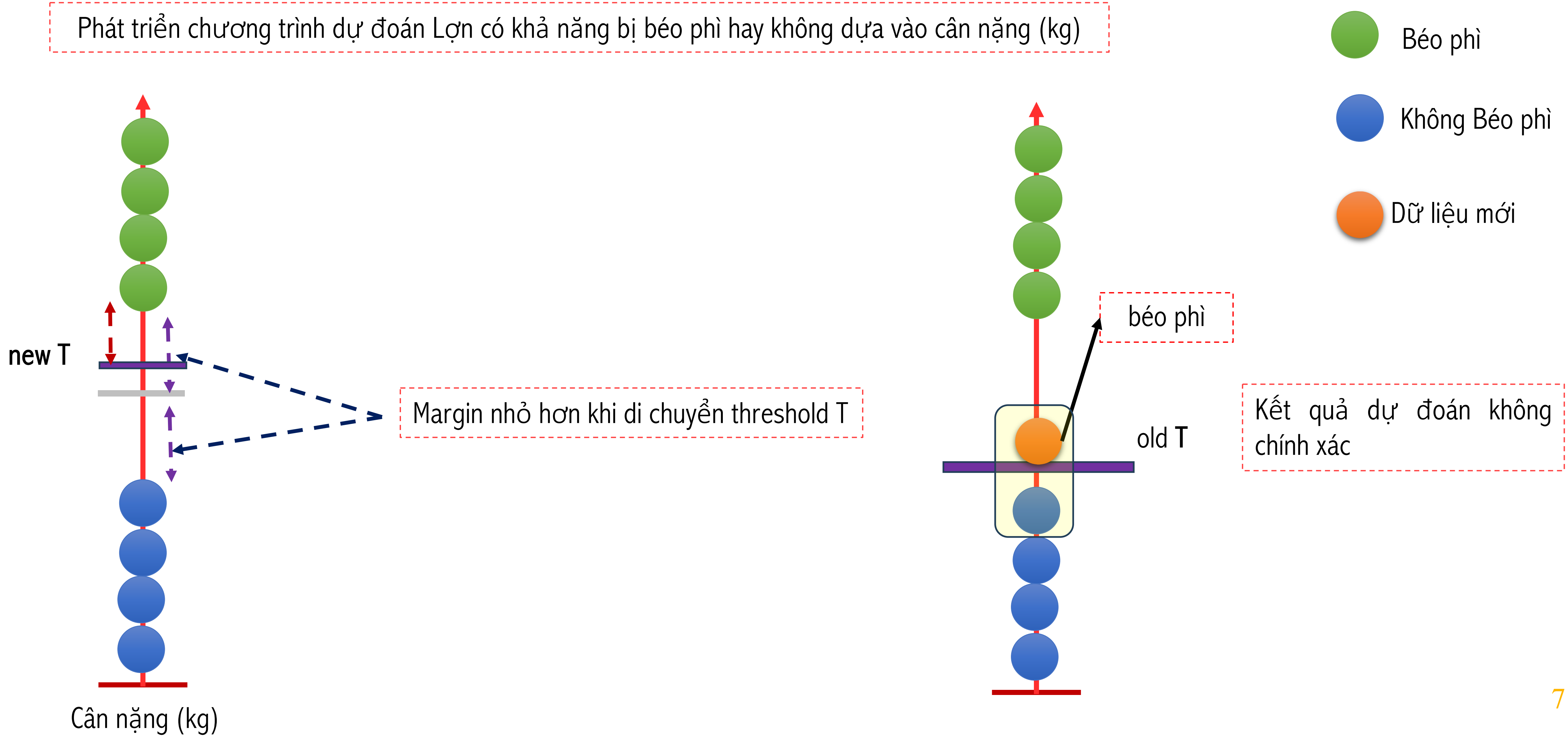
Margin đạt giá trị lớn nhất trong trường hợp T nằm giữa 2 observations



Kết quả dự đoán không chính xác

SVM Motivation

Phát triển chương trình dự đoán Lợn có khả năng bị béo phì hay không dựa vào cân nặng (kg)



SVM Motivation

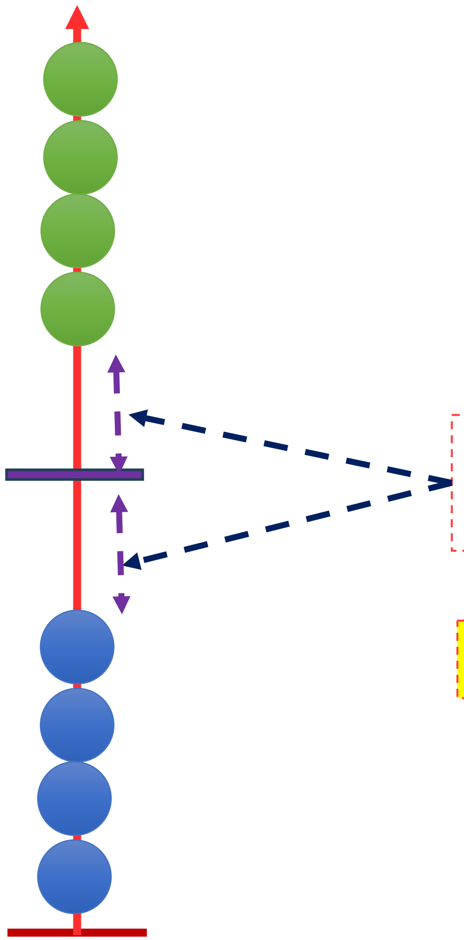
Phát triển chương trình dự đoán Lợn có khả năng bị béo phì hay không dựa vào cân nặng (kg)

Béo phì

Không Béo phì

Dữ liệu mới

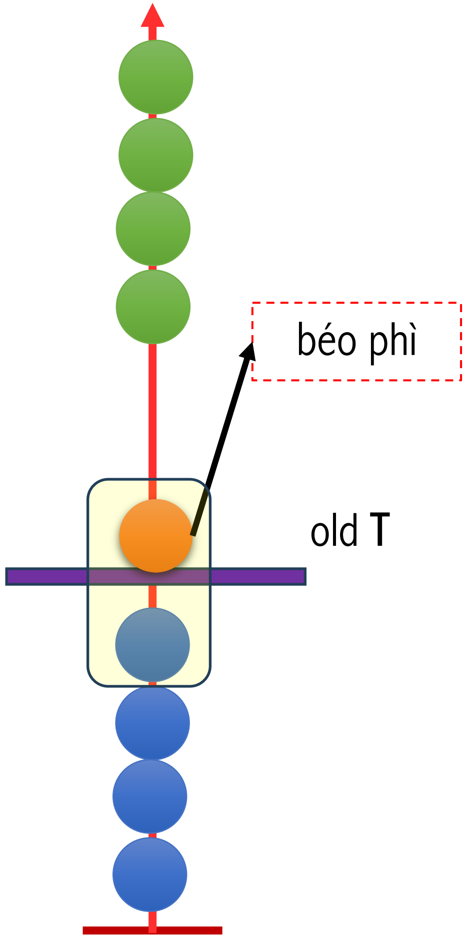
new T



Cân nặng (kg)

This margin give us the largest margin to make classification

Maximal Margin Classifier



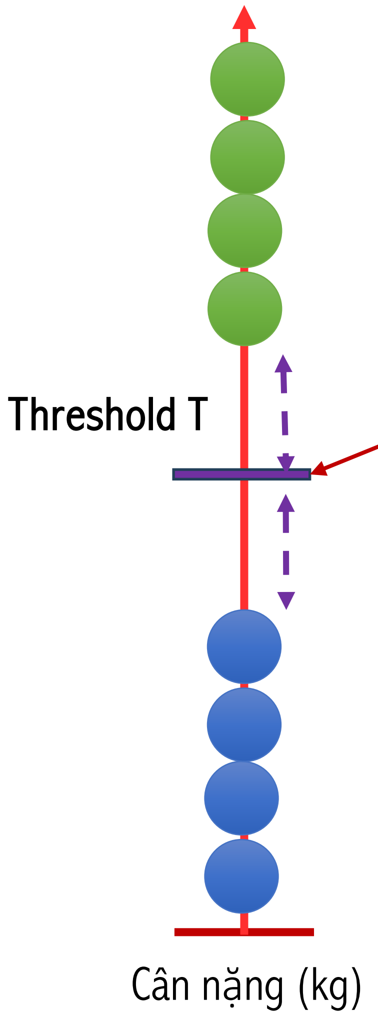
Kết quả dự đoán không chính xác

Outline

- Maximal Margin Classifier
- Support Vector Classifier
- Support Vector Machine
- Polynomial Kernel
- Radial Basic Function Kernel (RBF)
- Example

SVM Motivation

Phát triển chương trình dự đoán Lợn có khả năng bị béo phì hay không dựa vào cân nặng (kg)



Maximal Margin Classifier

↓

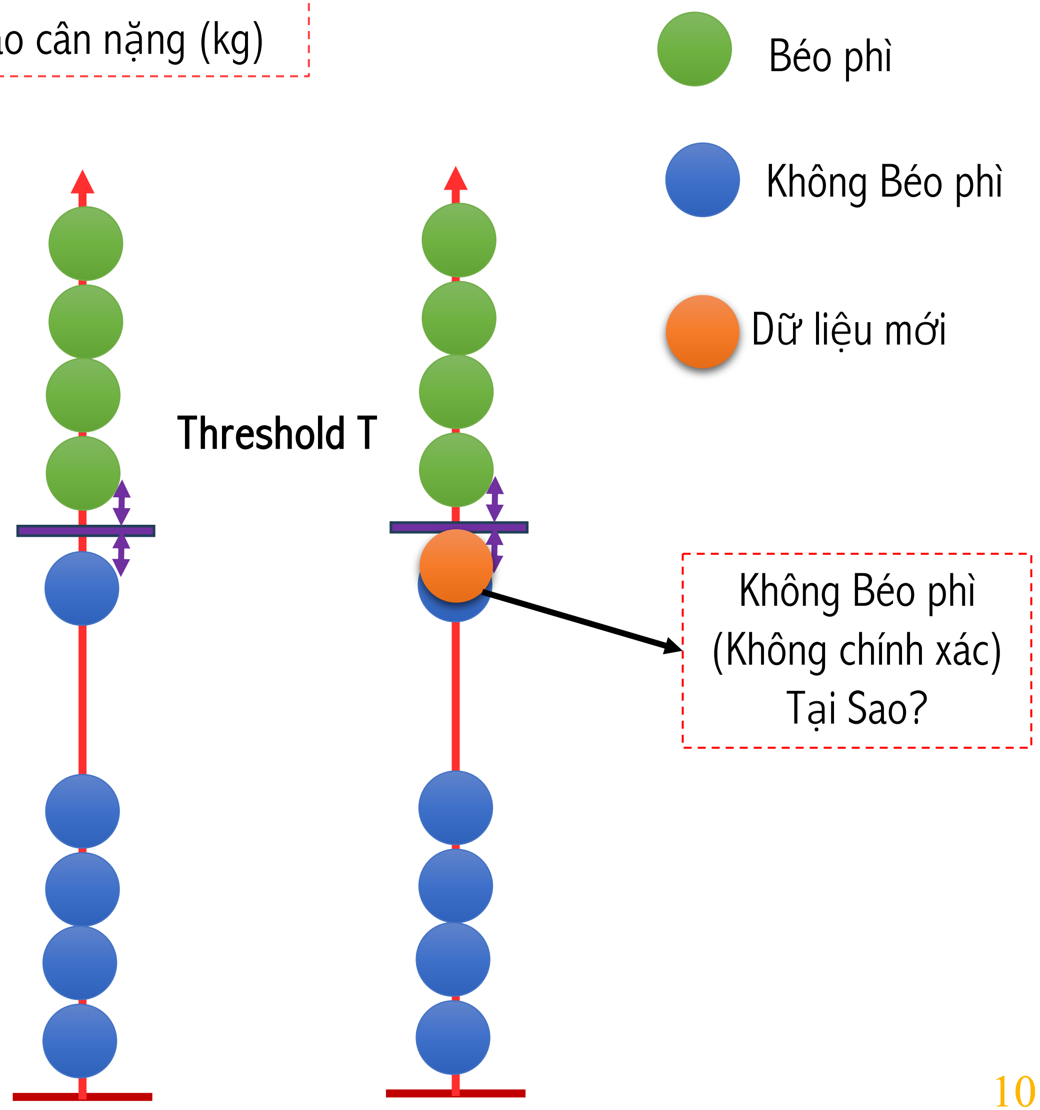
MMC is sensitive to outliers in the training data

↓

Can we select a Threshold T that not sensitive to outliers

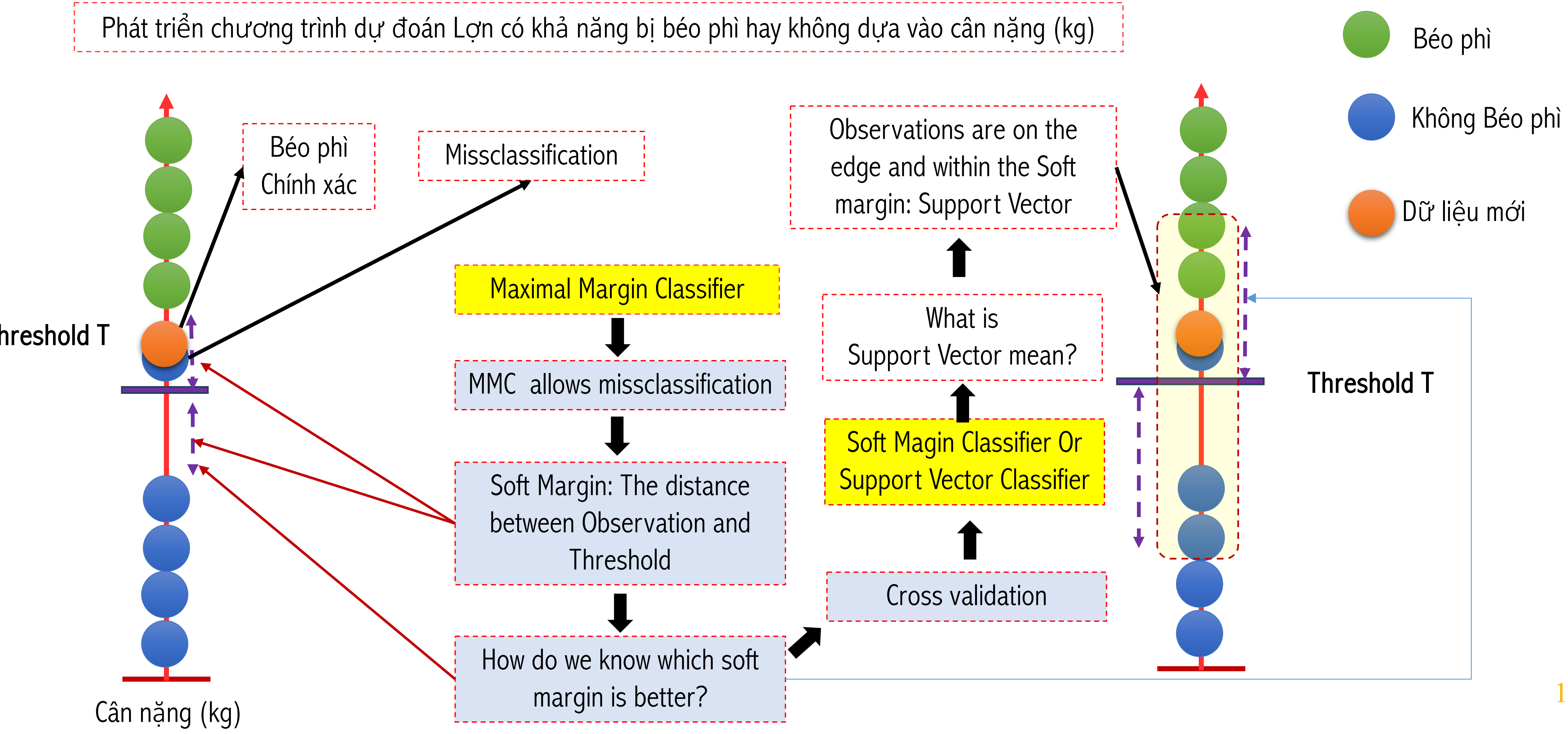
↓

Yes. Allow missclassification

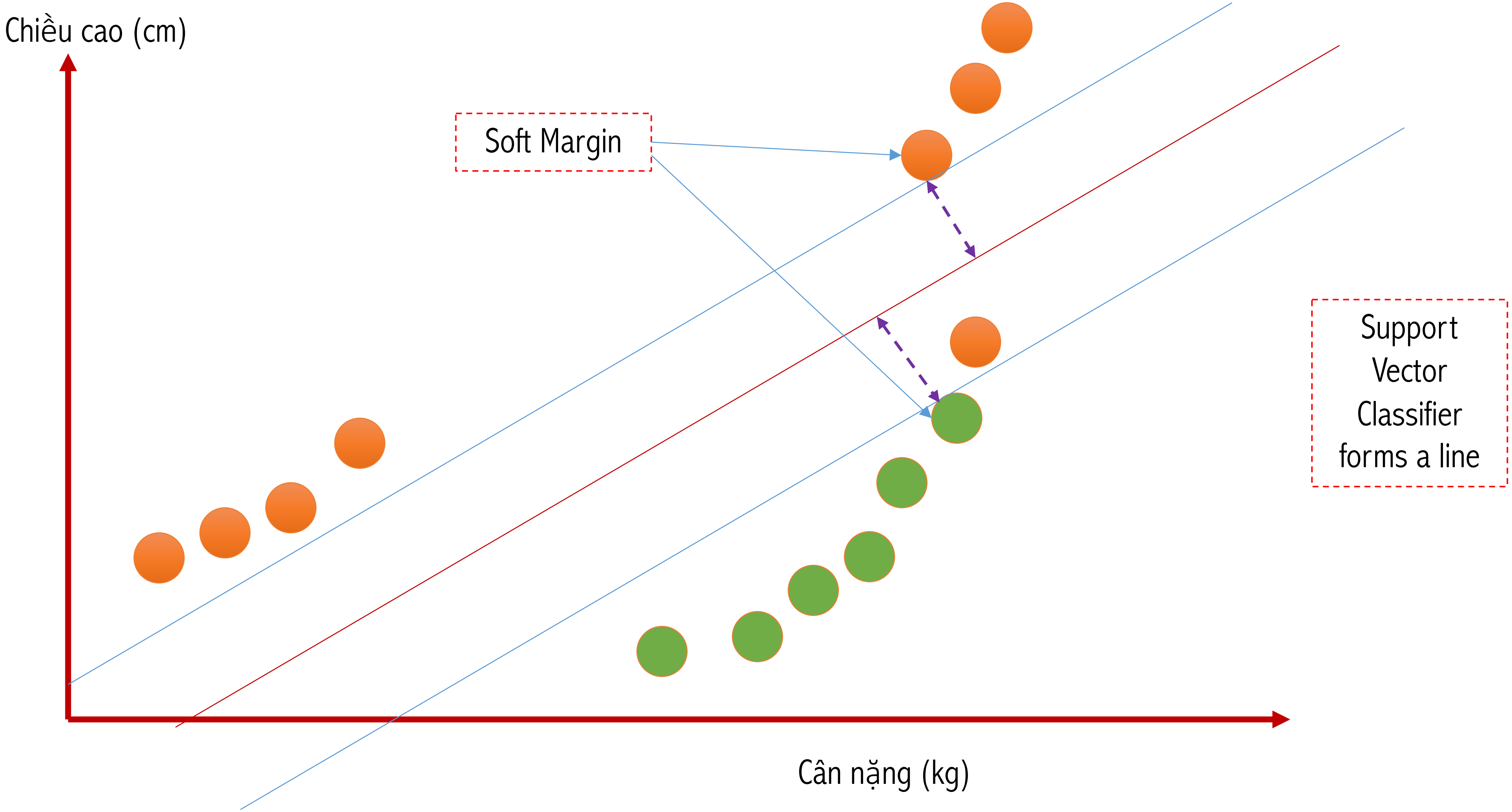


SVM Motivation

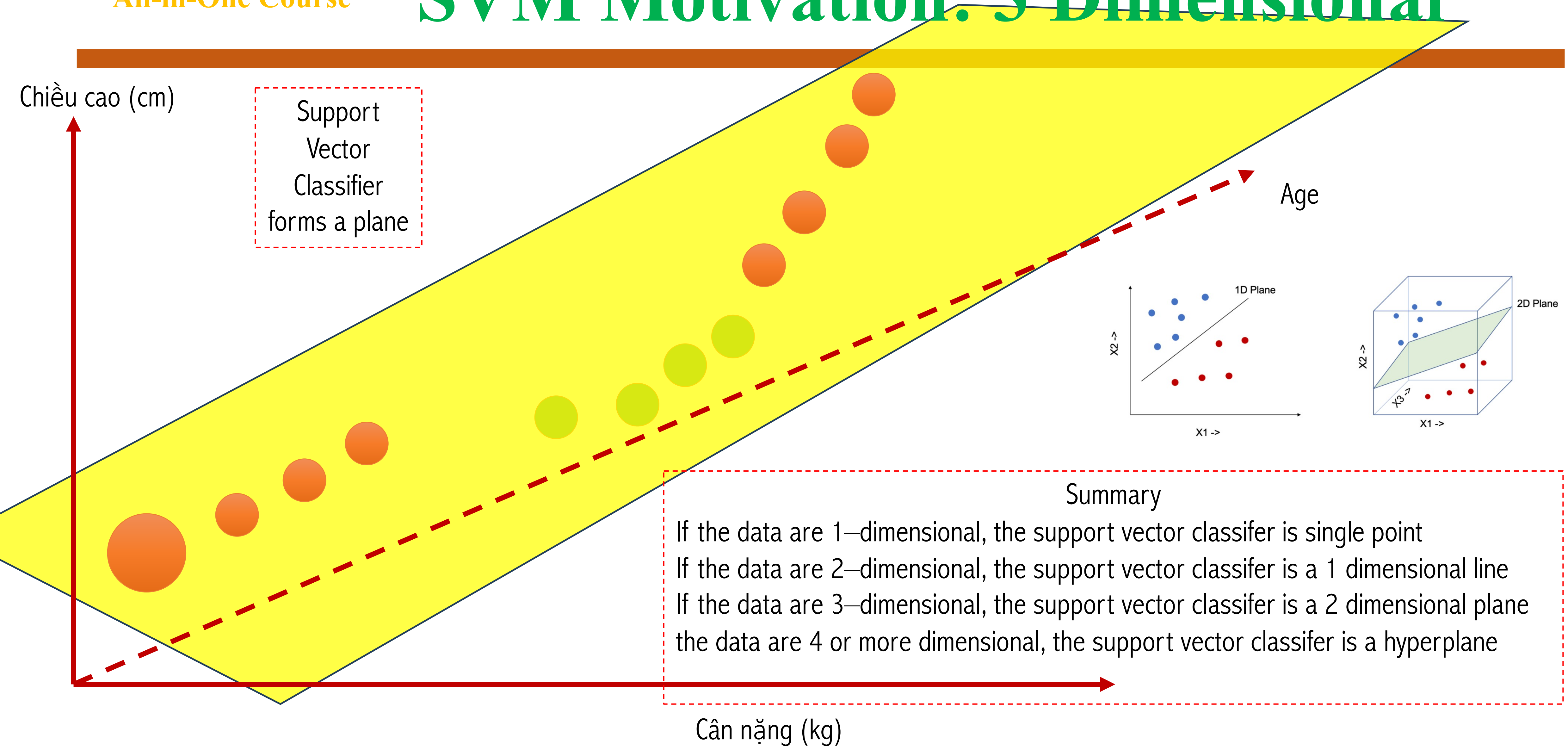
Phát triển chương trình dự đoán Lợn có khả năng bị béo phì hay không dựa vào cân nặng (kg)



SVM Motivation: 2 Dimensional



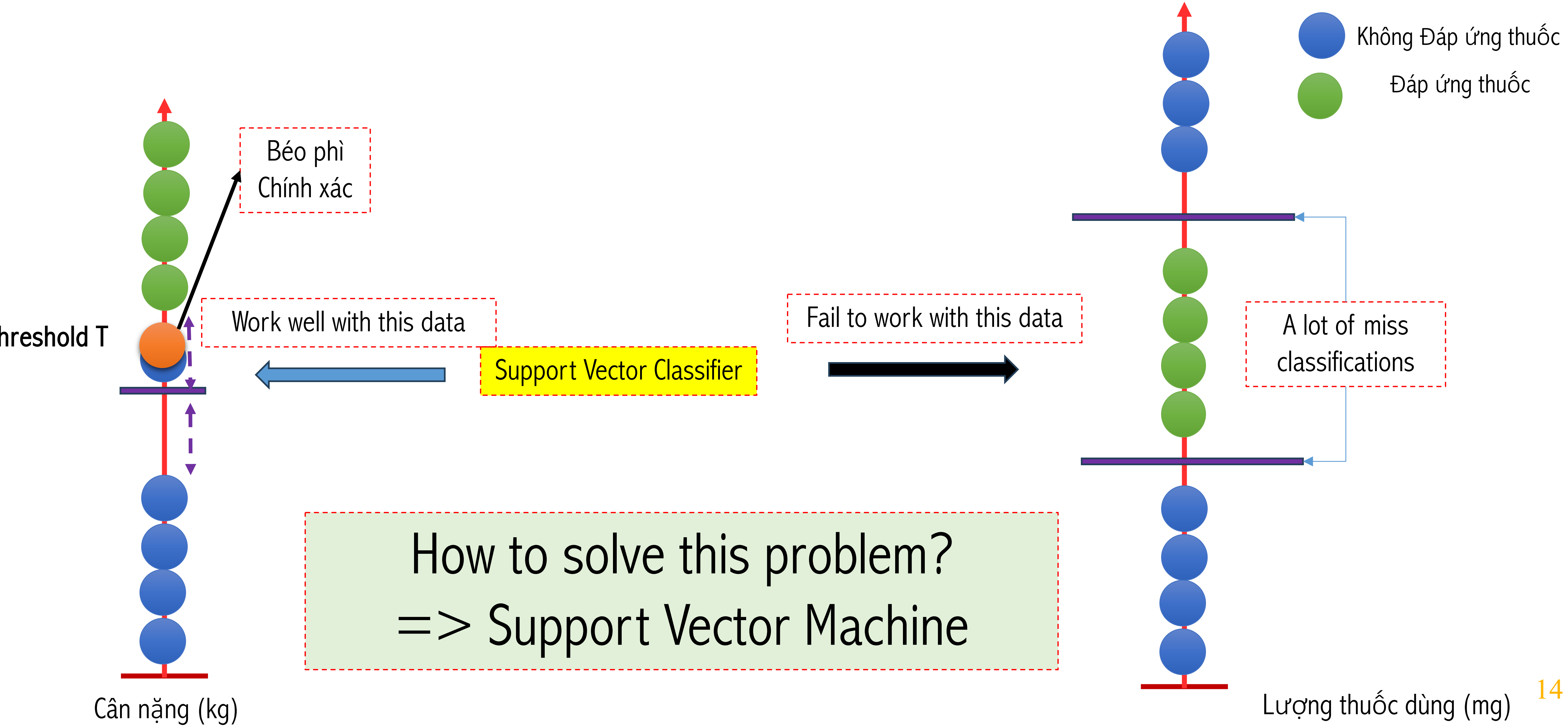
SVM Motivation: 3 Dimensional



Summary

If the data are 1–dimensional, the support vector classifer is single point
If the data are 2–dimensional, the support vector classifer is a 1 dimensional line
If the data are 3–dimensional, the support vector classifer is a 2 dimensional plane
the data are 4 or more dimensional, the support vector classifer is a hyperplane

Support Vector Classifier: Limitation

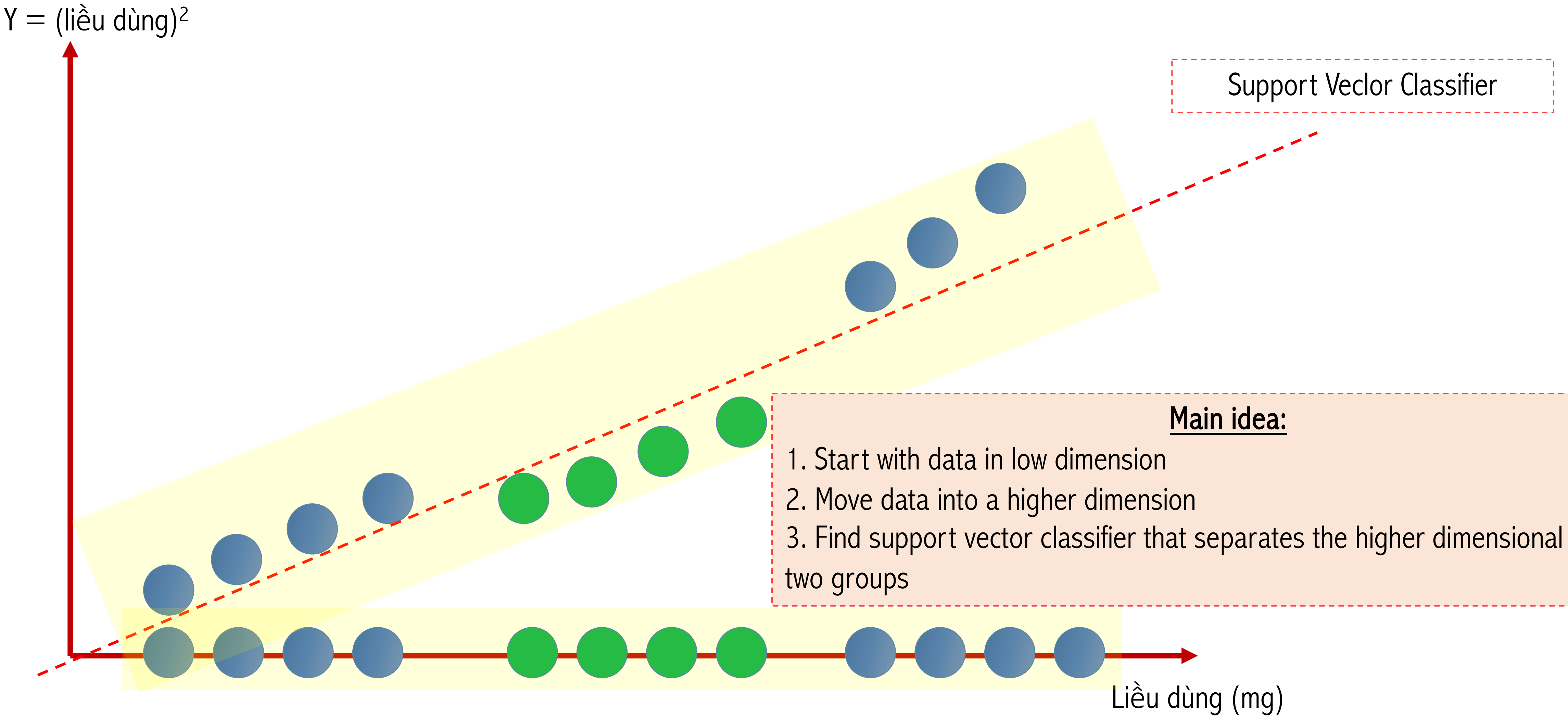


How to solve this problem?
=> Support Vector Machine

Outline

- Maximal Margin Classifier
- Support Vector Classifier
- Support Vector Machine
- Polynomial Kernel
- Radial Basic Function Kernel (RBF)
- Example

Support Vector Machine: Main Idea

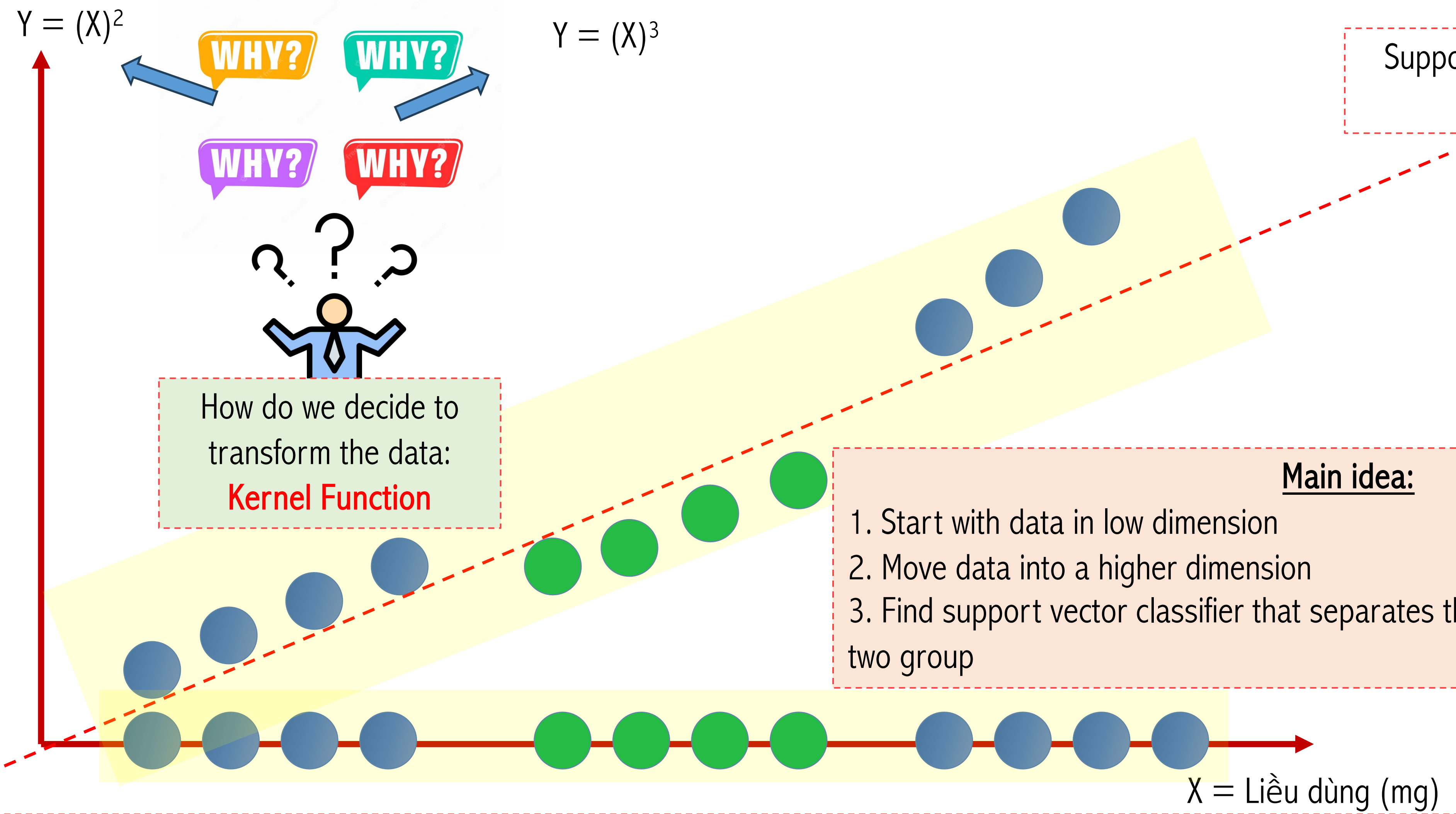


Main idea:

1. Start with data in low dimension
2. Move data into a higher dimension
3. Find support vector classifier that separates the higher dimensional data into two groups

AI VIETNAM
All-in-One Course

Support Vector Machine: Kernel Function



Support Veclor Classfier find this line

How do we decide to transform the data: **Kernel Function**

Main idea:

1. Start with data in low dimension
2. Move data into a higher dimension
3. Find support vector classifier that separates the higher dimensional data into two group

Kernel functions: None-linear functions that help us to transform data from lower dimension to higher dimension

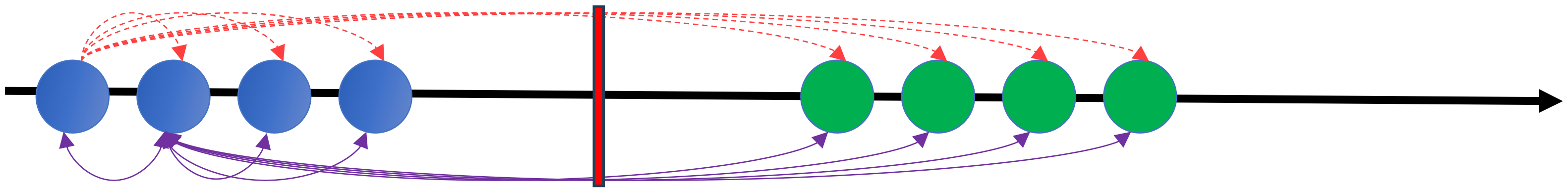
Outline

- Maximal Margin Classifier
- Support Vector Classifier
- Support Vector Machine
- Polynomial Kernel
- Radial Basic Function Kernel (RBF)
- Example

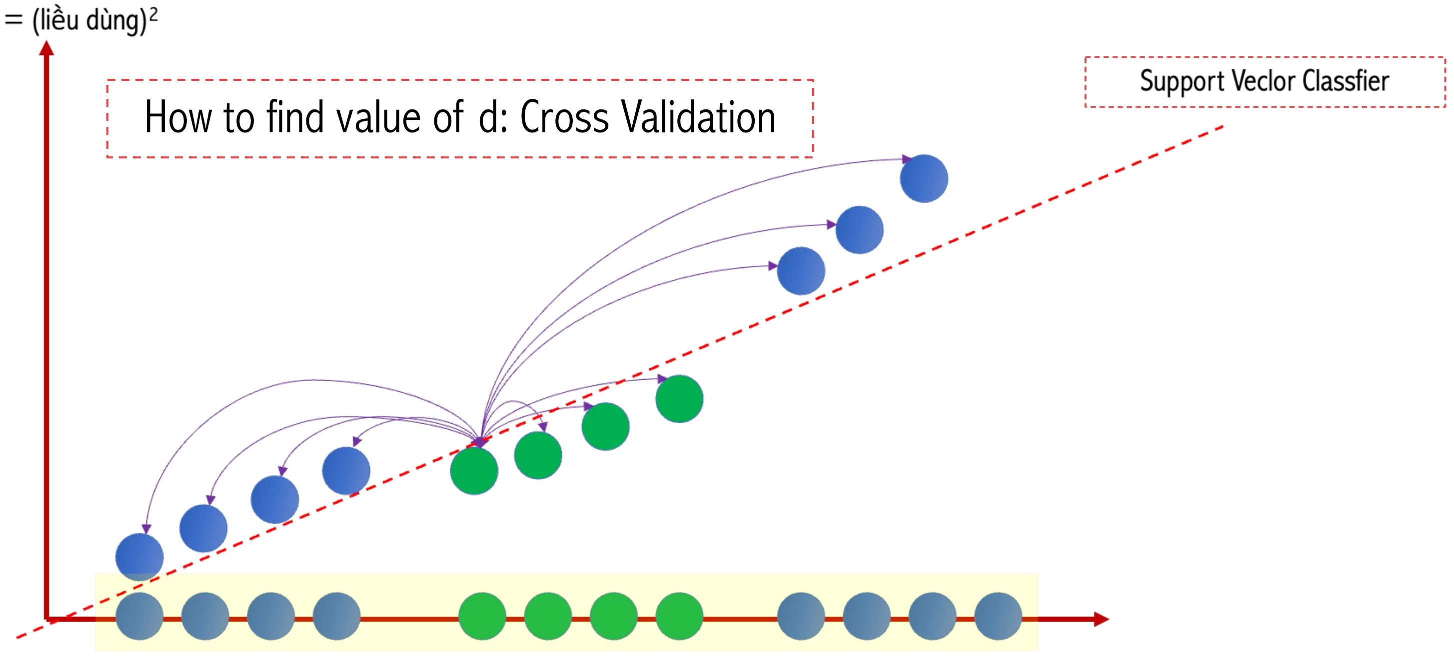
Polynomial Kernel

Polynomial Kernel: the degree of the polynomial *d*

d = 1. Compute the relationship between each pair of observations in 1-Dimensional tò find SVC



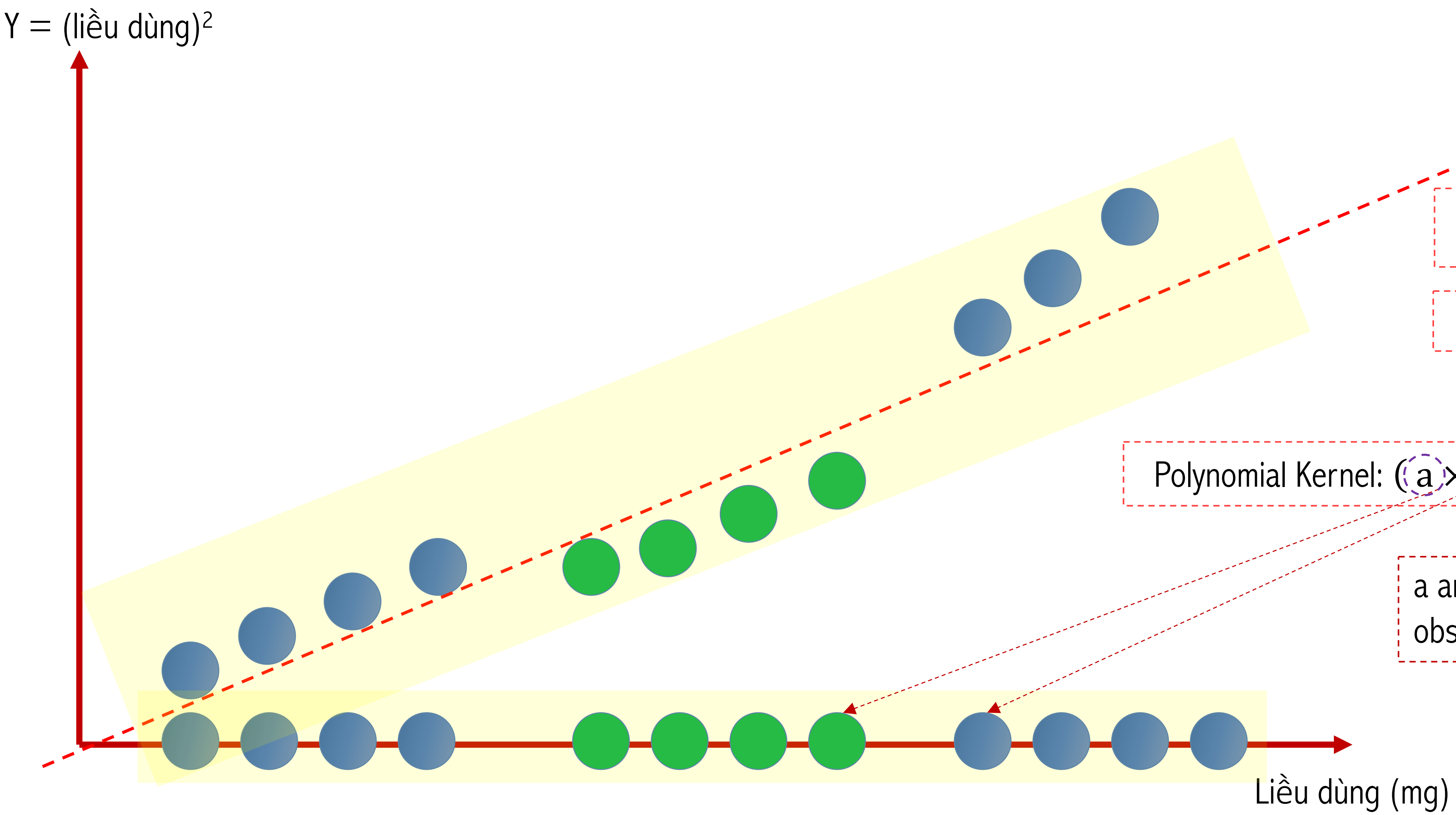
d = 2. Compute 2-Dimensional relationship between each pair of observations find SVC



d = n. Compute n-Dimensional relationship between each pair of observations. Those relationship are used to find SVC

Other kernel: Radial Basic Function Kernel (RBF)

Polynomial Kernel



Example: $r = \frac{1}{2}$, $d = 2$

Coefficient of the polynomial kernel

Degree of polyno

Polynomial Kernel: $(a \times b + r)^d$

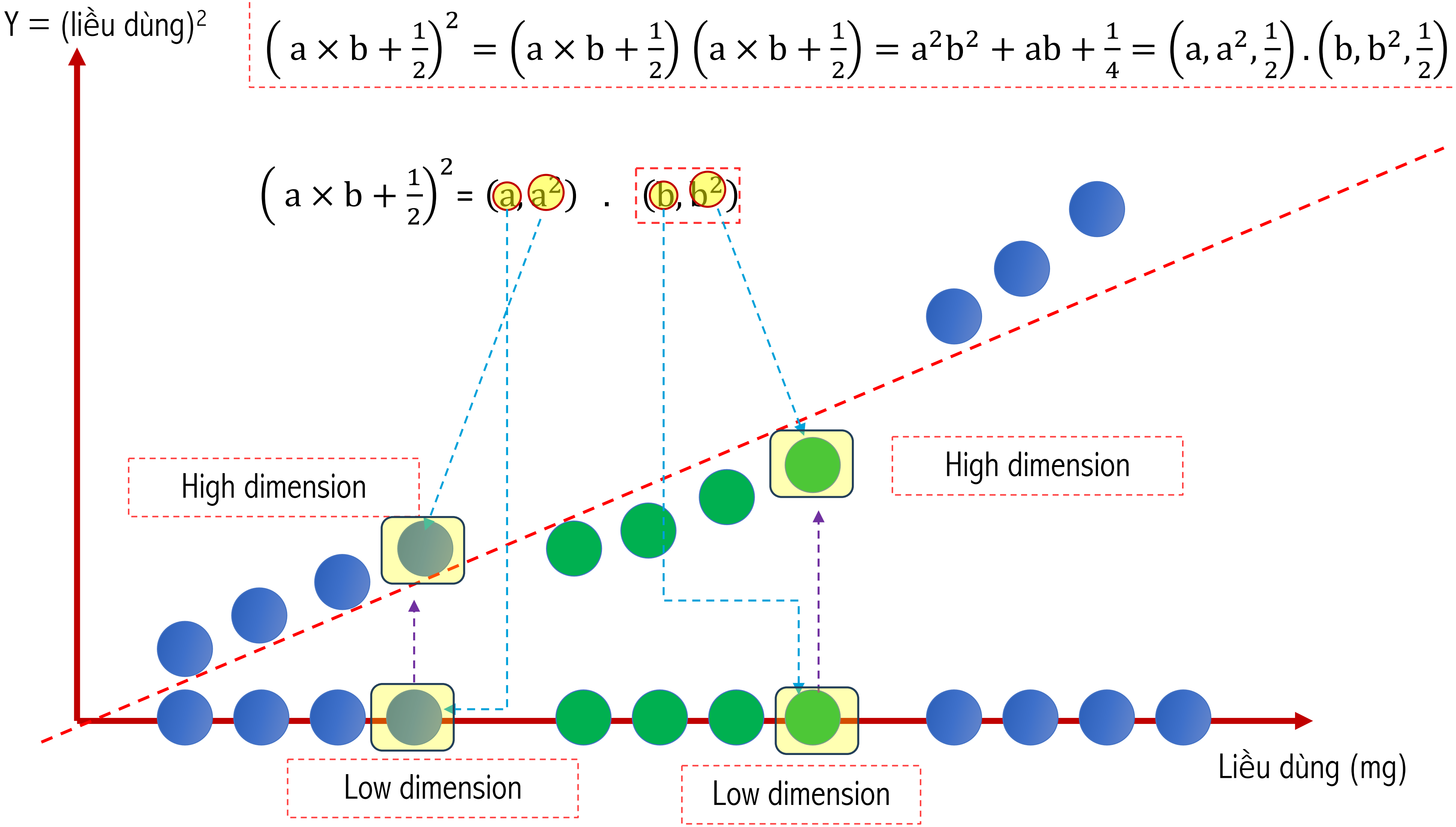
a and b refer to two different observations in the dataset

Polynomial Kernel

$$\left(a \times b + \frac{1}{2}\right)^2 = \left(a \times b + \frac{1}{2}\right)\left(a \times b + \frac{1}{2}\right) = a^2b^2 + ab + \frac{1}{4} = \left(a, a^2, \frac{1}{2}\right).\left(b, b^2, \frac{1}{2}\right)$$

$r = \frac{1}{2}, d = 2$

$$\left(a \times b + \frac{1}{2}\right)^2 = (a, a^2) \;.\; (b, b^2)$$

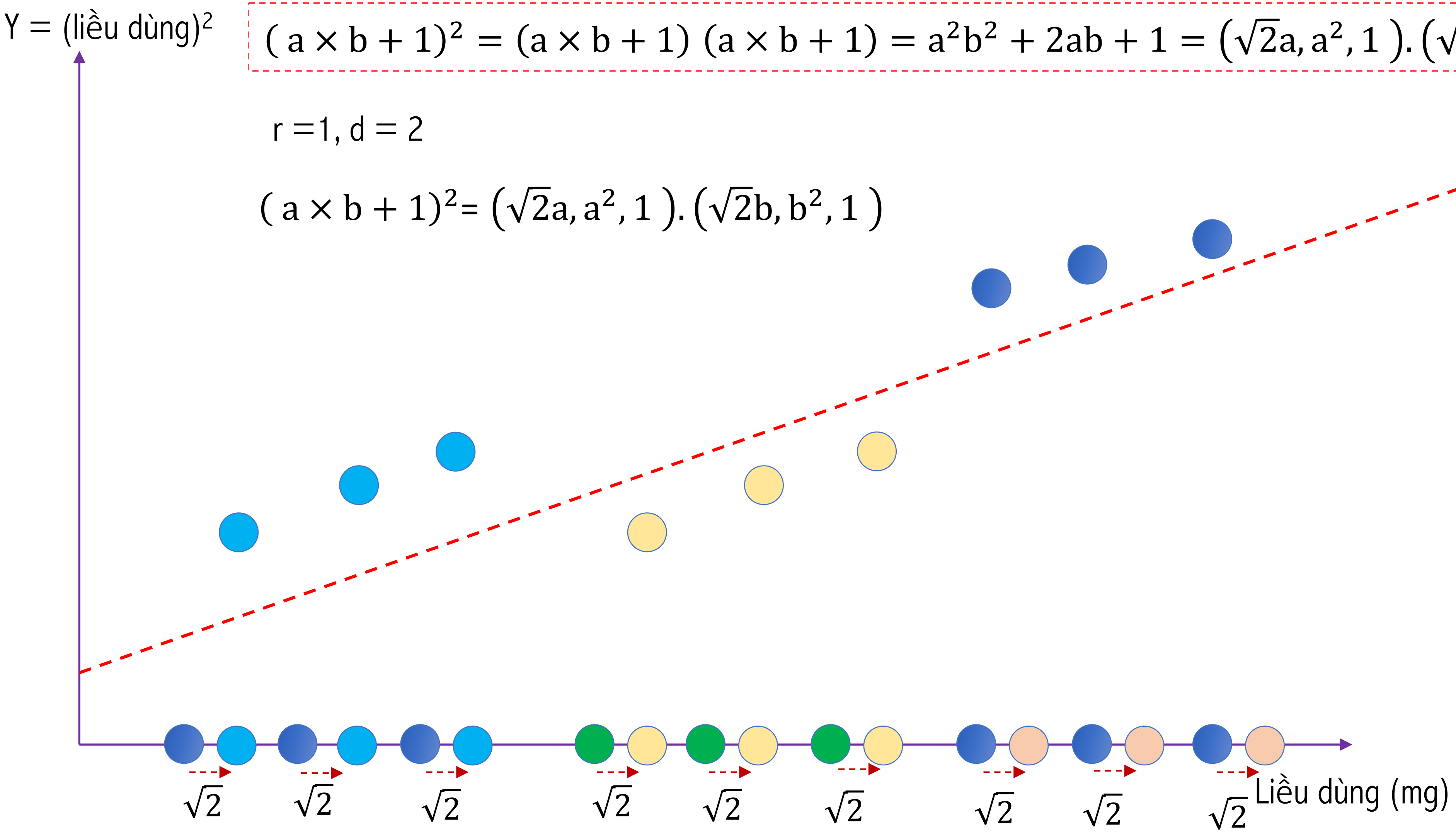


Polynomial Kernel

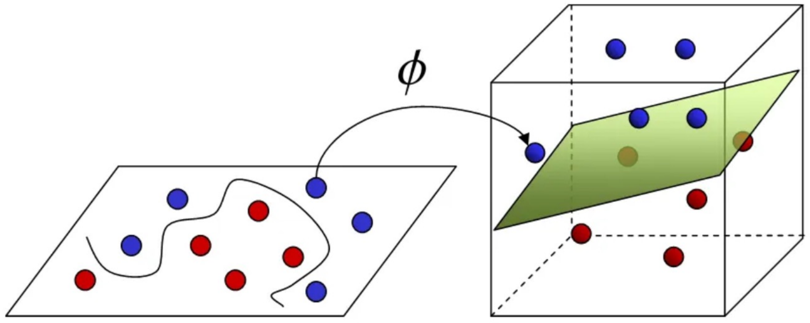
$$(a \times b + 1)^2 = (a \times b + 1)\,(a \times b + 1) = a^2b^2 + 2ab + 1 = \left(\sqrt{2}a, a^2, 1\right).\left(\sqrt{2}b, b^2, 1\right)$$

r =1, d = 2

$$(a \times b + 1)^2 = \left(\sqrt{2}a, a^2, 1\right).\left(\sqrt{2}b, b^2, 1\right)$$



Kernel Trick



Low Dimension Data

Kernel Function

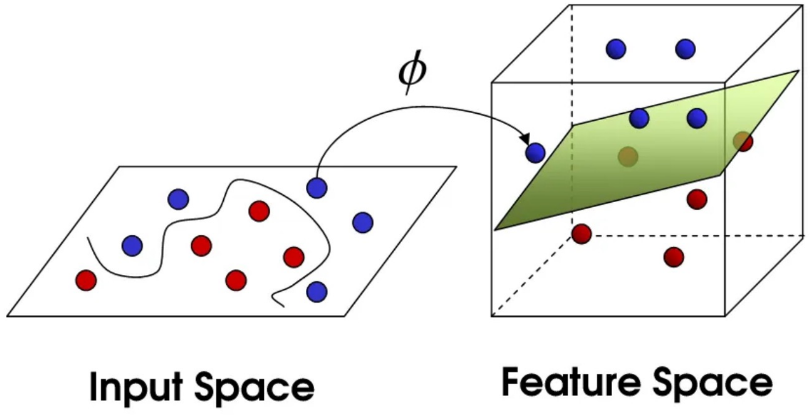
High Dimension Data

$$\phi(\mathbf{x}) = \phi\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) = \begin{pmatrix} x_1^{\,2} \\ \sqrt{2}\,x_1 x_2 \\ x_2^{\,2} \end{pmatrix}$$

A function that takes its input vector in the original space and returns the dot product of the vectors in the feature space

Test Data → SVM → Test Data

Kernel Trick



Low Dimension Data

↓

Kernel Function

↓

High Dimension Data

↓

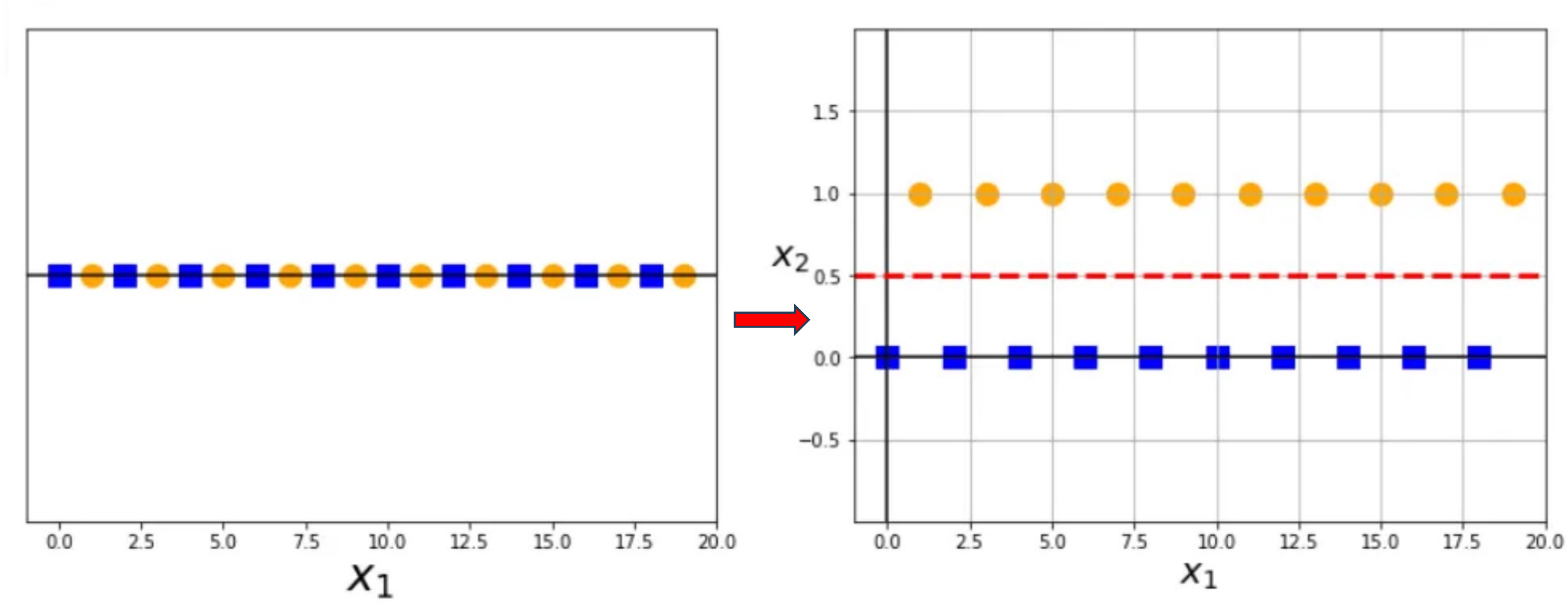
Test Data → SVM → Test Data

$$\phi(\mathbf{x}) = \phi\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) = \begin{pmatrix} x_1{}^2 \\ \sqrt{2}\,x_1 x_2 \\ x_2{}^2 \end{pmatrix}$$

A function that takes its input vector in the original space and returns the dot product of the vectors in the feature space

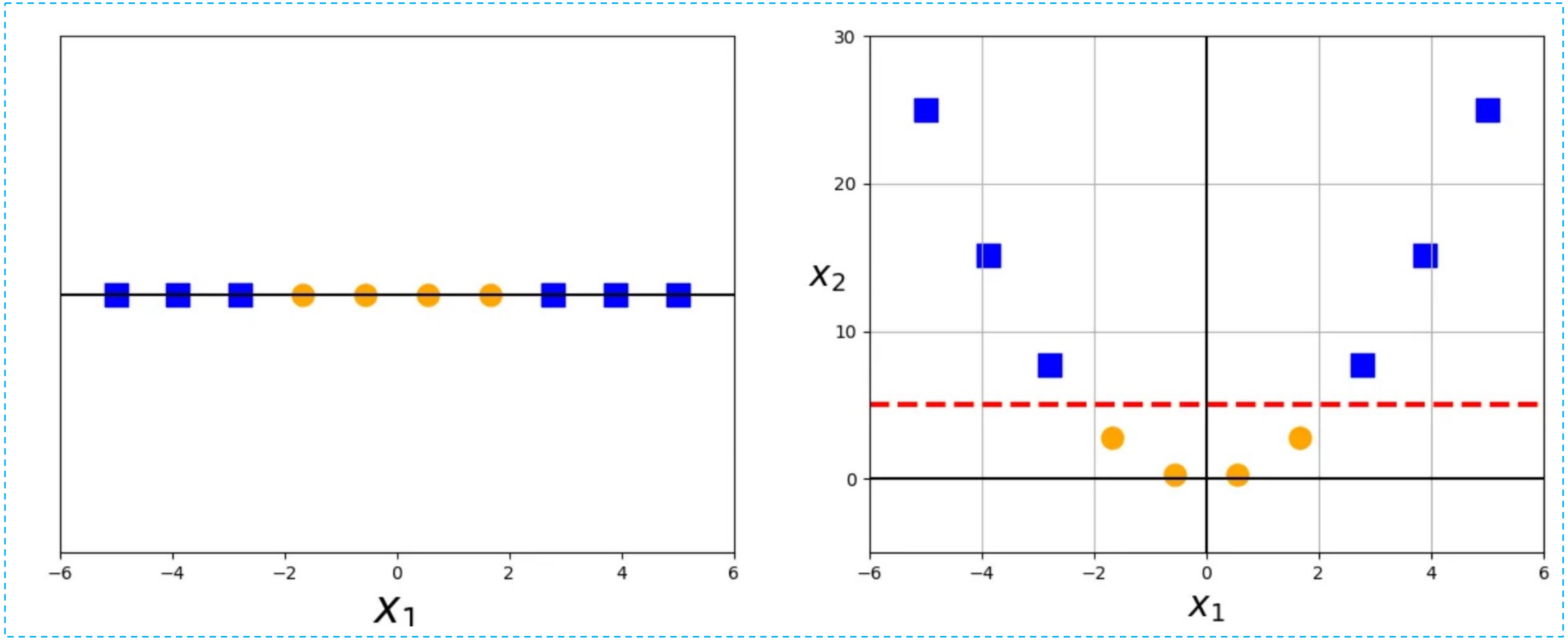
Kernel Trick

Apply the transformation $\phi(x) = x \bmod 2$



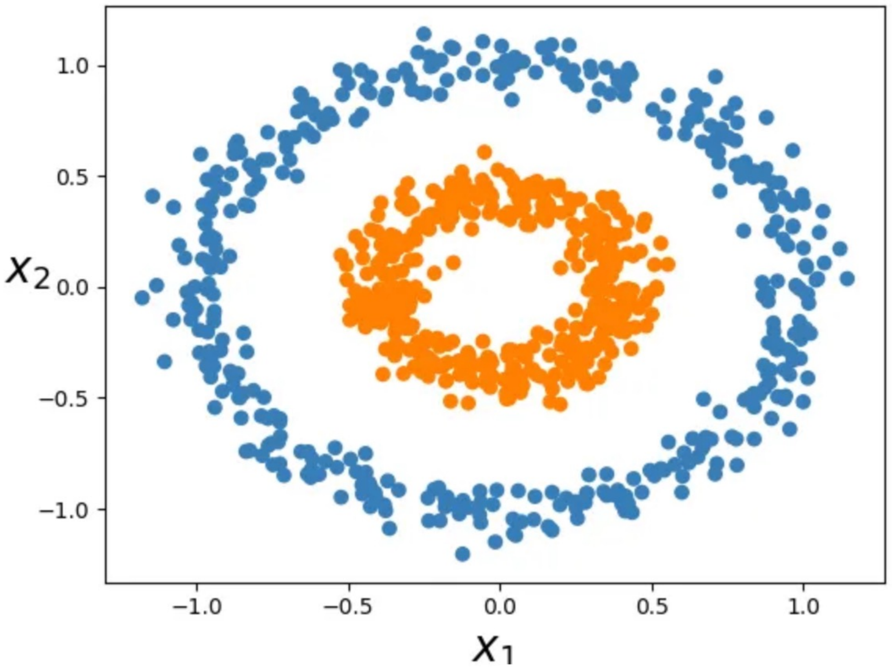
Kernel Trick

Apply the transformation $\phi(x) = x^2$



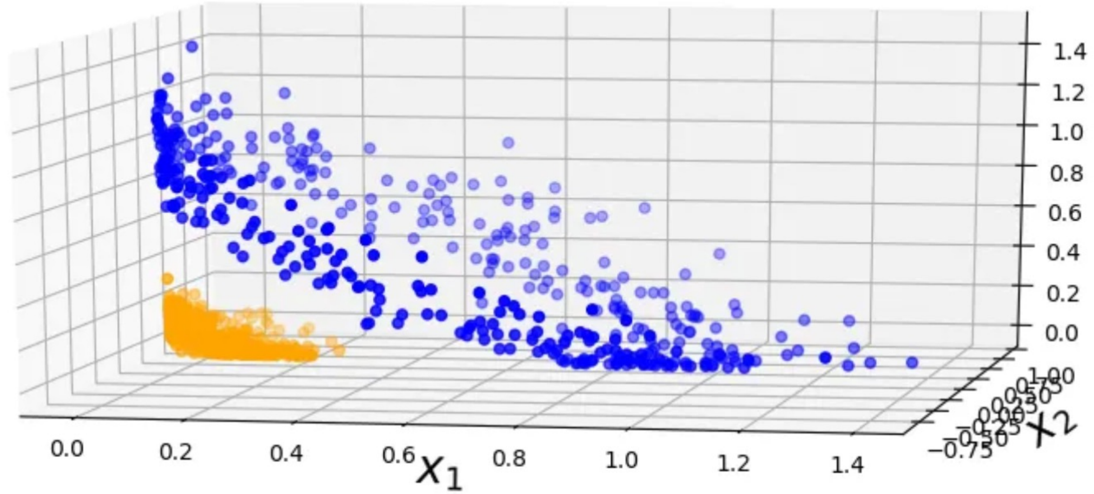
Kernel Trick

The **kernel trick** provides a solution to this problem.
It allows us to operate in the original feature space without computing the coordinates of the data in a higher dimensional space.



Second-degree polynomial mapping

$$\phi(\mathbf{x}) = \phi\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) = \begin{pmatrix} {x_1}^2 \\ \sqrt{2}x_1x_2 \\ {x_2}^2 \end{pmatrix}$$



We have seen how higher dimensional transformations can allow us to separate data in order to make classification predictions. It seems that in order to train a support vector classifier and optimize our objective function, we would have to perform operations with the higher dimensional vectors in the transformed feature space => extremely high and impractical computational costs

Kernel Trick

The “trick” is that kernel methods represent the data only through a set of pairwise similarity comparisons between the original data observations **x** (with the original coordinates in the lower dimensional space), instead of explicitly applying the transformations $\phi(\mathbf{x})$ and representing the data by these transformed coordinates in the higher dimensional feature space.

Kernel Function:

More formally, if we have data $\mathbf{x}, \mathbf{z} \in X$ and a map $\phi : X \rightarrow \Re^N$ then

$$k(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$$

is a kernel function

The ultimate benefit of the kernel trick is that the objective function we are optimizing to fit the higher dimensional decision boundary only includes the dot product of the transformed feature vectors. Therefore, we can just substitute these **dot product terms** with the kernel function, and we don’t even use $\phi(\mathbf{x})$.

$$\phi(\mathbf{a})^T \cdot \phi(\mathbf{b}) = \begin{pmatrix} a_1^2 \\ \sqrt{2}\, a_1 a_2 \\ a_2^2 \end{pmatrix}^T \cdot \begin{pmatrix} b_1^2 \\ \sqrt{2}\, b_1 b_2 \\ b_2^2 \end{pmatrix} = a_1^2 b_1^2 + 2 a_1 b_1 a_2 b_2 + a_2^2 b_2^2$$

$$\phi(\mathbf{x}) = \phi\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) = \begin{pmatrix} x_1^2 \\ \sqrt{2}\, x_1 x_2 \\ x_2^2 \end{pmatrix}$$

$$= (a_1 b_1 + a_2 b_2)^2 = \left(\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}^T \cdot \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}\right)^2 = (\mathbf{a}^T \cdot \mathbf{b})^2$$

The kernel function here is the polynomial kernel

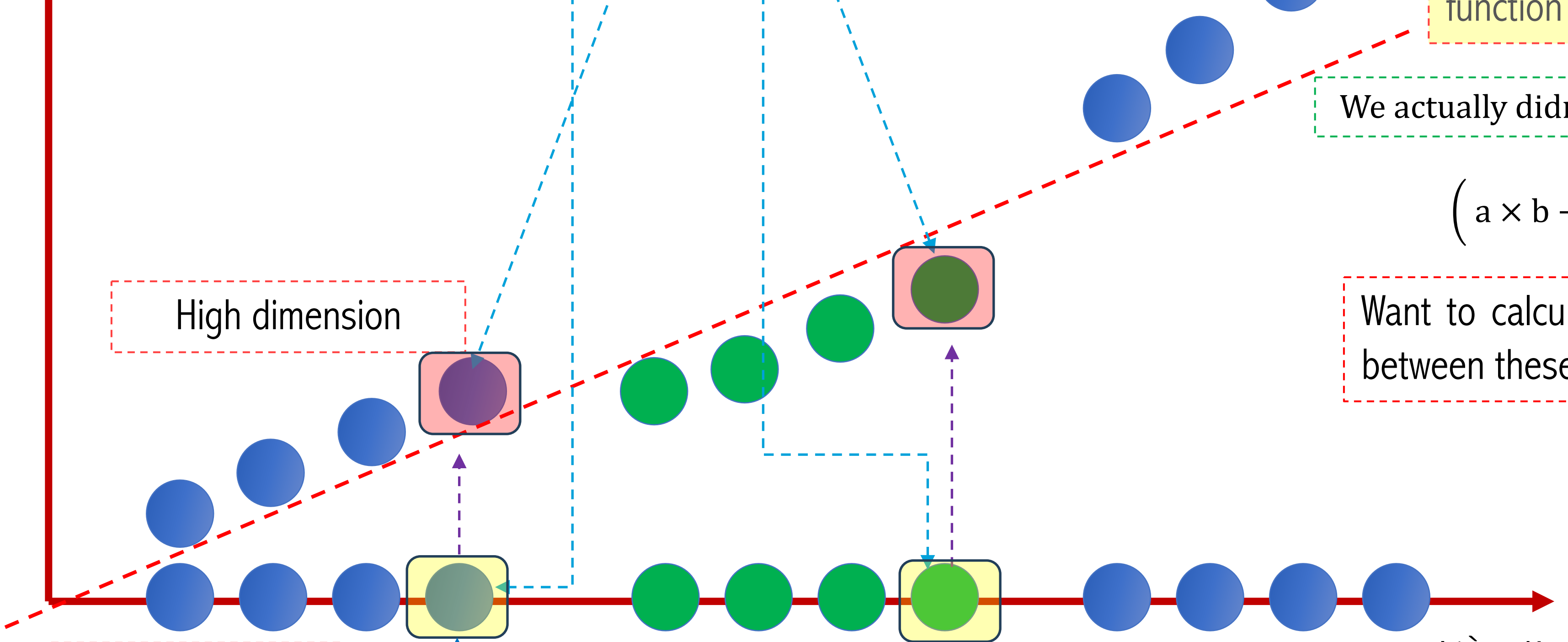
$k(a,b) = (a^T . b)^2$

Kernel Trick

$$\left(a \times b + \frac{1}{2}\right)^2 = \left(a \times b + \frac{1}{2}\right)\left(a \times b + \frac{1}{2}\right) = a^2b^2 + ab + \frac{1}{4} = \left(a, a^2, \frac{1}{2}\right).\left(b, b^2, \frac{1}{2}\right)$$

$r = \frac{1}{2}, d = 2$

$$\left(a \times b + \frac{1}{2}\right)^2 = (a, a^2) \; . \; \boxed{(b, b^2)}$$



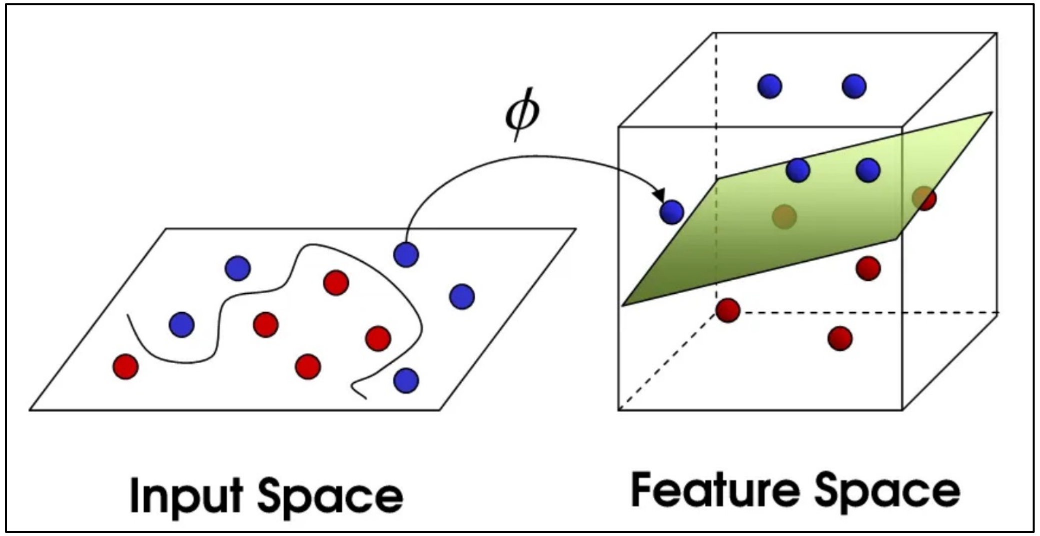
Kernel trick is to convert dot product of support vectors to the dot product of mapping function

We actually didn't transform the data to 2 − Dimensions.

$$\left(a \times b + \frac{1}{2}\right)^2 = \left(5 \times 10 + \frac{1}{2}\right)^2 = 110.25$$

Want to calculate the high dimensional relationship between these two observations (samples)

Kernel Trick



Input space $x = (x_1, x_2)$

Feature space $\phi(x) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2)$

$$
\begin{aligned}
\phi(x).\,\phi(z) &= (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2).\,(1, \sqrt{2}z_1, \sqrt{2}z_2, z_1^2, z_2^2, \sqrt{2}z_1z_2) \\
&= 1 + 2x_1z_1 + 2x_2z_2 + x_1^2z_1^2 + x_2^2z_2^2 + 2x_1x_2z_1z_2 \\
&= (1 + x_1z_1 + x_2z_2)^2 \\
&= (1 + x.\,z)^2
\end{aligned}
$$

Did you notice the magic of mathematics ? We can represent the dot product $\phi(x).\,\phi(z)$ in feature space just by using a simple formula $(1 + x.\,z)^2$ in input space.

So we do not have to perform any complex transformations or store the feature space in memory, if the dot product of feature space can be represented using dot product of input space.

Kernel Trick

If the feature space is **abstract vector space** then the Kernel is represented using $K(x_i, x_j) =< \phi(x_i), \phi(x_j) >$ and when the feature space is **vector space** then the **transpose** operator can be used. So don't be confused seeing these two different representations, they are essentially the same.

$$\phi(x_i)^T \phi(x_j) \equiv < \phi(x_i), \phi(x_j) >$$

Any Machine Learning algorithm can use Kernel Method, however it's mainly used in **SVM** and **Clustering**.

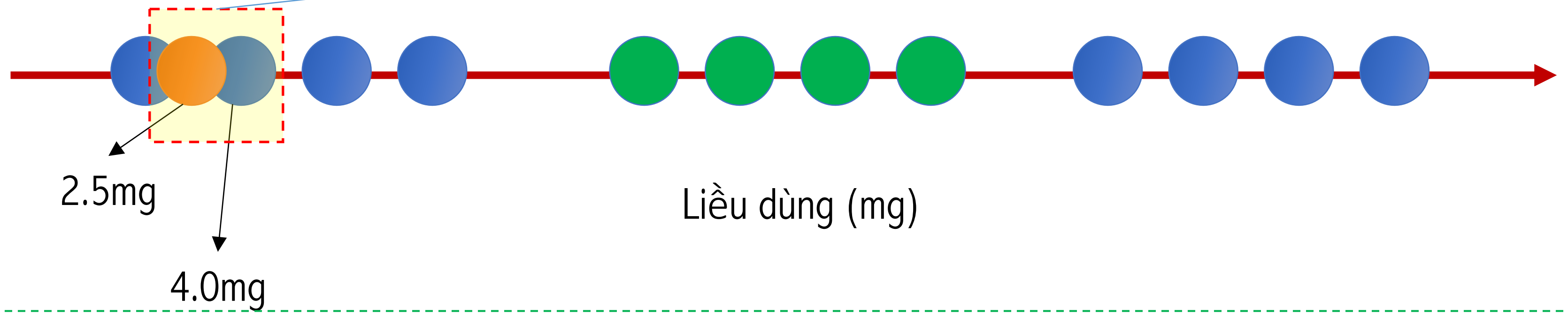
Radial Kernel

Find support vector classifer in infinite dimensions

Radial kernel: $e^{-\gamma(a-b)^2}$

$\gamma = 1$	$\gamma = 2$
$e^{-(2.5-4)^2} = 0.11$	$e^{-(2.5-4)^2} = 0.01$

γ scales the amount of influence two points have each other



The nearest neighbors have a lot of influence on how we classify the new observation.
Radiel kernel determines how much influence each observation in the Training Dataset has on classifying new observation

Outline

- Maximal Margin Classifier
- Support Vector Classifier
- Support Vector Machine
- Polynomial Kernel
- Radial Basic Function Kernel (RBF)
- Example

Radial Kernel

Find support vector classifer in infinite dimensions

Radial kernel: $e^{-\gamma(a-b)^2}$

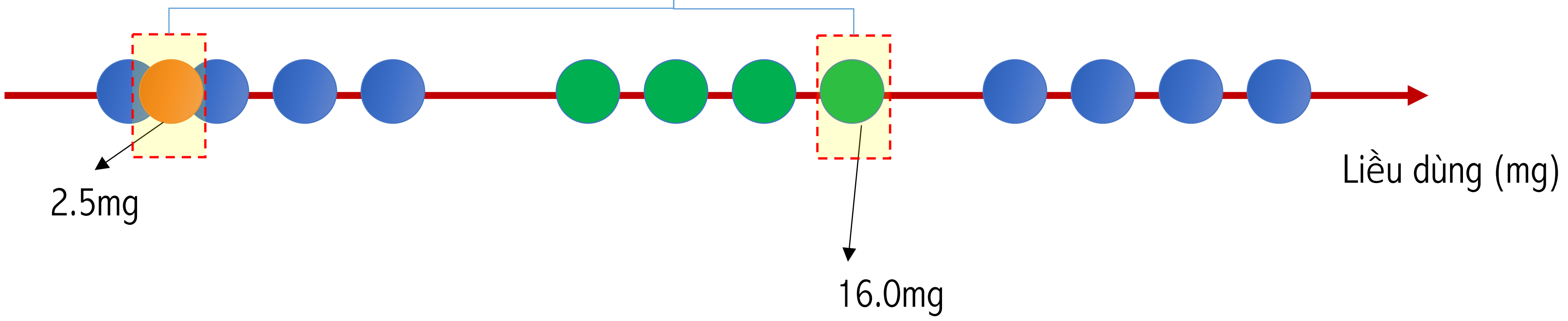
$\gamma = 1$

$e^{-(2.5-16)^2} \approx 0$

$\gamma = 2$

$e^{-(2.5-16)^2} \approx 0$

γ scales the amount of influence two points have each other



The further two observation are from each other, the less influence they have on each other

Radial Kernel: Intuition

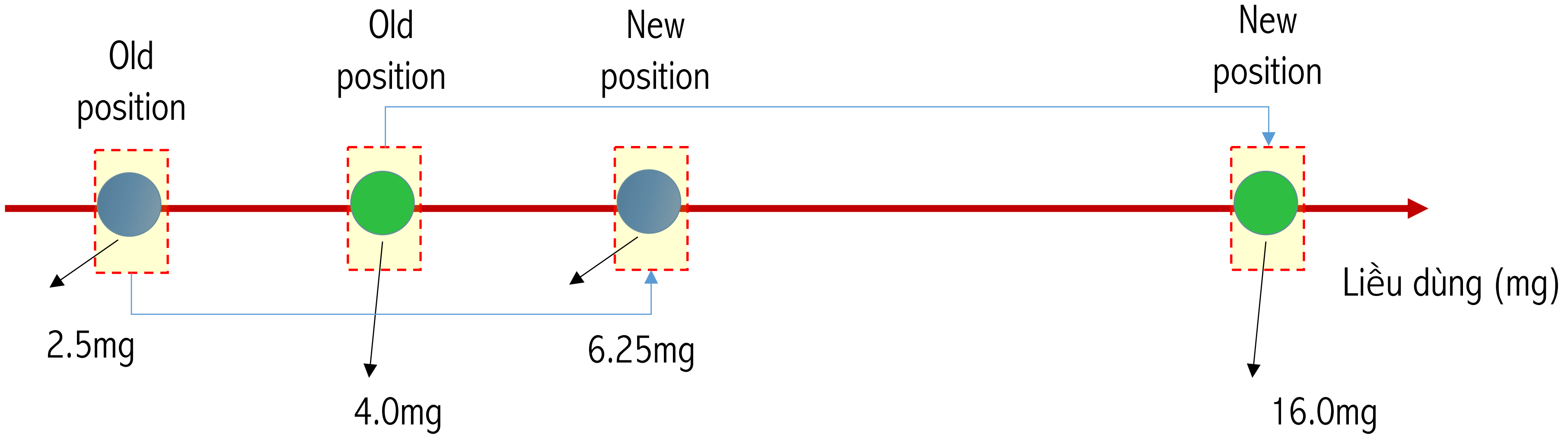
Polynomial Kernel: $(a \times b + r)^d$ with $r = 0$

$$(a \times b + r)^d = (a \times b)^d = (a^d)\,(b^d)$$

$$d = 2 \Rightarrow (a \times b + r)^2 = (a^2)\,(b^2)$$

$$d = 2 \Rightarrow (a \times b + r)^2 = (2.5^2)\,(4^2)$$

This dot product only has one coordinate. The new coordinate is square of the original measurement on the original axis



The original data are shift with r = 0 and d = 2

Radial Kernel: Intuition

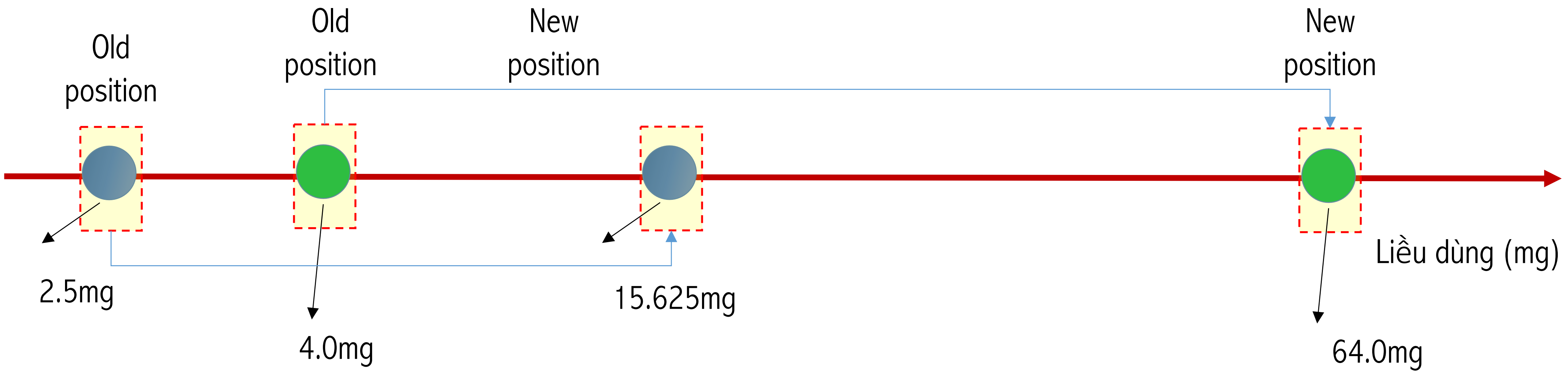
Polynomial Kernel: $(a \times b + r)^d$ with $r = 0$

$$(a \times b + r)^d = (a \times b)^d = (a^d)\,(b^d)$$

$$d = 3 \Rightarrow (a \times b + r)^3 = (a^3)\,(b^3)$$

$$d = 3 \Rightarrow (a \times b + r)^3 = (2.5^3)\,(4^3)$$

This dot product only has one coordinate. The new coordinate is square of the original measurement on the original axis



The original data are shift further with r = 0 and d = 3

Radial Kernel: Intuition

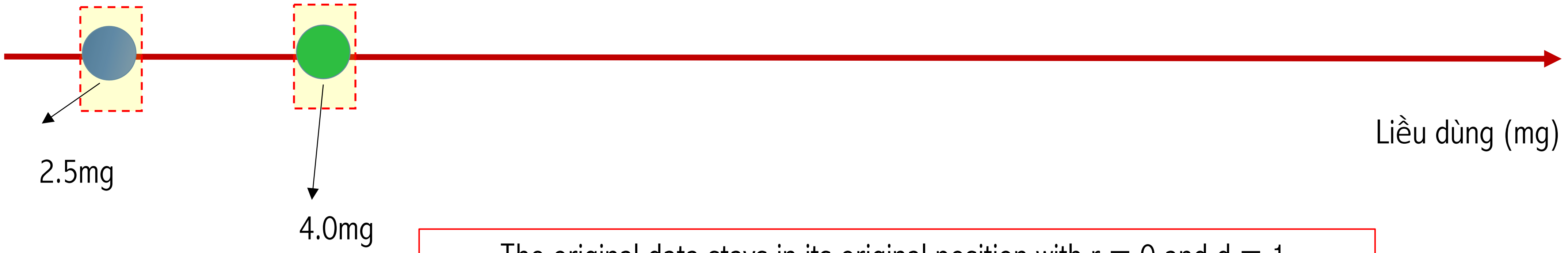
Polynomial Kernel: $(a \times b + r)^d$ with $r = 0$

$$(a \times b + r)^d = (a \times b)^d = (a^d)\,(b^d)$$

$$d = 3 \Rightarrow (a \times b + r)^3 = (a^1)\,(b^1)$$

$$d = 3 \Rightarrow (a \times b + r)^3 = (2.5^1)\,(4^1)$$

This dot product only has one coordinate. The new coordinate is square of the original measurement on the original axis



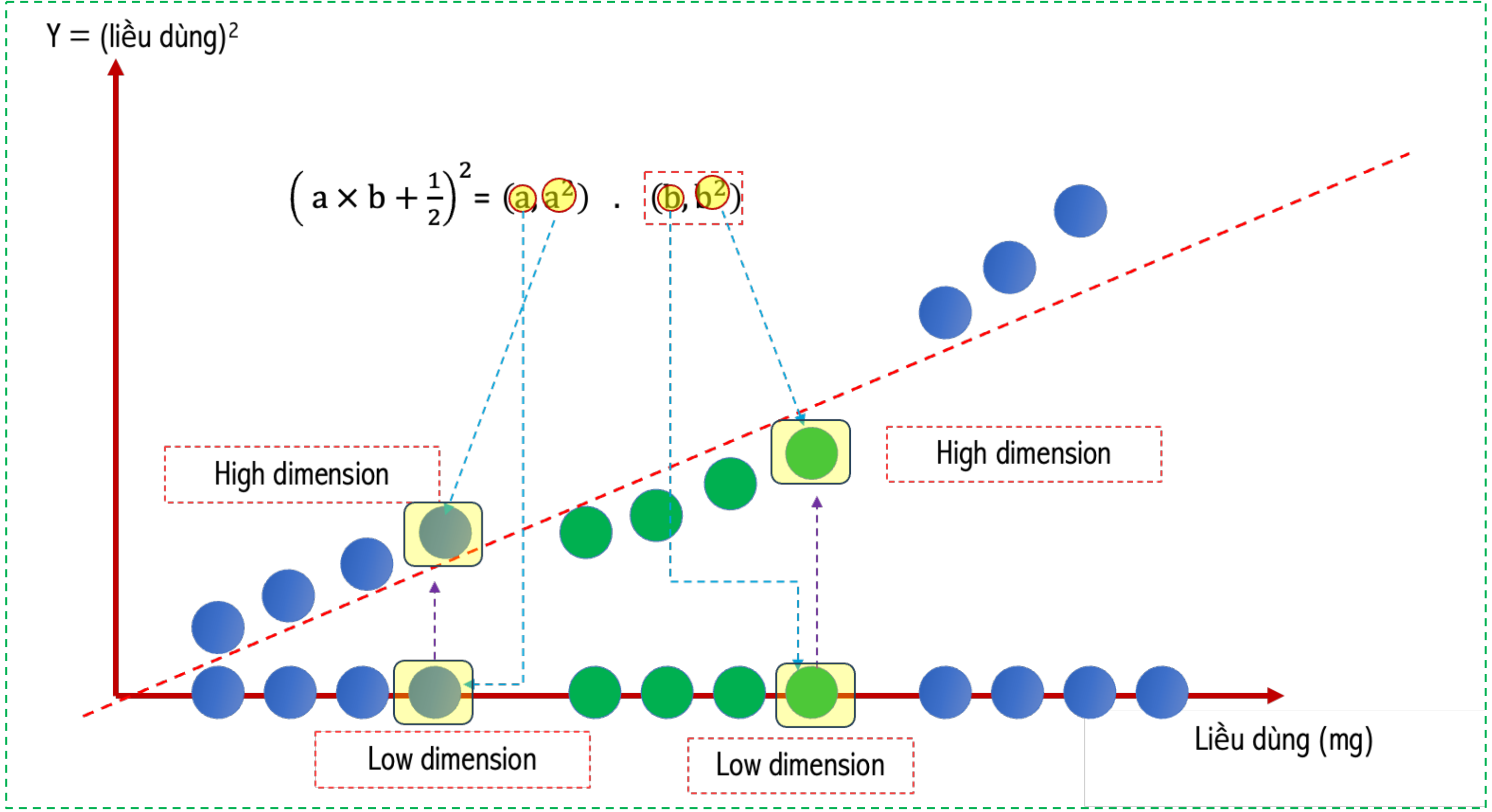
The original data stays in its original position with r = 0 and d = 1.
The data stays on the same 1-dimensional line regardless the value of d

Radial Kernel: Intuition

Polynomial Kernel: $(a \times b)^1 = (a^1)(b^1)$ with $r = 0$ and $d = 1$

Polynomial Kernel: $(a \times b)^2 = (a^2)(b^2)$ with $r = 0$ and $d = 2$

$$(a^1)(b^1) + (a^2)(b^2) = (a, a^2) \cdot (b, b^2)$$



We do not acctually do the transformation, we just solve for Dot Product the get high dimensional relationship!

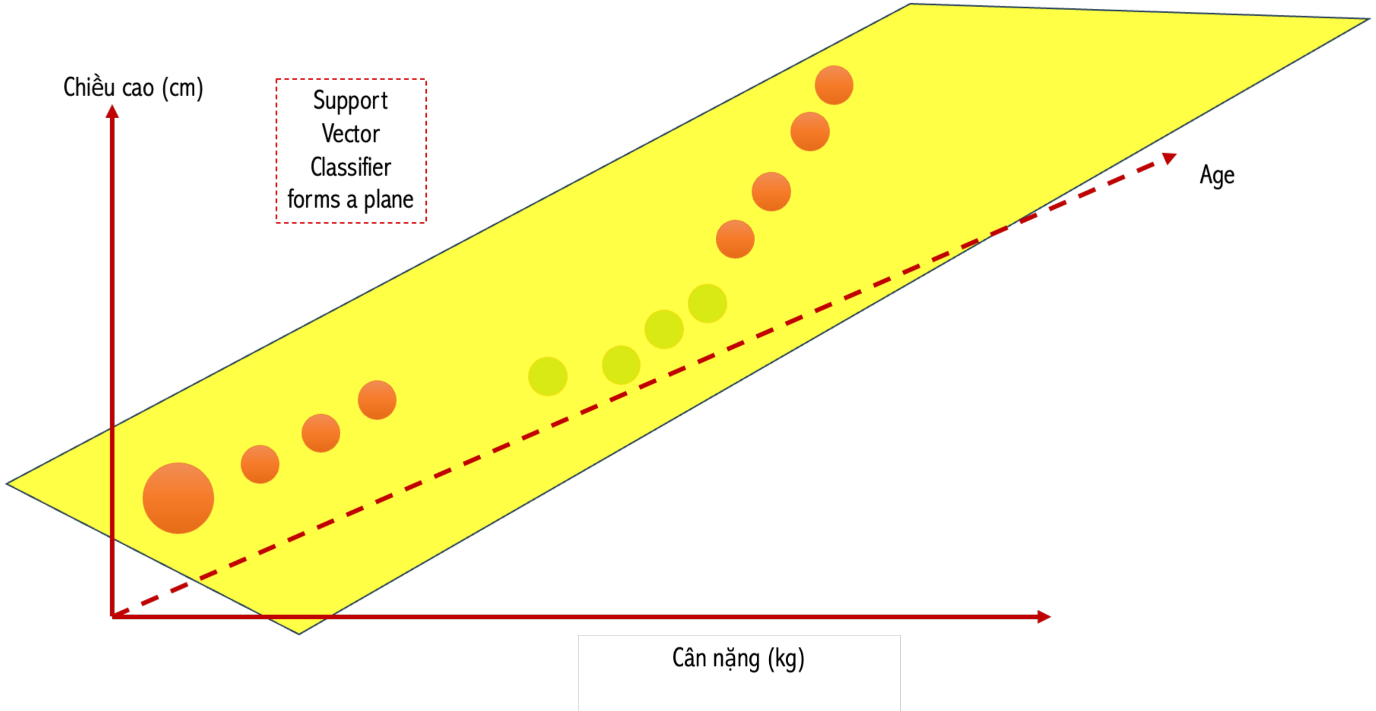
Radial Kernel: Intuition

Polynomial Kernel: $(a \times b)^1 = (a^1)\,(b^1)$ with $r = 0$ and $d = 1$

Polynomial Kernel: $(a \times b)^2 = (a^2)\,(b^2)$ with $r = 0$ and $d = 2$

Polynomial Kernel: $(a \times b)^3 = (a^3)\,(b^3)$ with $r = 0$ and $d = 3$

$$(a^1)\,(b^1) + (a^2)\,(b^2) + (a^3)\,(b^3) = (a, a^2, a^3)\ .\ (b, b^2, b^3)$$



We do not acctually do the transformation, we just solve for Dot Product the get high dimensional relationship!

Radial Kernel: Intuition

Polynomial Kernel: $(a \times b)^1 = (a^1)\,(b^1)$ with $r = 0$ and $d = 1$

Polynomial Kernel: $(a \times b)^2 = (a^2)\,(b^2)$ with $r = 0$ and $d = 2$

Polynomial Kernel: $(a \times b)^3 = (a^3)\,(b^3)$ with $r = 0$ and $d = 3$

Polynomial Kernel: $(a \times b)^{\cdots} = (a^{\cdots})\,(b^{\cdots})$ with $r = 0$ and $d = \cdots$

Polynomial Kernel: $(a \times b)^\infty = (a^\infty)\,(b^\infty)$ with $r = 0$ and $d = \infty$

$$(a^1)\,(b^1) + (a^2)\,(b^2) + (a^3)\,(b^3) + \ldots + (a^\infty)\,(b^\infty) = (a, a^2, a^3, \ldots, a^\infty,) \,.\, (b, b^2, b^3, \ldots, b^\infty)$$

Radial kernel: $e^{-\gamma(a-b)^2} = e^{-\frac{1}{2}(a^2-2ab+b^2)} = e^{-\frac{1}{2}(a^2+b^2)}\; e^{ab}$

Taylor Series:

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots + \frac{f^\infty(a)}{\infty!}(x-a)^\infty$$

$$e^x = e^a + \frac{e^a}{1!}(x-a) + \frac{e^a}{2!}(x-a)^2 + \frac{e^a}{3!}(x-a)^3 + \ldots + \frac{e^a}{\infty!}(x-a)^\infty$$

$$e^x = e^0 + \frac{e^0}{1!}(x-a) + \frac{e^0}{2!}(x-0)^2 + \frac{e^0}{3!}(x-0)^3 + \ldots + \frac{e^0}{\infty!}(x-0)^\infty$$

Radial Kernel: Intuition

Polynomial kernel

$$(a^0)(b^0) + (a^1)(b^1) + (a^2)(b^2) + (a^3)(b^3) + \ldots + (a^\infty)(b^\infty) = (a, a^2, a^3, \ldots, a^\infty,) \,.\, (b, b^2, b^3, \ldots, b^\infty)$$

Radial kernel

Radial kernel: $e^{-\gamma(a-b)^2} = e^{-\frac{1}{2}(a^2-2ab+b^2)} = e^{-\frac{1}{2}(a^2+b^2)}\, e^{ab}$

$$e^{ab} = e^0 + \frac{e^0}{1!}ab + \frac{e^0}{2!}(ab)^2 + \frac{e^0}{3!}(ab)^3 + \ldots + \frac{e^0}{\infty!}(ab)^\infty$$

$$e^{ab} = \boxed{1} + \frac{1}{1!}\boxed{ab} + \frac{1}{2!}\boxed{(ab)^2} + \frac{1}{3!}\boxed{(ab)^3} + \ldots + \frac{1}{\infty!}\boxed{(ab)^\infty}$$

$$\boxed{(a^0)(b^0)} + \boxed{(a^1)(b^1)} + \boxed{(a^2)(b^2)} + \boxed{(a^3)(b^3)} + \ldots + \boxed{(a^\infty)(b^\infty)} = (1, a, a^2, a^3, \ldots, a^\infty,) \,.\, (1, b, b^2, b^3, \ldots, b^\infty)$$

$$e^{ab} = \left(1, \sqrt{\frac{1}{1!}}a, \sqrt{\frac{1}{2!}}a^2, \sqrt{\frac{1}{3!}}a^3, \ldots, \sqrt{\frac{1}{\infty!}}a^\infty,\right) \cdot \left(1, \sqrt{\frac{1}{1!}}b, \sqrt{\frac{1}{2!}}b^2, \sqrt{\frac{1}{3!}}b^3, \ldots, \sqrt{\frac{1}{\infty!}}b^\infty\right)$$

Radial Kernel: Intuition

Radial kernel: $e^{-\gamma(a-b)^2} = e^{-\frac{1}{2}(a^2-2ab+b^2)} = e^{-\frac{1}{2}(a^2+b^2)}\ e^{ab}$

$$e^{ab} = \left(1, \sqrt{\frac{1}{1!}}a, \sqrt{\frac{1}{2!}}a^2, \sqrt{\frac{1}{3!}}a^3, \dots, \sqrt{\frac{1}{\infty!}}a^\infty, \right) \cdot \left(1, \sqrt{\frac{1}{1!}}b, \sqrt{\frac{1}{2!}}b^2, \sqrt{\frac{1}{3!}}b^3, \dots, \sqrt{\frac{1}{\infty!}}b^\infty\right)$$

$$e^{-\frac{1}{2}(a^2+b^2)}\ e^{ab} = e^{-\frac{1}{2}(a^2+b^2)}\left(1, \sqrt{\frac{1}{1!}}a, \sqrt{\frac{1}{2!}}a^2, \sqrt{\frac{1}{3!}}a^3, \dots, \sqrt{\frac{1}{\infty!}}a^\infty, \right) \cdot \left(1, \sqrt{\frac{1}{1!}}b, \sqrt{\frac{1}{2!}}b^2, \sqrt{\frac{1}{3!}}b^3, \dots, \sqrt{\frac{1}{\infty!}}b^\infty\right)$$

$$e^{-\frac{1}{2}(a^2+b^2)}\ e^{ab} = \left(\delta, \delta\sqrt{\frac{1}{1!}}a, \delta\sqrt{\frac{1}{2!}}a^2, \delta\sqrt{\frac{1}{3!}}a^3, \dots, \delta\sqrt{\frac{1}{\infty!}}a^\infty, \right) \cdot \left(\delta, \delta\sqrt{\frac{1}{1!}}b, \delta\sqrt{\frac{1}{2!}}b^2, \delta\sqrt{\frac{1}{3!}}b^3, \dots, \delta\sqrt{\frac{1}{\infty!}}b^\infty\right)$$

$$\delta = \sqrt{e^{-\frac{1}{2}(a^2+b^2)}}$$

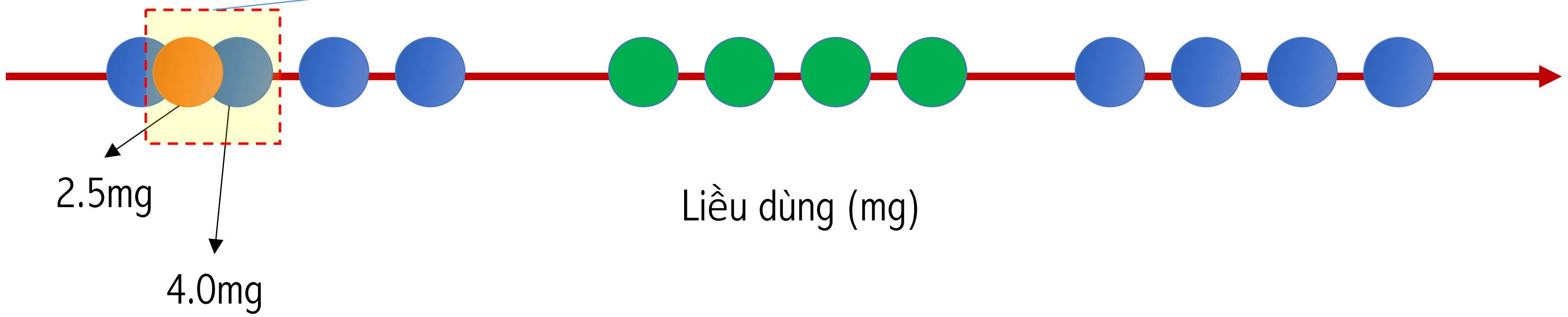
Radial kernel is equal to a dot product that has coordinates for infinite number of dimensions

Radial Kernel: Example

$$The\ relationship\ between\ two\ points\ in\ infinite - dimensions$$

$\gamma = 1$	$\gamma = 2$
$\mathrm{e}^{-(2.5-4)^2} = 0.11$	$\mathrm{e}^{-(2.5-4)^2} = 0.01$

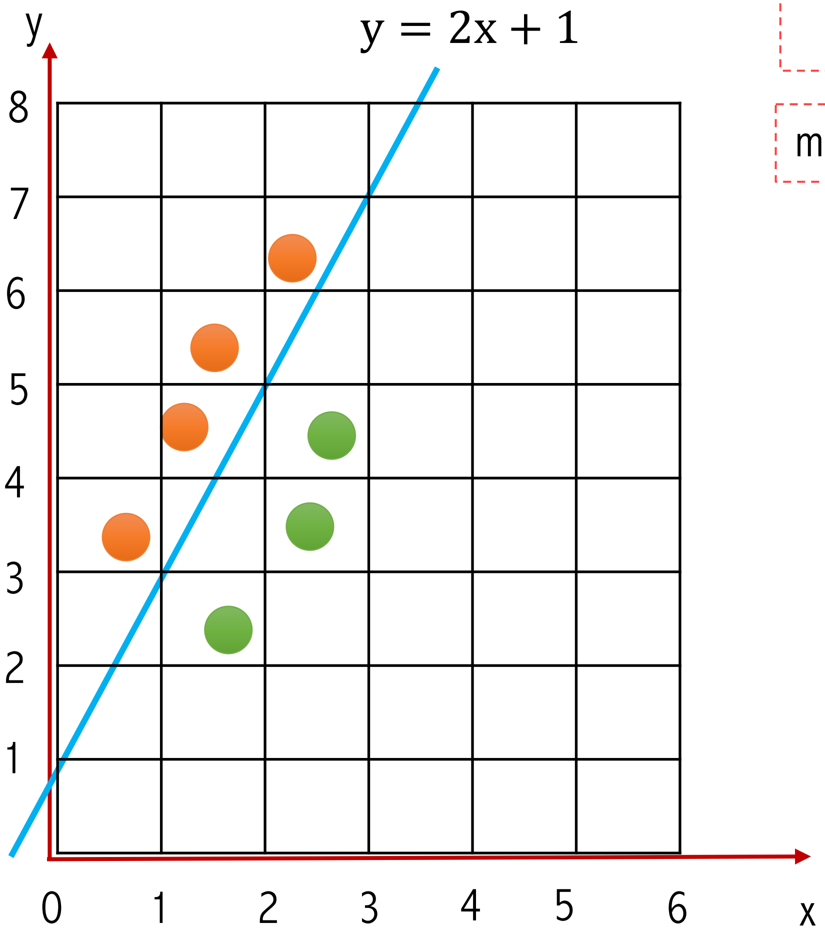
γ scales the amount of influence two points have each other



Outline

- Maximal Margin Classifier
- Support Vector Classifier
- Support Vector Machine
- Polynomial Kernel
- Radial Basic Function Kernel (RBF)
- Example

Line Equation Review



$$y = mx + b$$

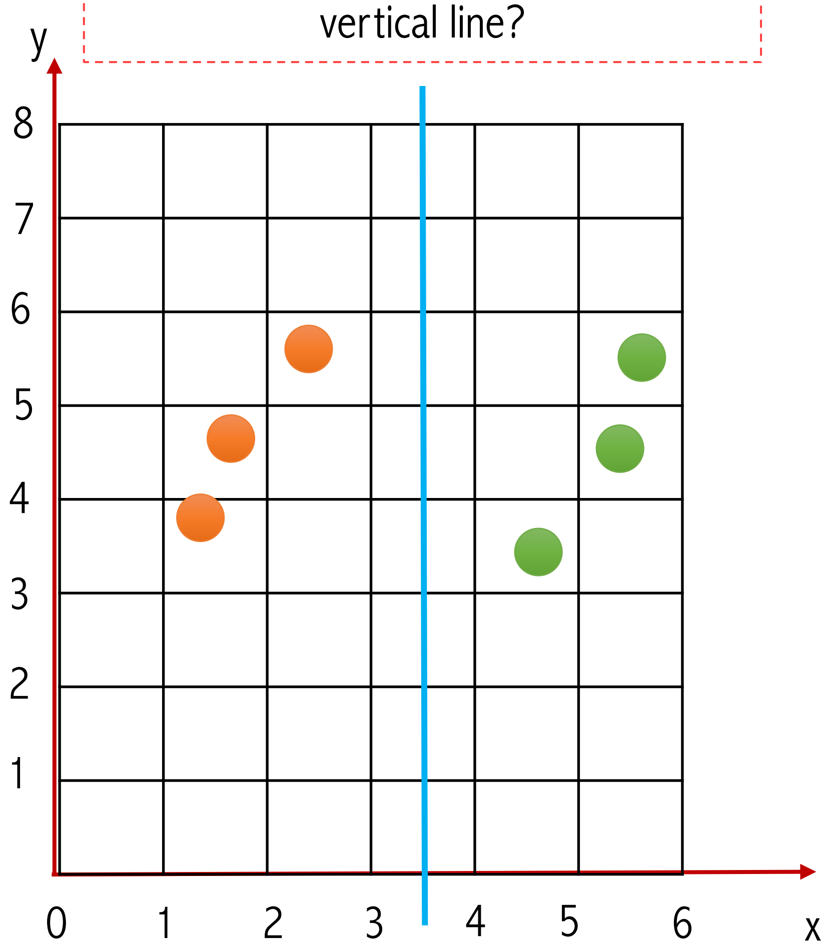
m : slope, b: intercept

General form of the equation of the straight line

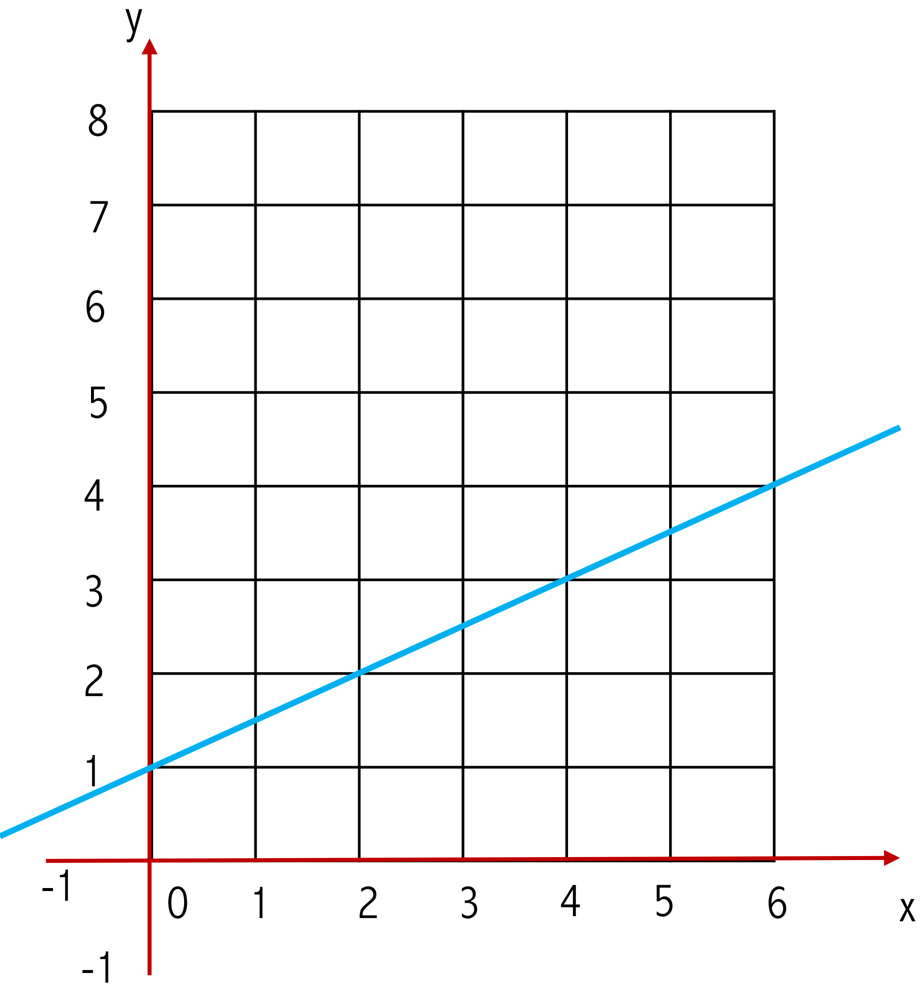
$$Ax + By + C = 0$$

$$y = -\frac{A}{B}x - \frac{C}{B}$$

What is the equation of the following vertical line?



Line Equation Review



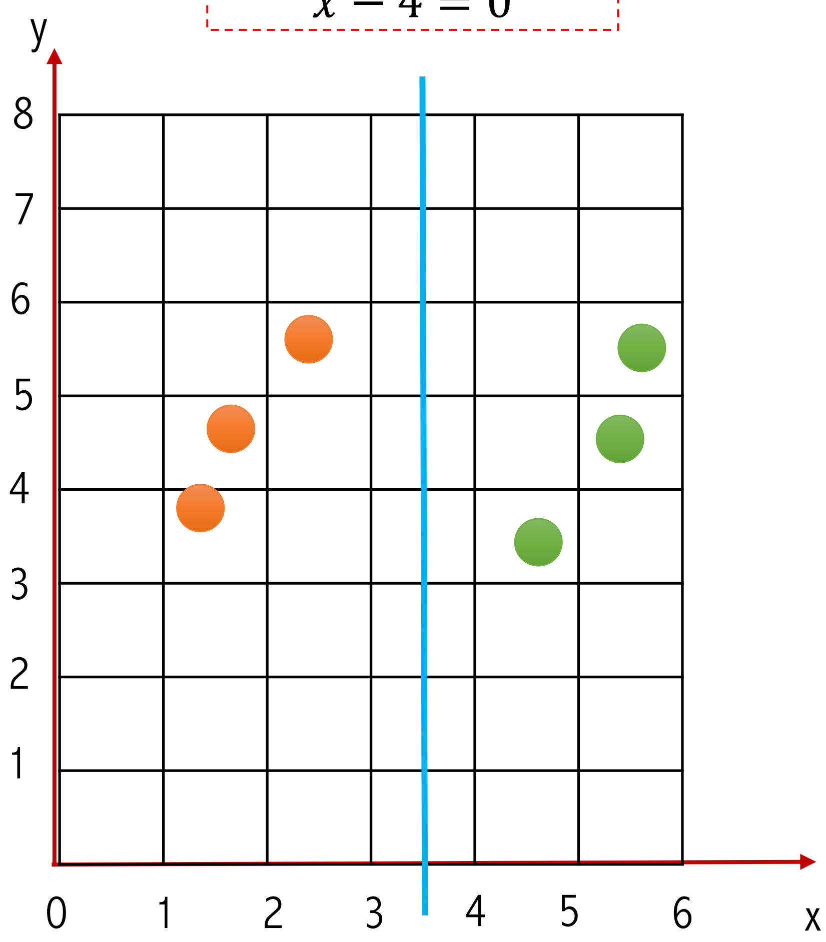
$y = 0.5x + 1$

$-0.5x + y - 1 = 0$

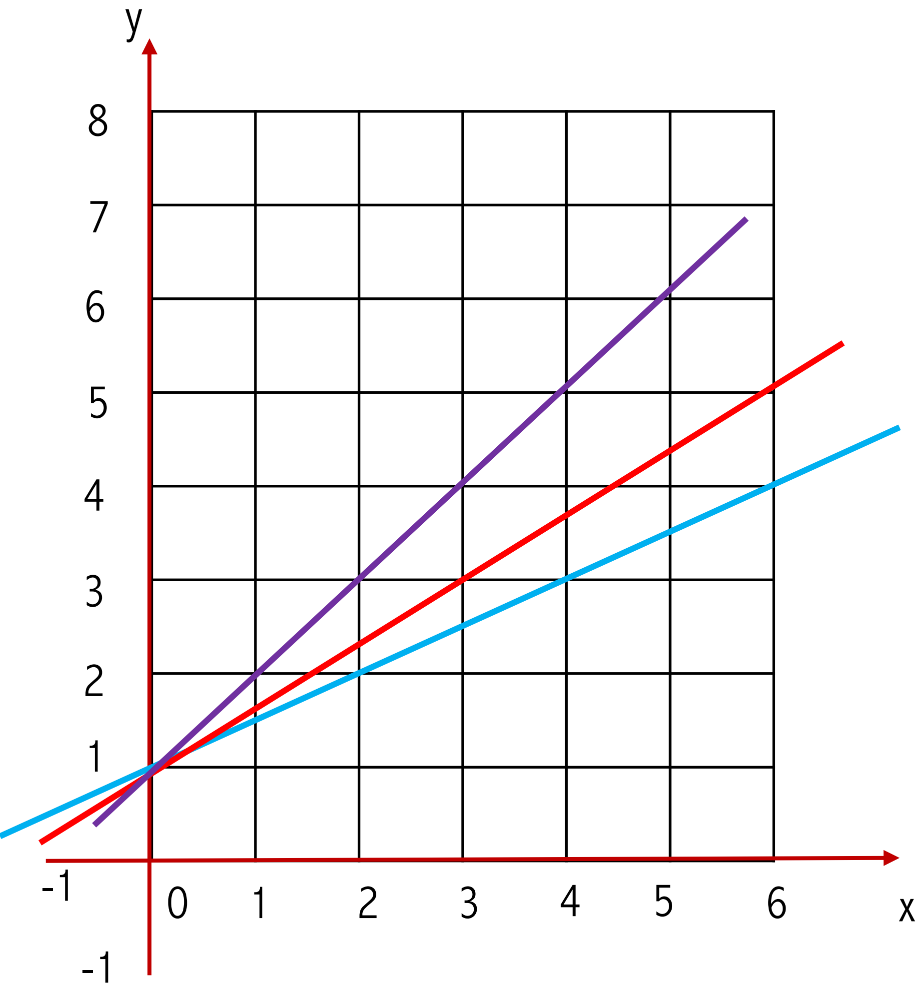
$-2x + 4y - 4 = 0$

$2x - 4y + 4 = 0$

$x - 4 = 0$



Line Equation Review: A, B, C Changes



$$Ax + By + C = 0$$

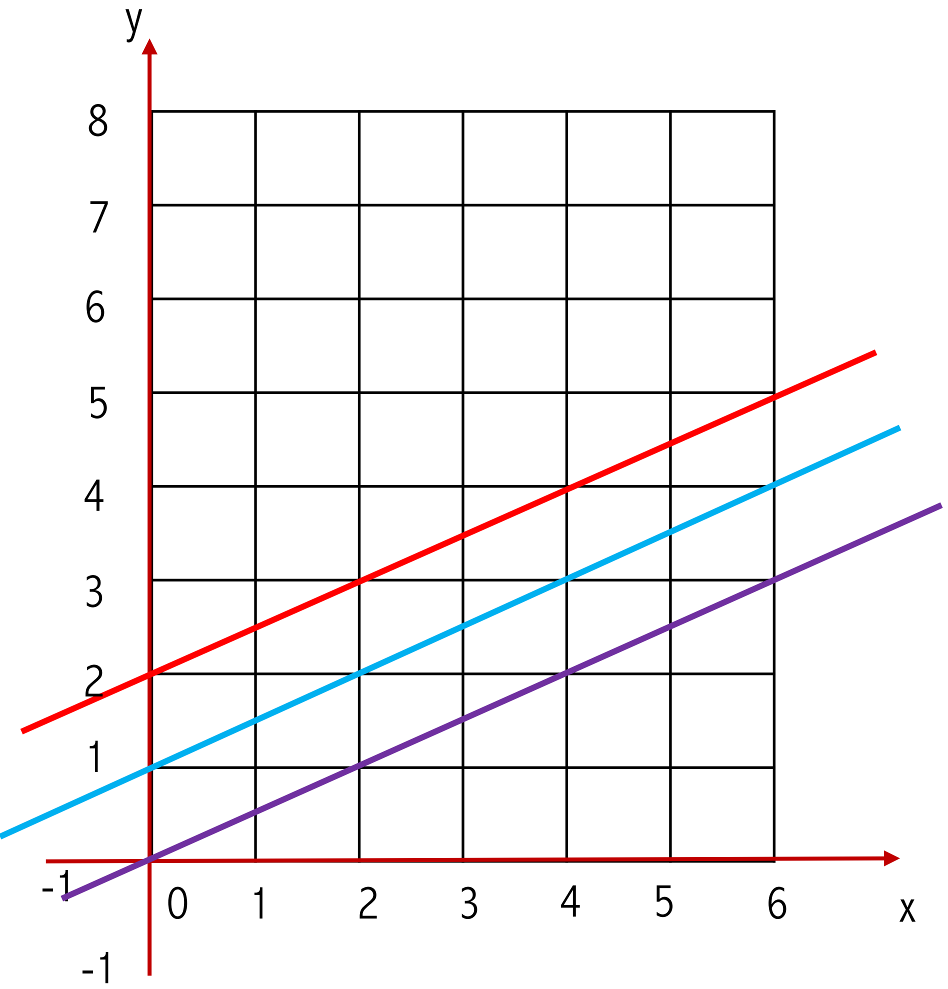
$$-2x + 4y - 4 = 0$$

$$-x + 4y - 4 = 0$$

$$-4x + 4y - 4 = 0$$

When we change A, the Line is rotating around 1.

Line Equation Review: A, B, C Changes



$$y = -\frac{A}{B}x - \frac{C}{B}$$

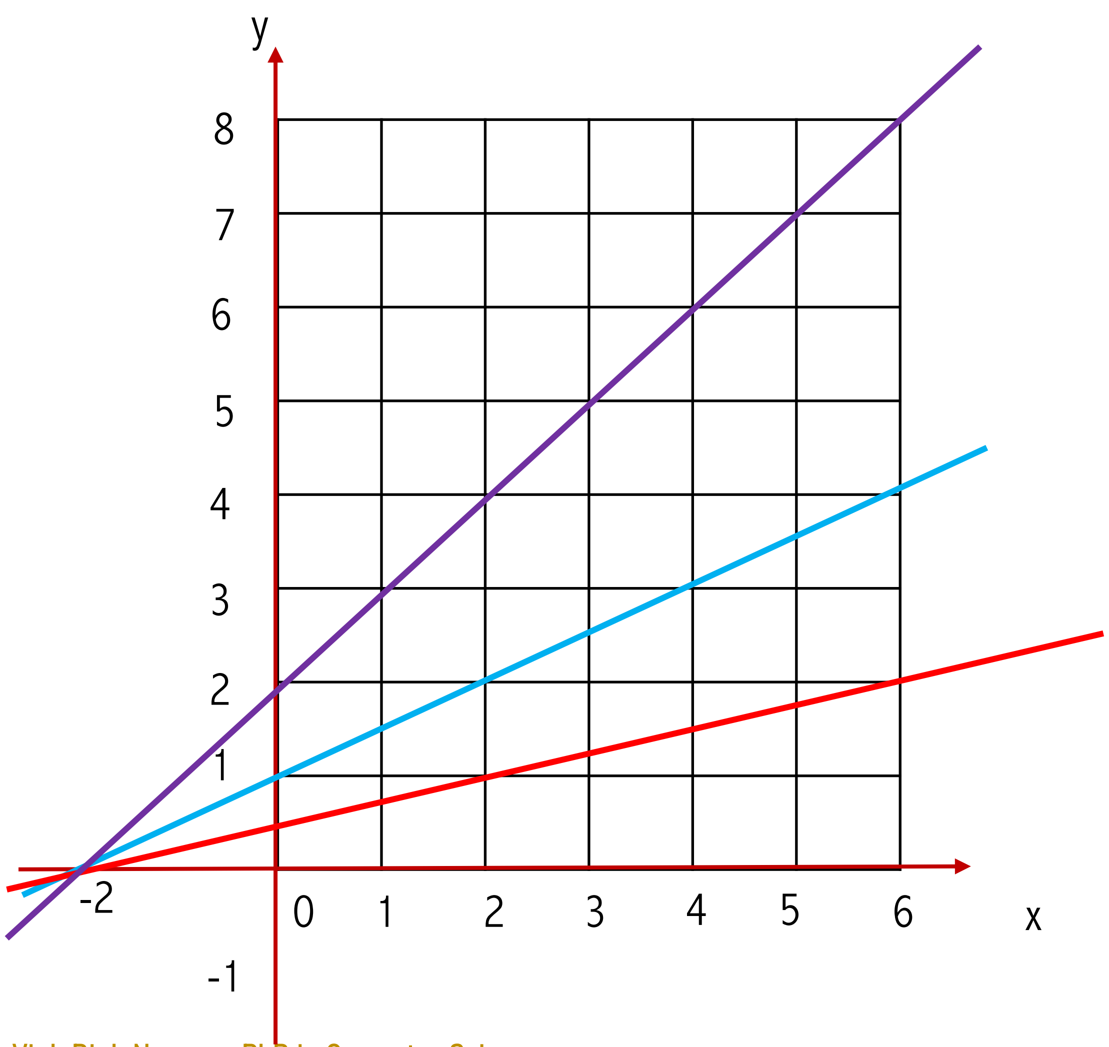
$$\mathrm{Ax + By + C = 0}$$

$$-2x + 4y - 4 = 0$$

$$-2x + 4y - 8 = 0$$

$$-2x + 4y - 0 = 0$$

Line Equation Review: A, B, C Changes



$$y = -\frac{A}{B}x - \frac{C}{B}$$

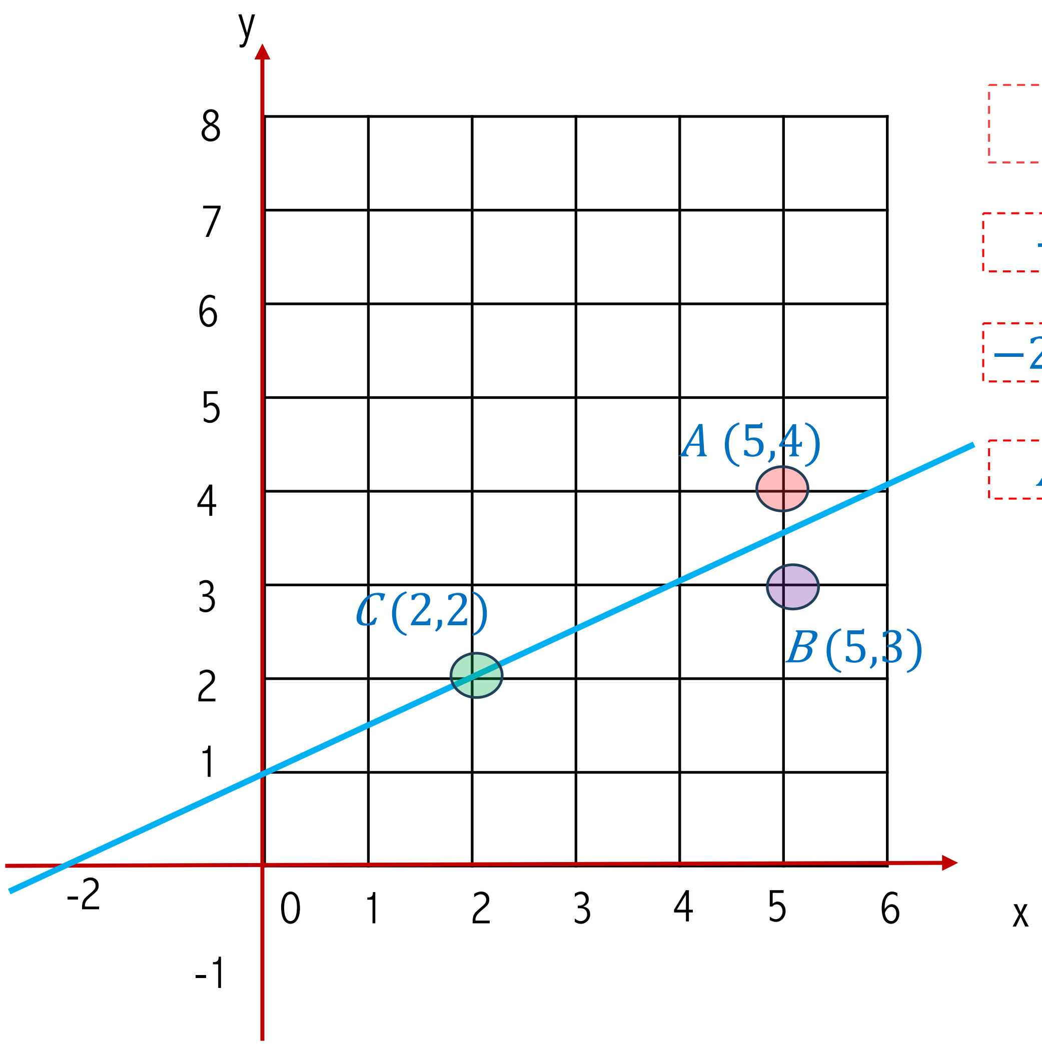
$$\mathrm{Ax} + \mathrm{By} + \mathrm{C} = 0$$

$$-2x + 4y - 4 = 0$$

$$-2x + 8y - 4 = 0$$

$$-4x + 2y - 4 = 0$$

General Form Line Equation



$$Ax + By + C = 0$$

$$-2x + 4y - 4 = 0$$

$$-2 * 5 + 4 * 4 - 4 = 2$$

A is above the line

$$Ax + By + C = 0$$

$$-2x + 4y - 4 = 0$$

$$-2 * 5 + 3 * 4 - 4 = -2$$

B is below the line

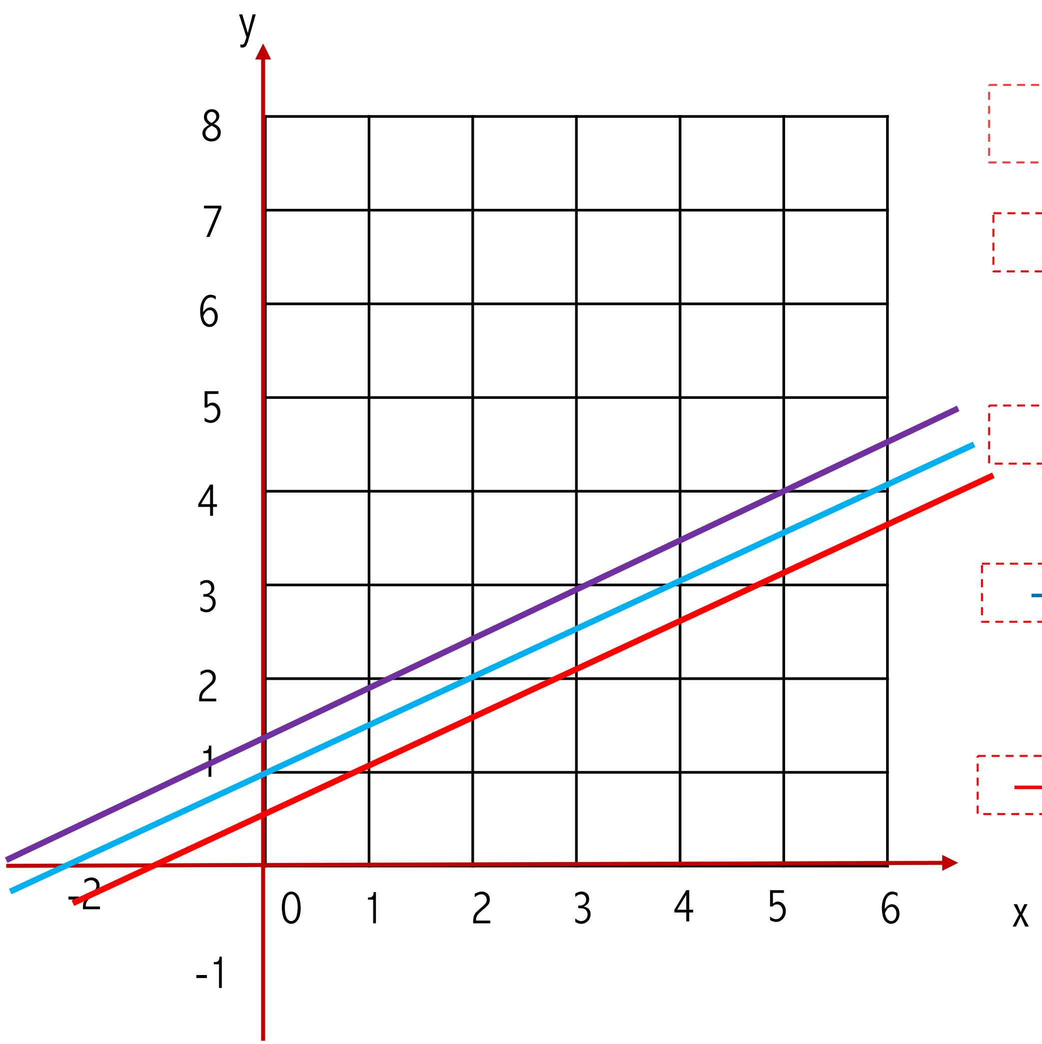
$$Ax + By + C = 0$$

$$-2x + 4y - 4 = 0$$

$$-2 * 2 + 2 * 4 - 4 = 0$$

C is on the line

General Form Line Equation



$Ax + By + C = 0$

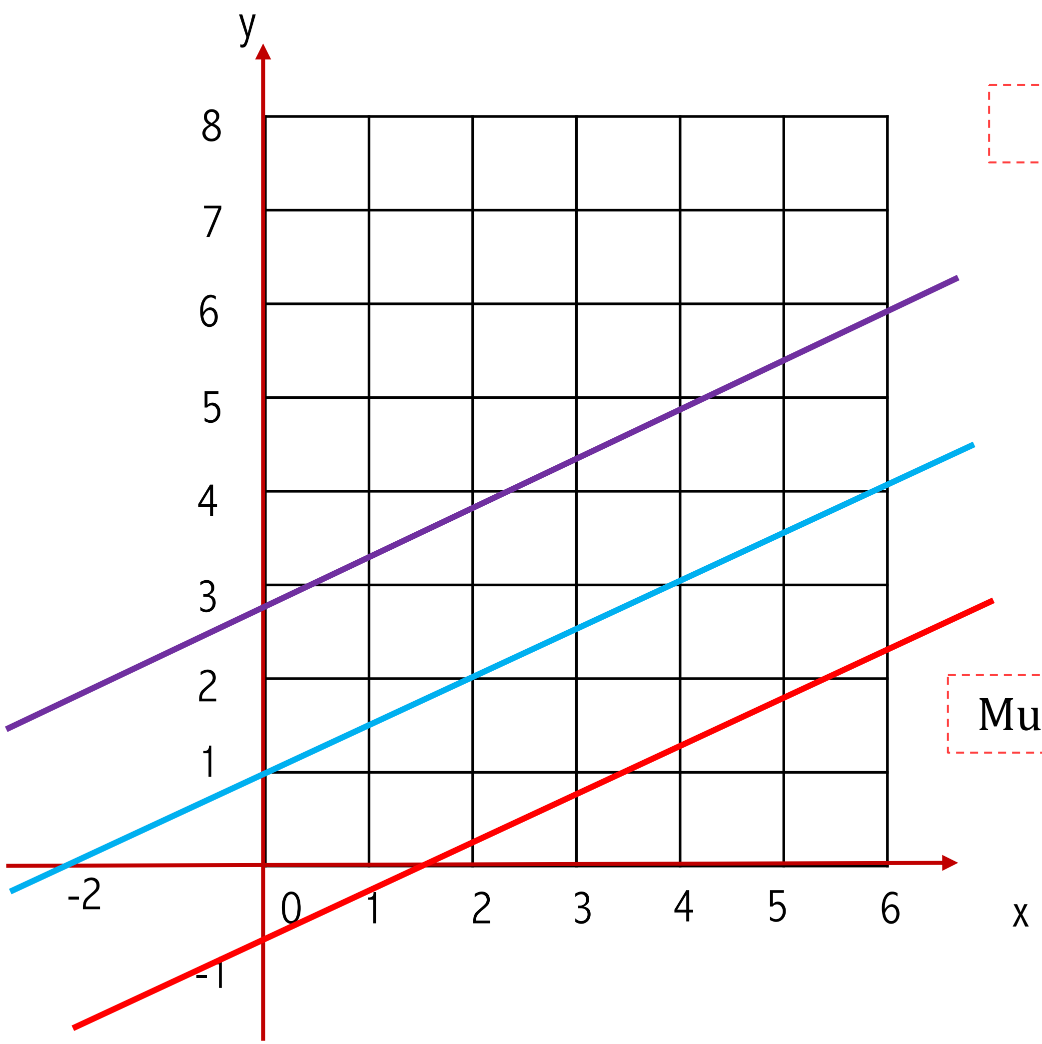
$-2x + 4y - 4 = 0$ → $-2x + 4y - 4 = 1$

$y = 0.5x + 1$ → $y = 0.5x + 1.25$

$-2x + 4y - 4 = 0$ → $-2x + 4y - 4 = -1$

$y = 0.5x + 1$ → $y = 0.5x + 0.75$

General Form Line Equation



$$Ax + By + C = 0$$

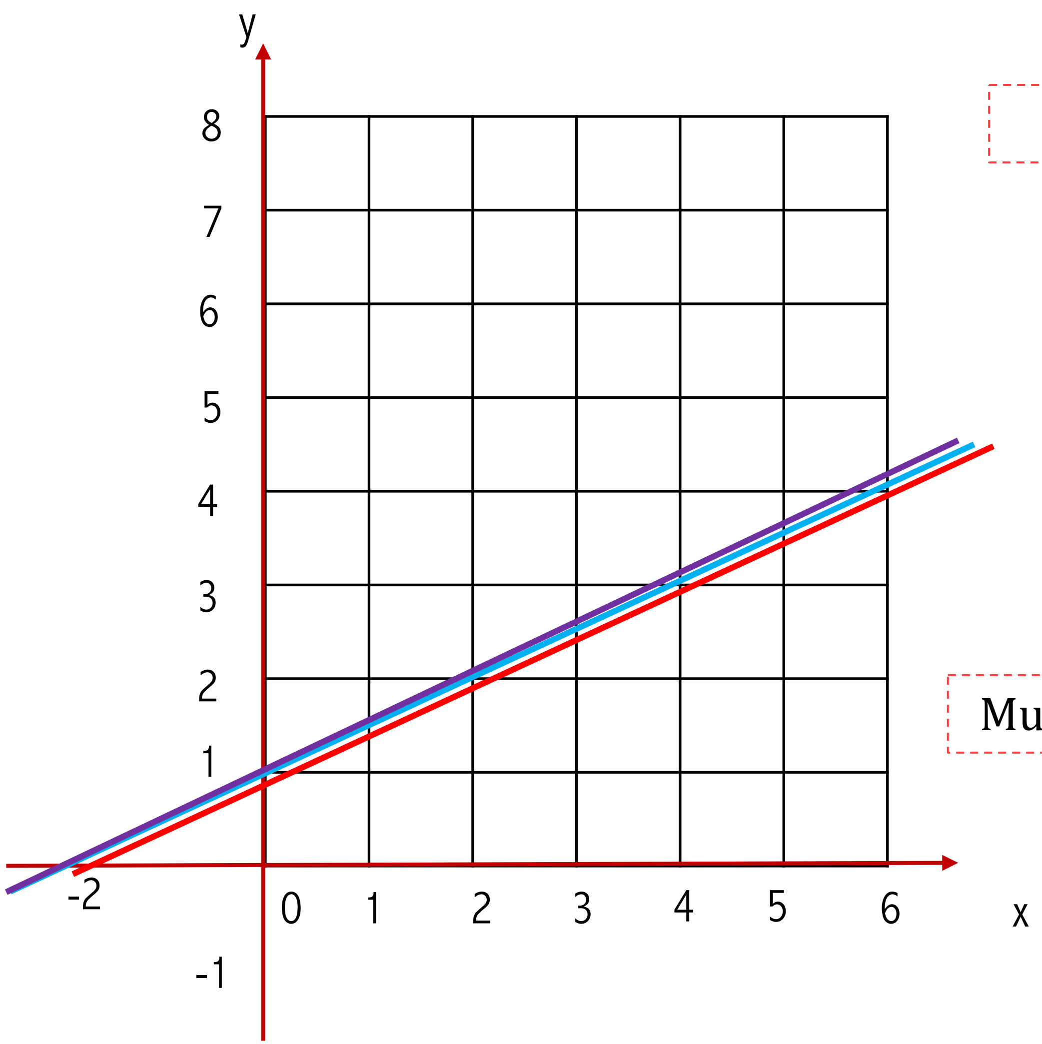
$$-2x + 4y - 4 = 0 \qquad y = 0.5x + 1$$

$$-2x + 4y - 4 = 1 \qquad y = 0.5x + 1.25$$

$$-2x + 4y - 4 = -1 \qquad y = 0.5x + 0.75$$

Multiple with a factor smaller than 1

General Form Line Equation



$$Ax + By + C = 0$$

$$-2x + 4y - 4 = 0 \qquad y = 0.5x + 1$$

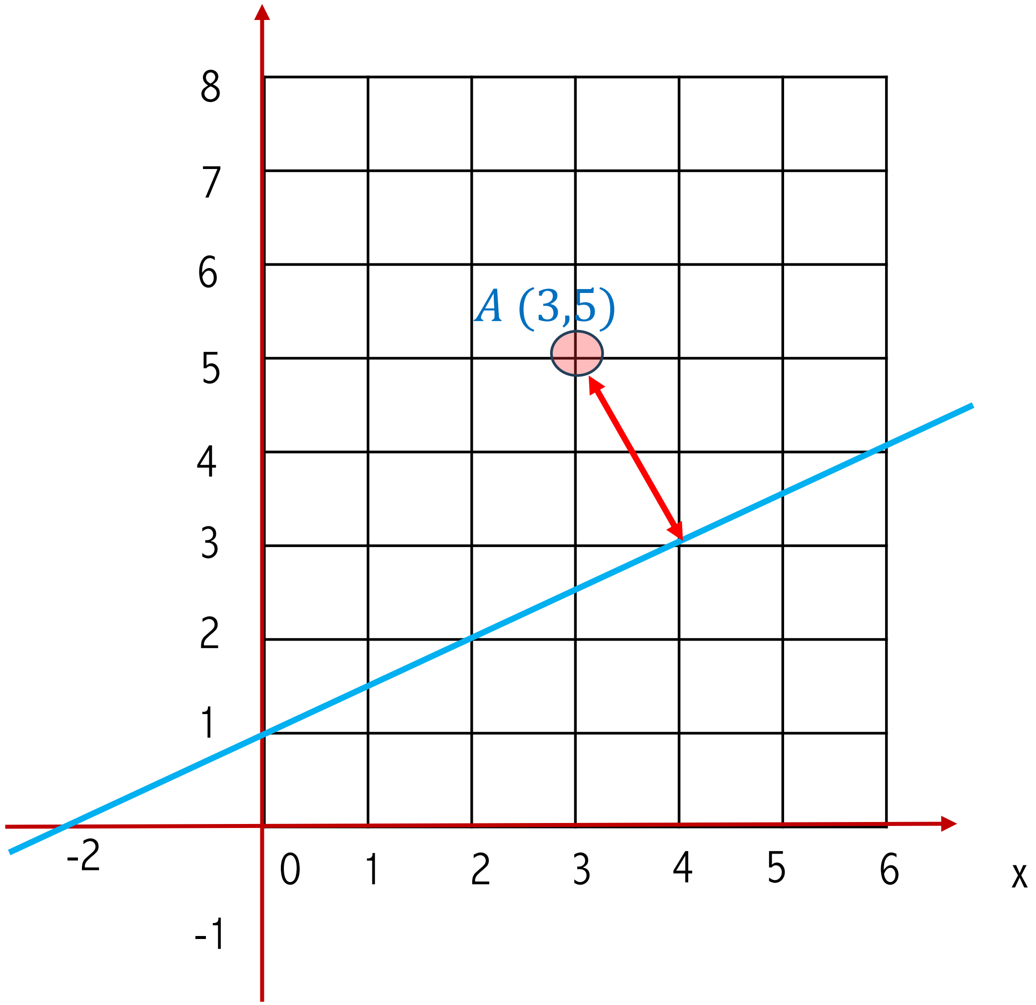
$$-2x + 4y - 4 = 1 \qquad y = 0.5x + 1.25$$

$$-2x + 4y - 4 = -1 \qquad y = 0.5x + 0.75$$

Multiple with a factor greater than 1

This method is used on SVM to in crease or decrease find the the Margin

Distance between a point and a line



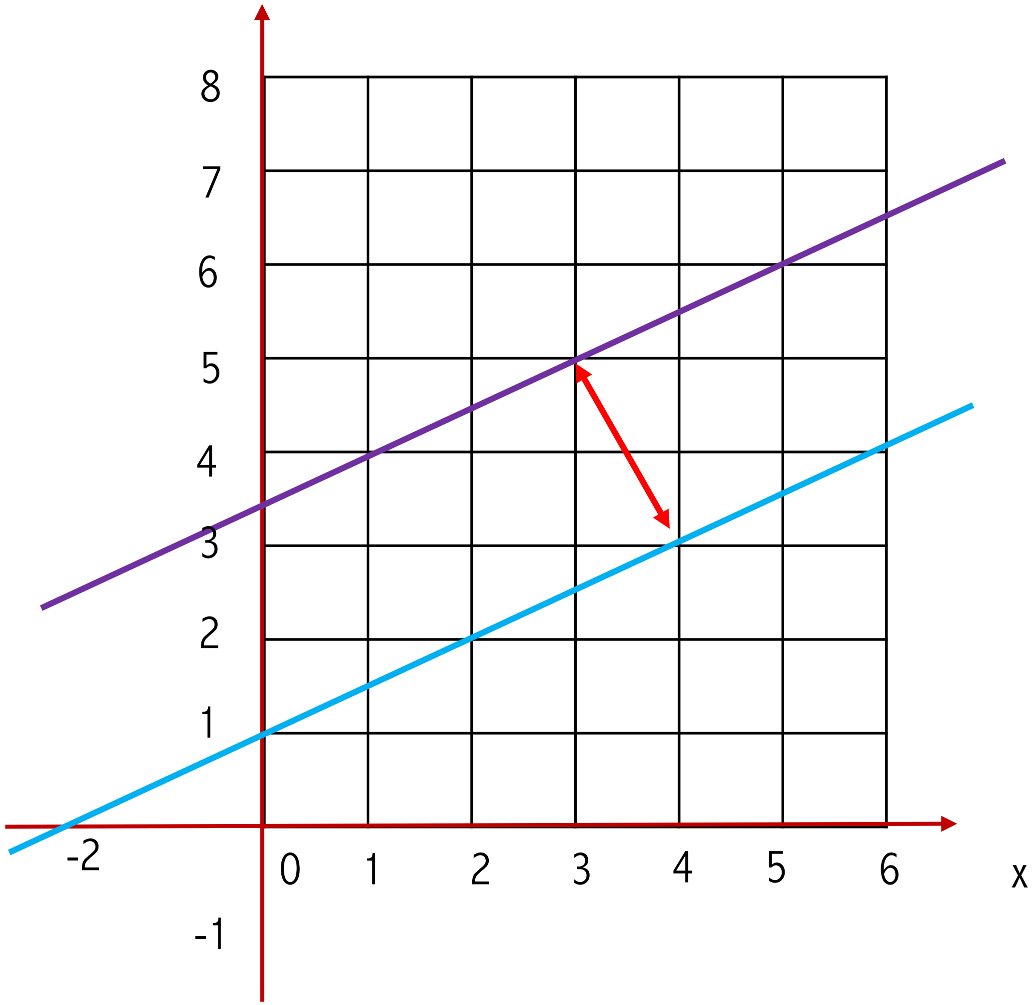
$$Ax + By + C = 0$$

$$-2x + 4y - 4 = 0$$

$$d = \frac{|Ax + By + C|}{\sqrt{A^2 + B^2}}$$

$$d = \frac{|(-2)*3+4*5+(-4)|}{\sqrt{(-2)^2+(4)^2}} = 2.236$$

Distance between parallel lines



$$Ax + By + C = 0$$

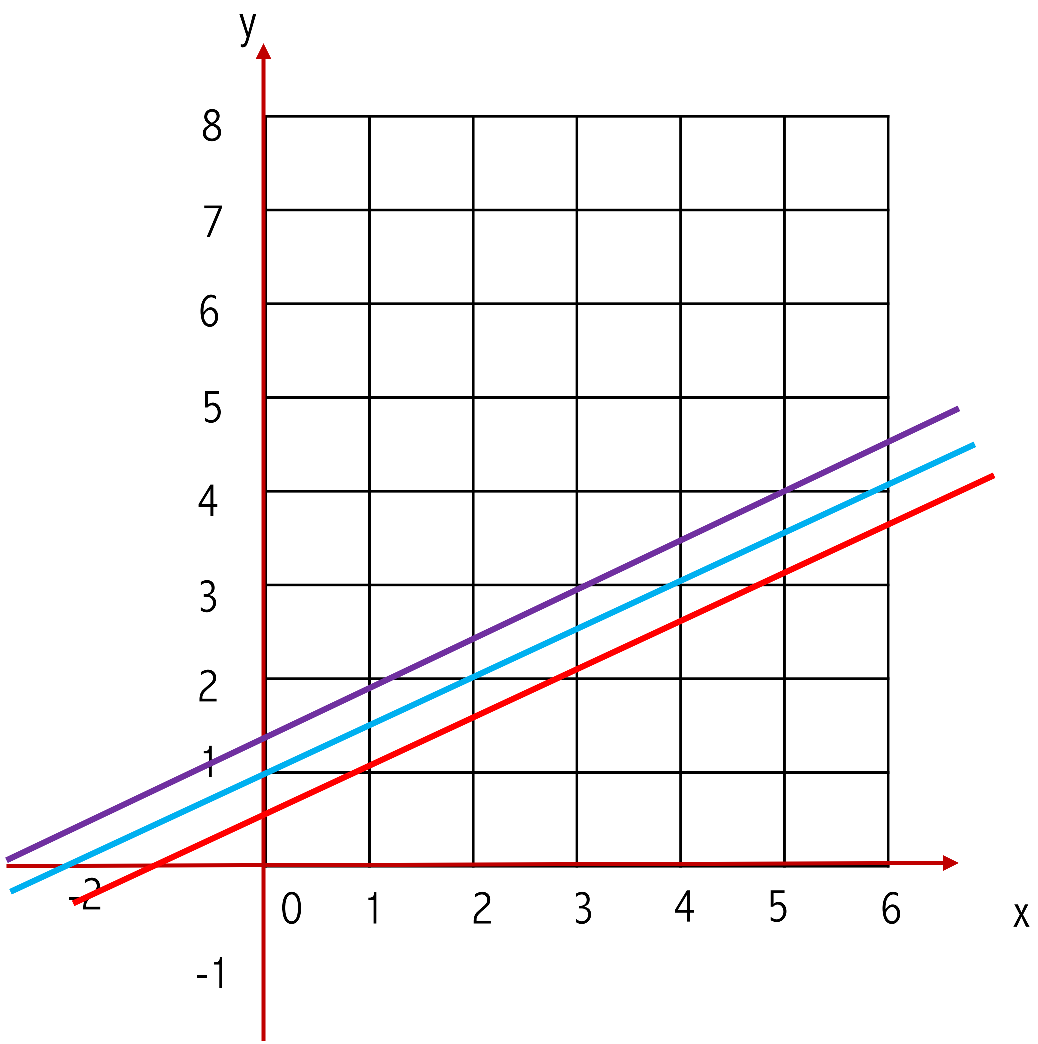
$$-2x + 4y - 14 = 0$$

$$-2x + 4y - 4 = 0$$

$$d = \frac{|C1 - C2|}{\sqrt{A^2 + B^2}}$$

$$d = \frac{|(-4)-(-14))|}{\sqrt{(-2)^2+(4)^2}} = \frac{10}{\sqrt{20}} = 2.236$$

Distance between parallel lines



$$Ax + By + C = 0$$

$$-2x + 4y - 4 = 1$$

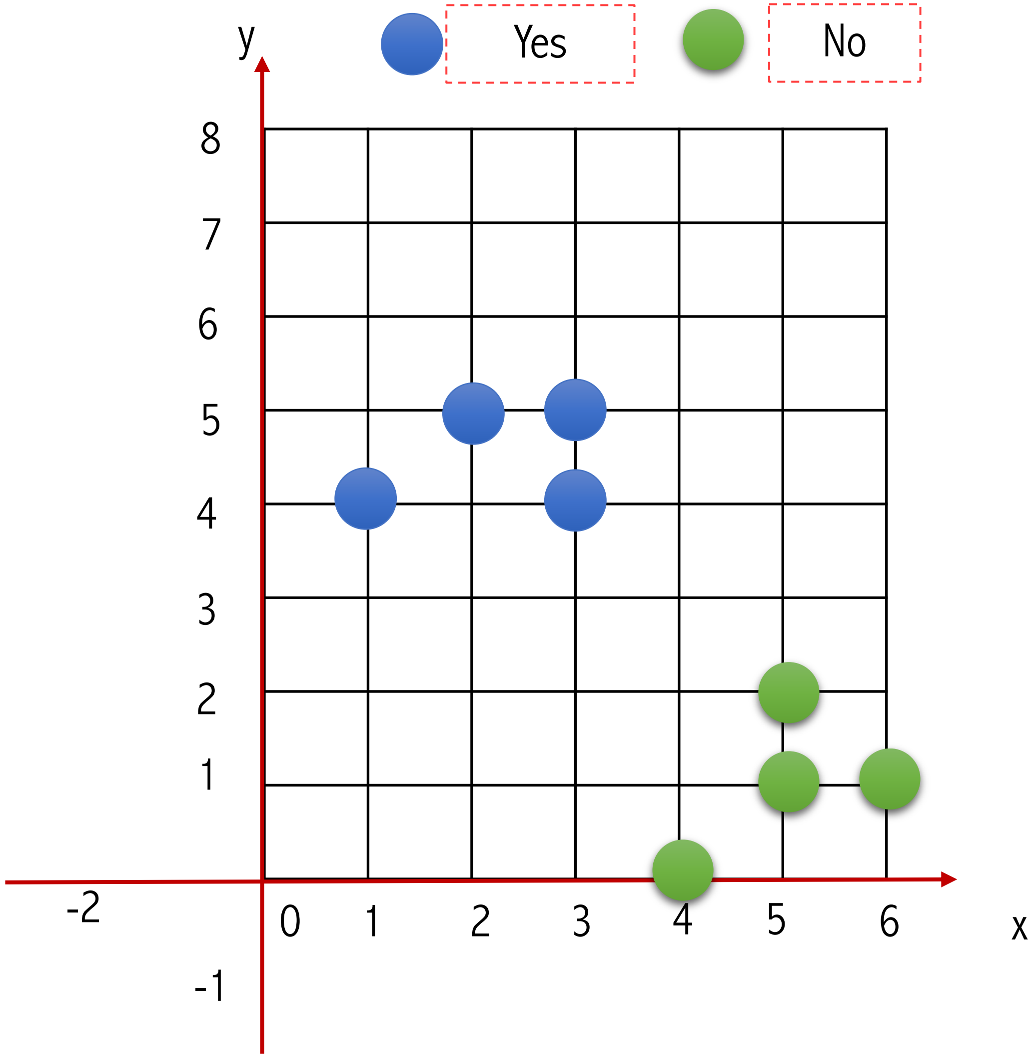
$$-2x + 4y - 4 = 0$$

$$-2x + 4y - 4 = -1$$

$$d = \frac{|(-3)-(-5))|}{\sqrt{(-2)^2+(4)^2}} = \frac{2}{\sqrt{20}} = 0.447$$

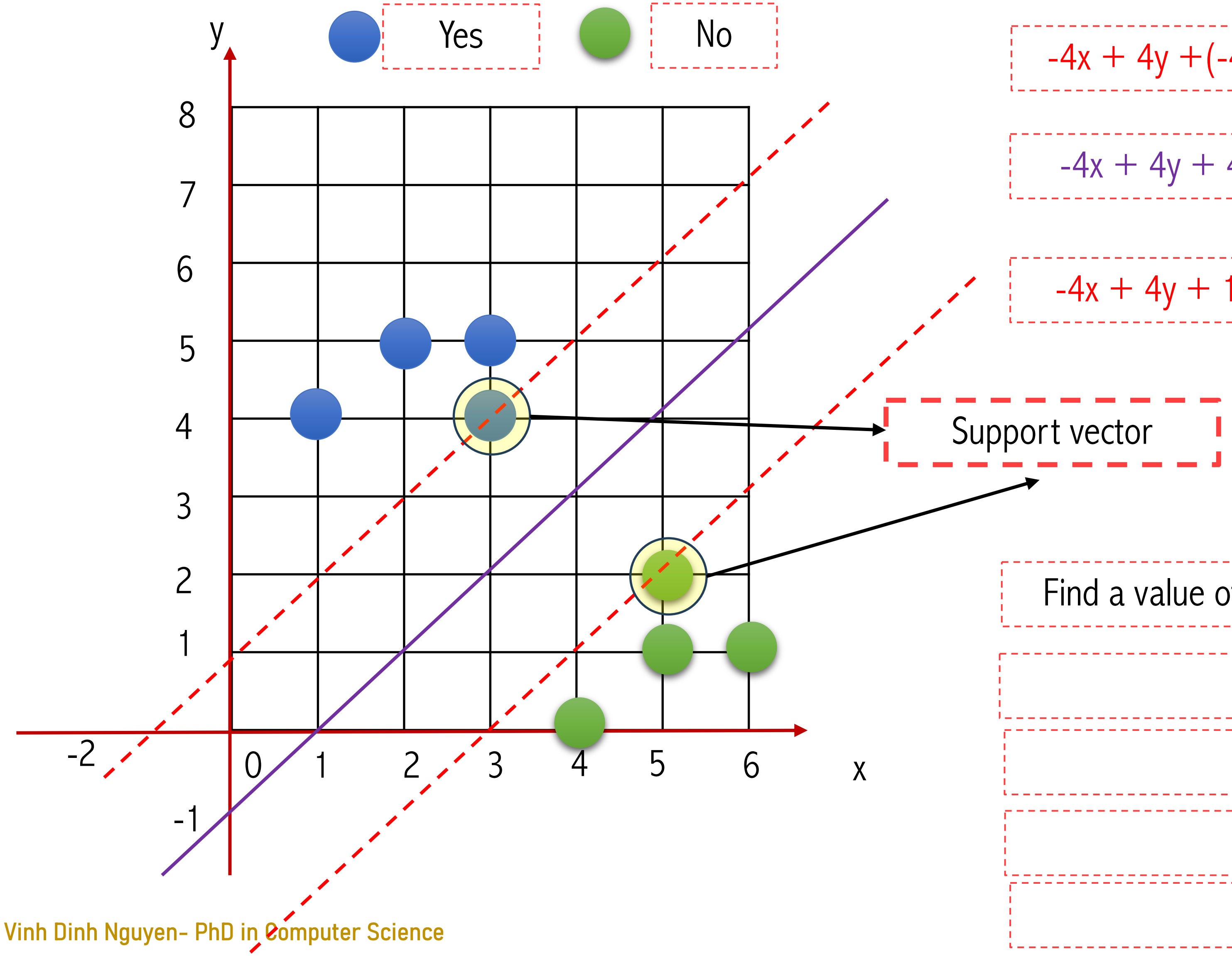
$$d = \frac{|2|}{\sqrt{A^2+B^2}} = \frac{2}{\|W\|_2}$$

Example



Blood Pressure	Cholesterol Level	Disease
1	2	Yes
2	5	Yes
3	5	Yes
3	4	Yes
6	1	No
4	0	No
5	2	No
5	1	No

Example



$-4x + 4y + (-4) = 0 \Longrightarrow -0.5x + 0.5y + 0.5 = 1$

$-4x + 4y + 4 = 0 \Longrightarrow -0.5x + 0.5y + 0.5 = 0$

$-4x + 4y + 12 = 0 \Longrightarrow -0.5x + 0.5y + 0.5 = -1$

Find a value of K so that the left-hand size is equal to one

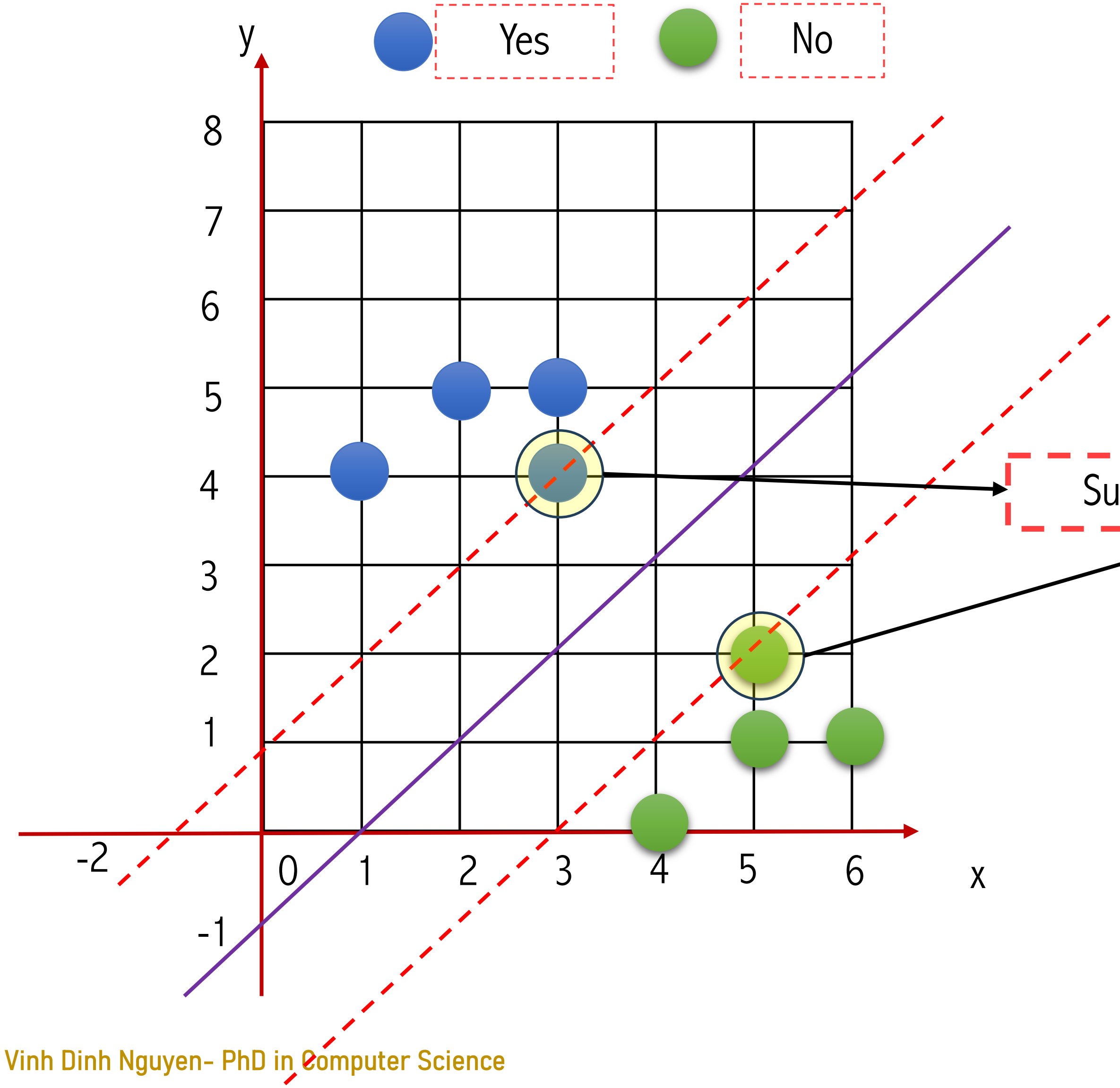
$k(-4x + 4y + 4) = 1$

$k(-4*3 + 4*4 + 4) = 1$

$k*8 = 1$

$k = 1/8$

Example



$-4x + 4y + (-4) = 0$ → k = 1/8 → $-0.5x + 0.5y + 0.5 = 1$

$-4x + 4y + 4 = 0$ → $-0.5x + 0.5y + 0.5 = 0$

$-4x + 4y + 12 = 0$ → $-0.5x + 0.5y + 0.5 = -1$

Find a value of K so that the left-hand size is equal to 1

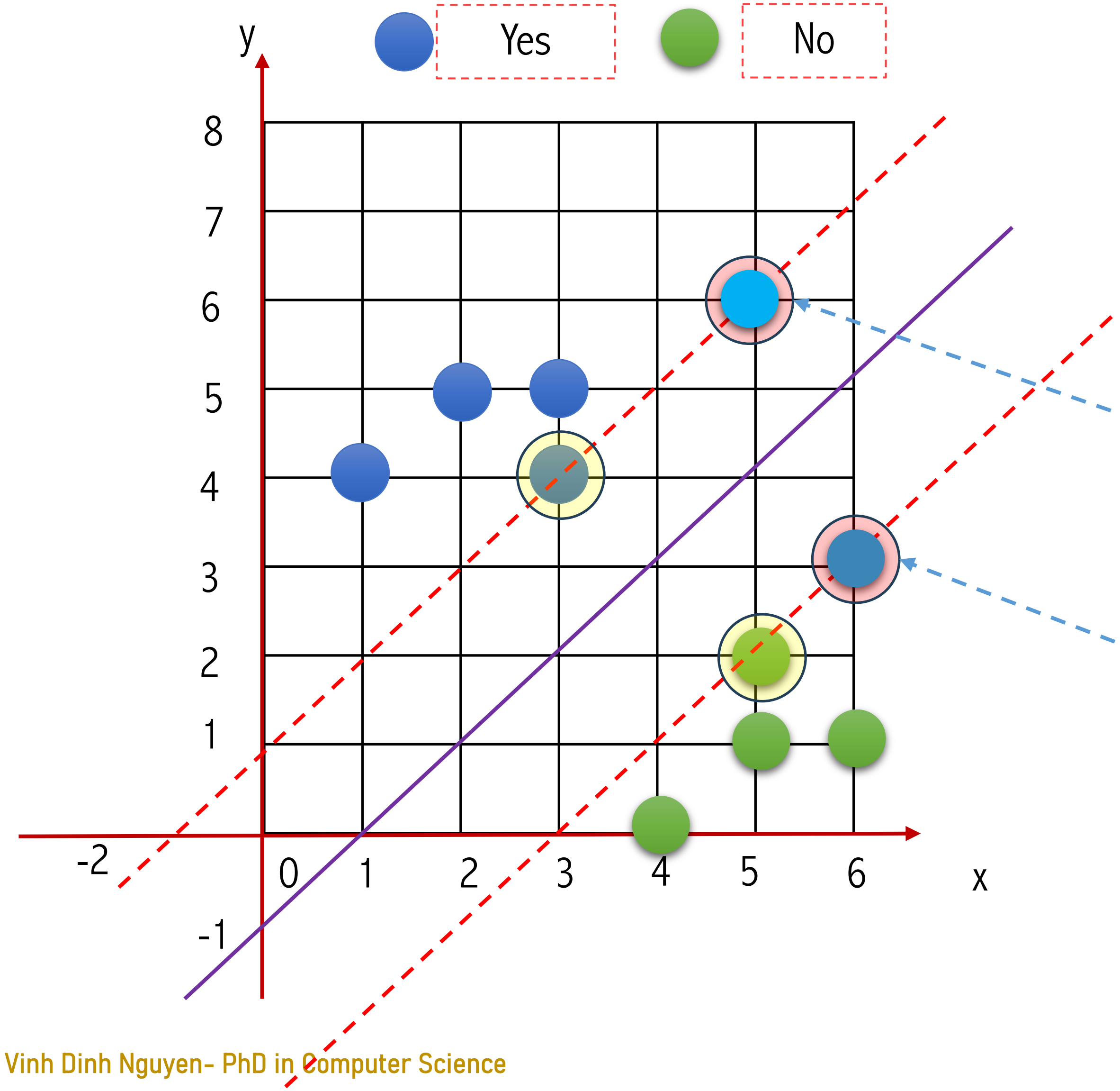
$k(-4x + 4y + 4) = 1$

$k(-4*3 + 4*4 + 4) = 1$

$k*8 = 1$

$k = 1/8$

Example



k = 1/8

-4x + 4y +(-4) = 0 → -0.5x + 0.5y + 0.5 = 1

-4x + 4y + 4 = 0 → -0.5x + 0.5y + 0.5 = 0

-4x + 4y + 12= 0 → -0.5x + 0.5y + 0.5 = -1

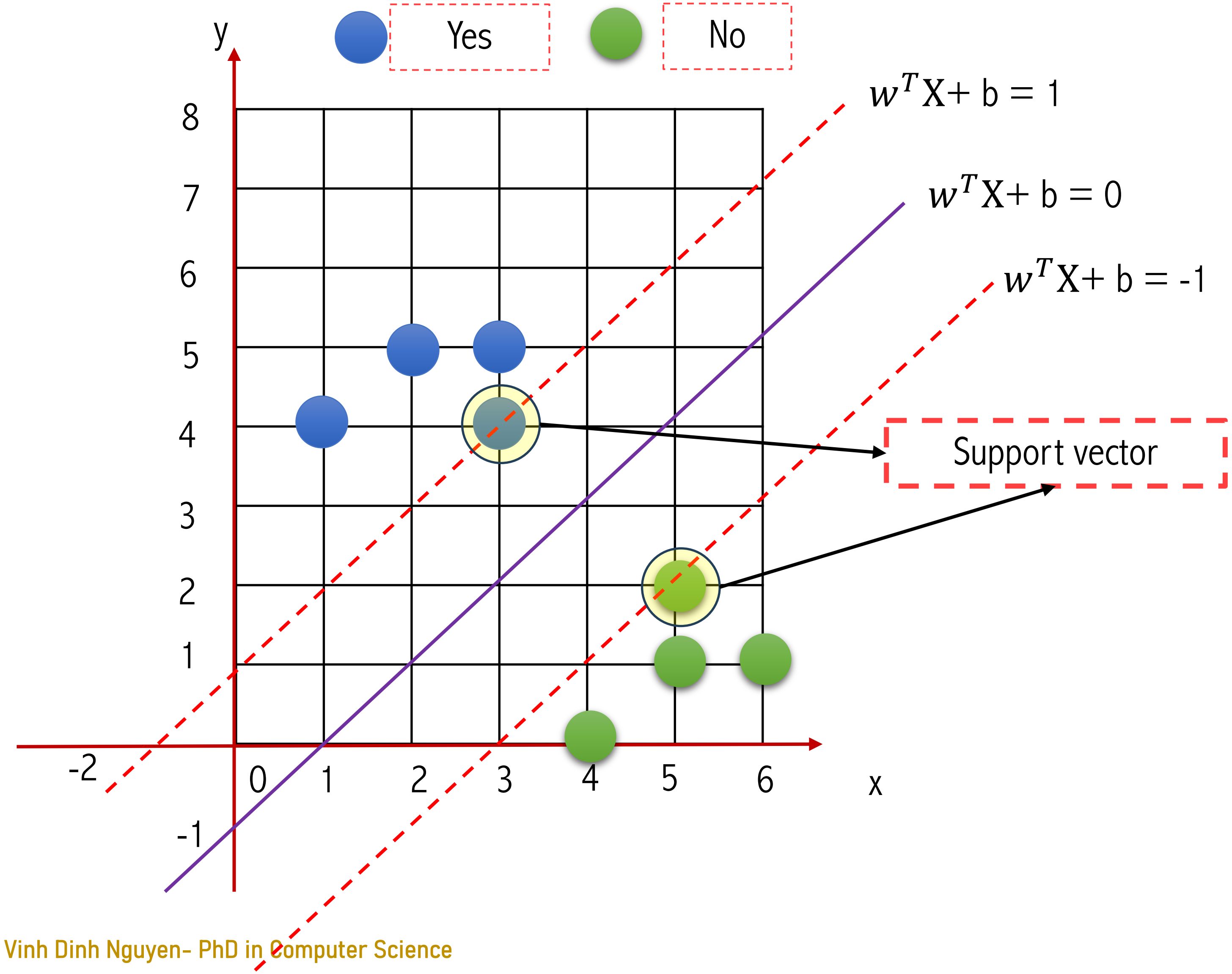
-0.5x + 0.5y + 0.5

-0.5x6 + 0.5x3 + 0.5 = -1

-0.5x + 0.5y + 0.5

-0.5x5 + 0.5x6 + 0.5 = 1

Example



b = 0.5

-0.5x + 0.5y + b = 1

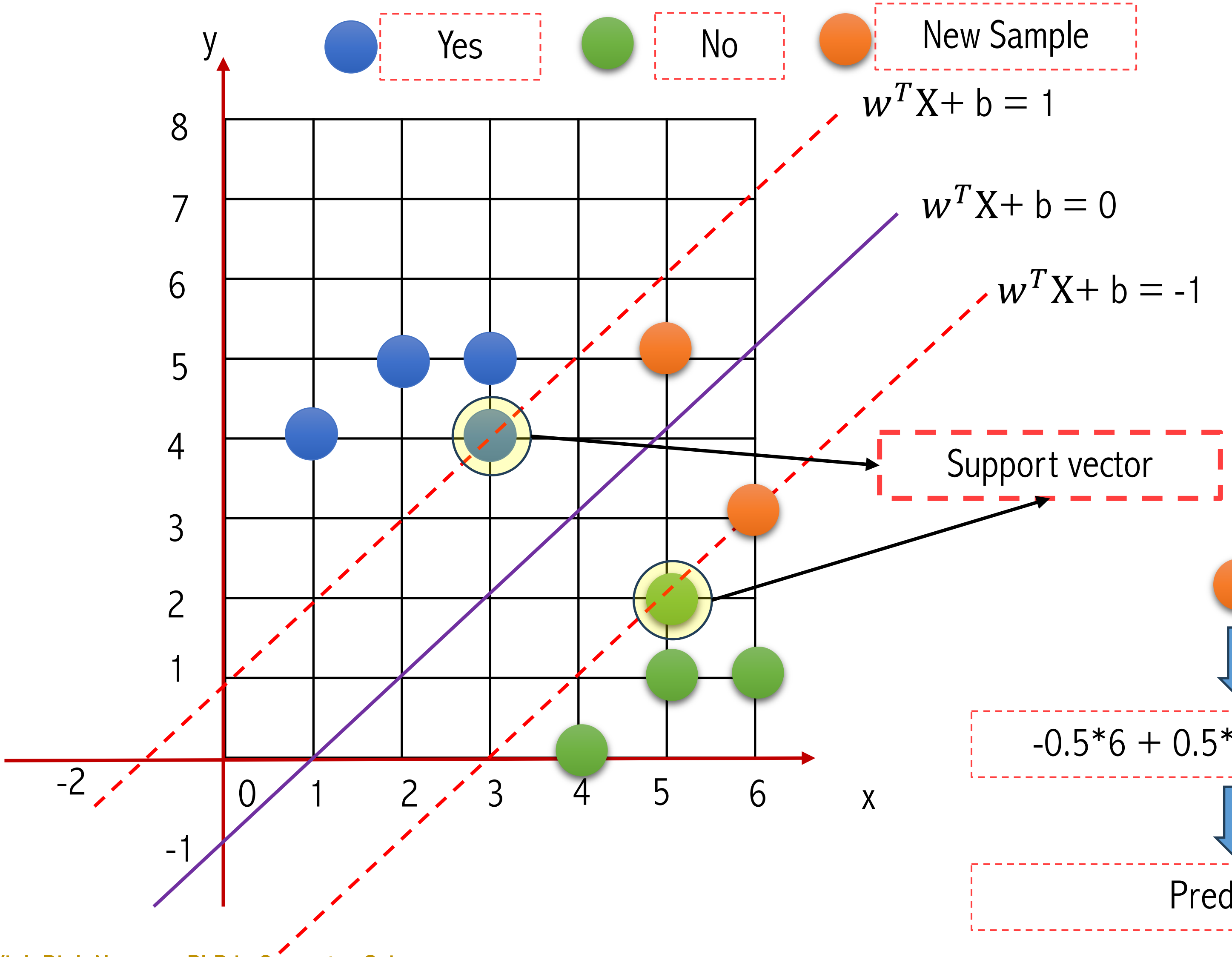
-0.5x + 0.5y + b = 0

-0.5x + 0.5y + b = -1

w^TX+ b = 0

$X = \begin{bmatrix} x \\ y \end{bmatrix}$ $W = \begin{bmatrix} -0.5 \\ 0.5 \end{bmatrix}$

Example



b = 0.5

-0.5x + 0.5y + b = 1

-0.5x + 0.5y + b = 0

-0.5x + 0.5y + b = -1

A(6,3)

-0.5*6 + 0.5*3 + 0.5 = -1.5

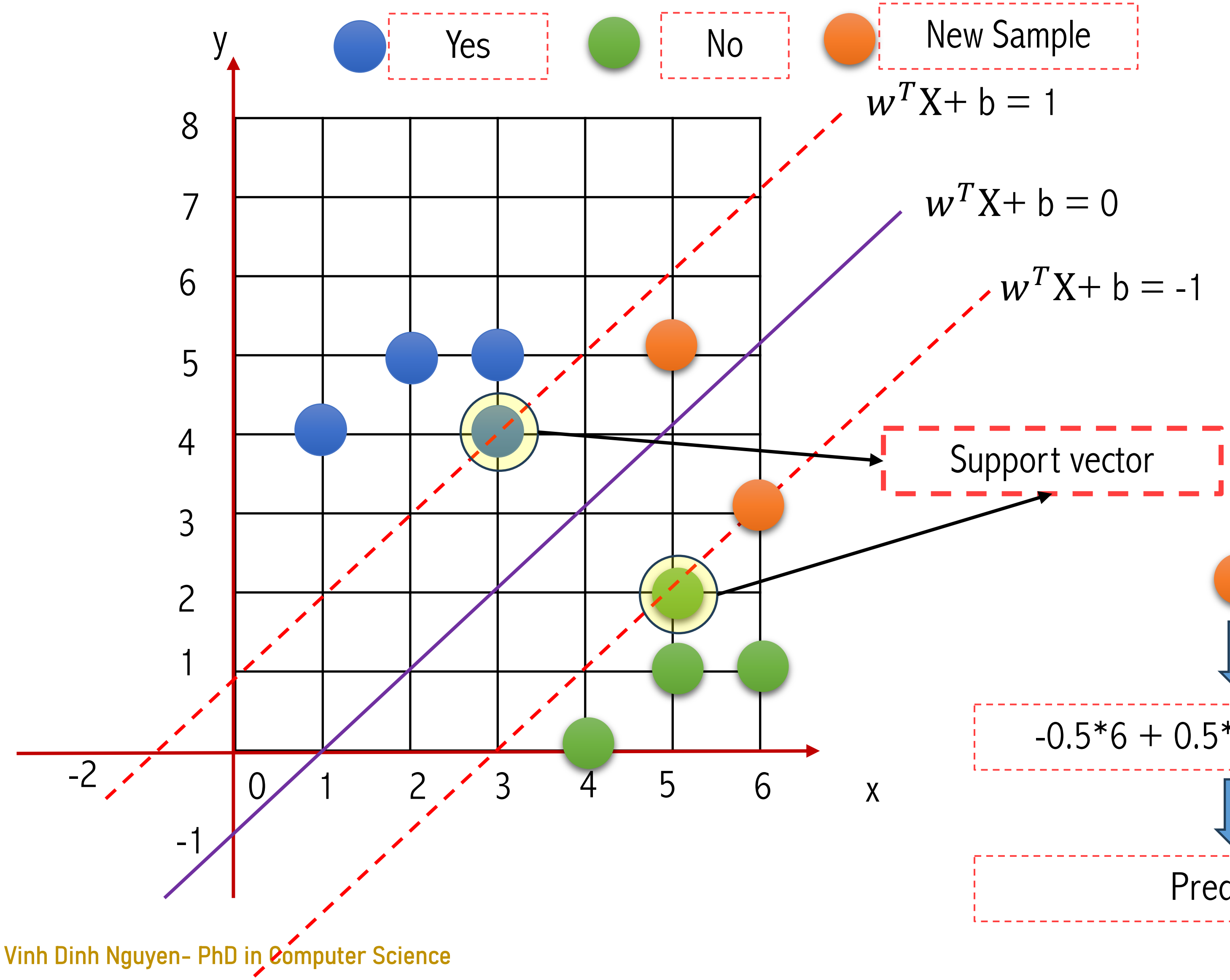
Predict No

A(5,5)

-0.5*5 + 0.5*5 + 0.5 = 0.5

Predict Yes

Example



$$\text{Predict} = \begin{cases} +1 \; if\, w^T\text{X+ b} \geq 0 \\ -1 \; if\, w^T\text{X+ b} < 0 \end{cases}$$

A(6,3)

-0.5*6 + 0.5*3 + 0.5 = -1.5

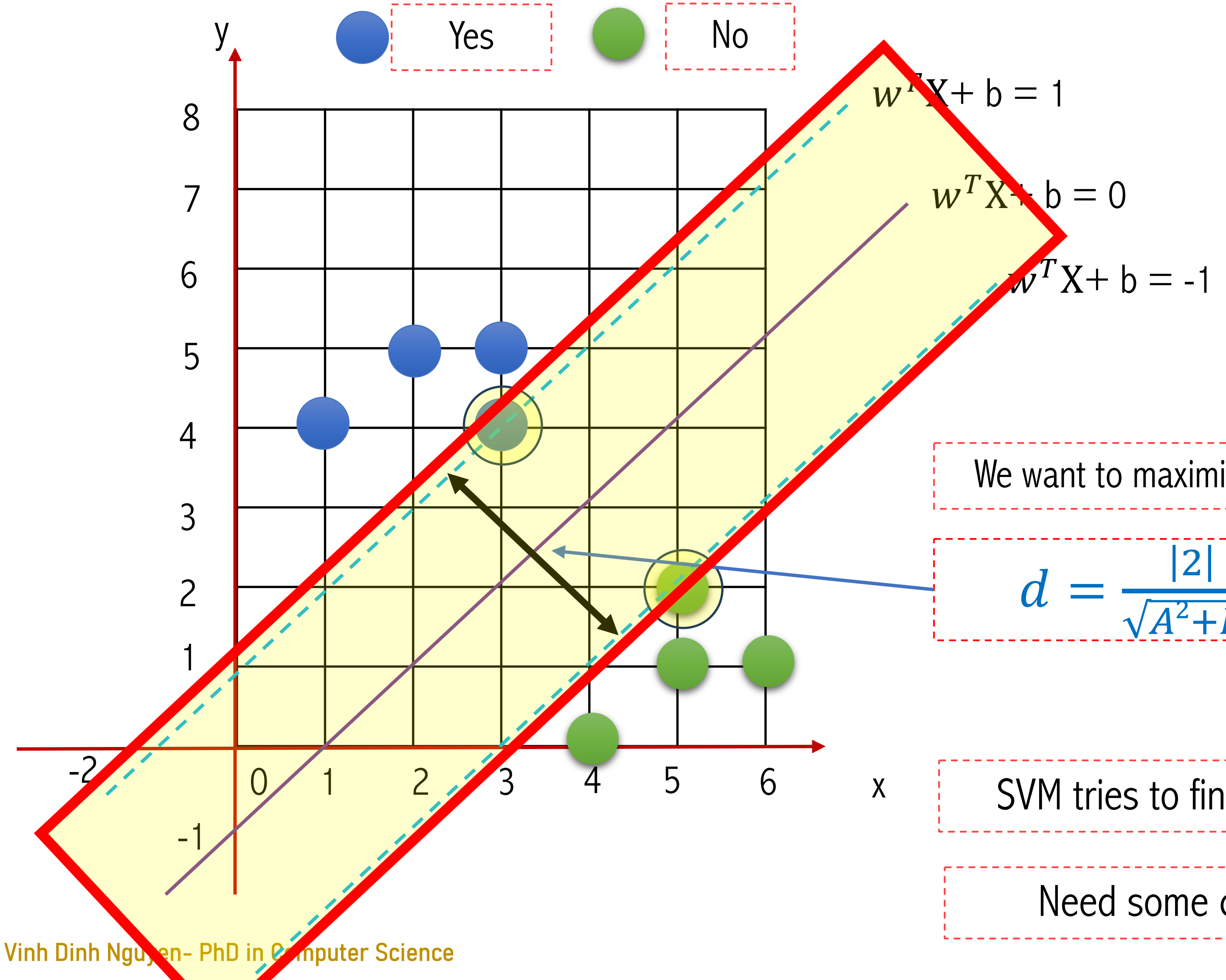
Predict No

A(5,5)

-0.5*5 + 0.5*5 + 0.5 = 0.5

Predict Yes

Example



$$Y = \begin{cases} +1 \; if \, w^T X + b \geq 0 \\ -1 \; if \, w^T X + b < 0 \end{cases}$$

We want to maximize the distance d

$$d = \frac{|2|}{\sqrt{A^2+B^2}} = \frac{2}{\|W\|_2}$$

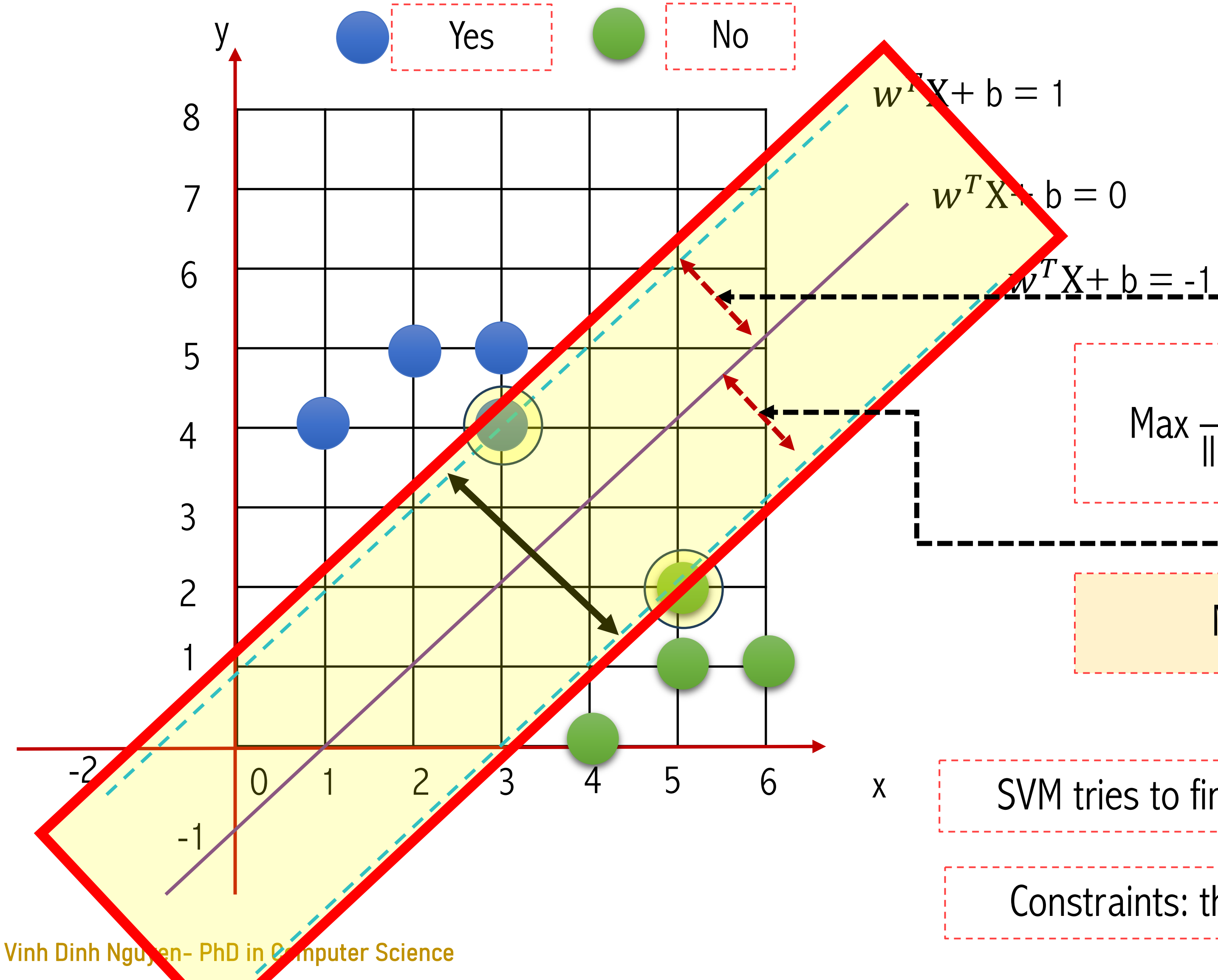
We want to minimize $\|W\|_2$

$$d = \frac{|2|}{\sqrt{A^2+B^2}} = \frac{2}{\|W\|_2}$$

SVM tries to find a hyperplane that maximizes the width of this margin

Need some constraints because the margin can be infinitely large

Example



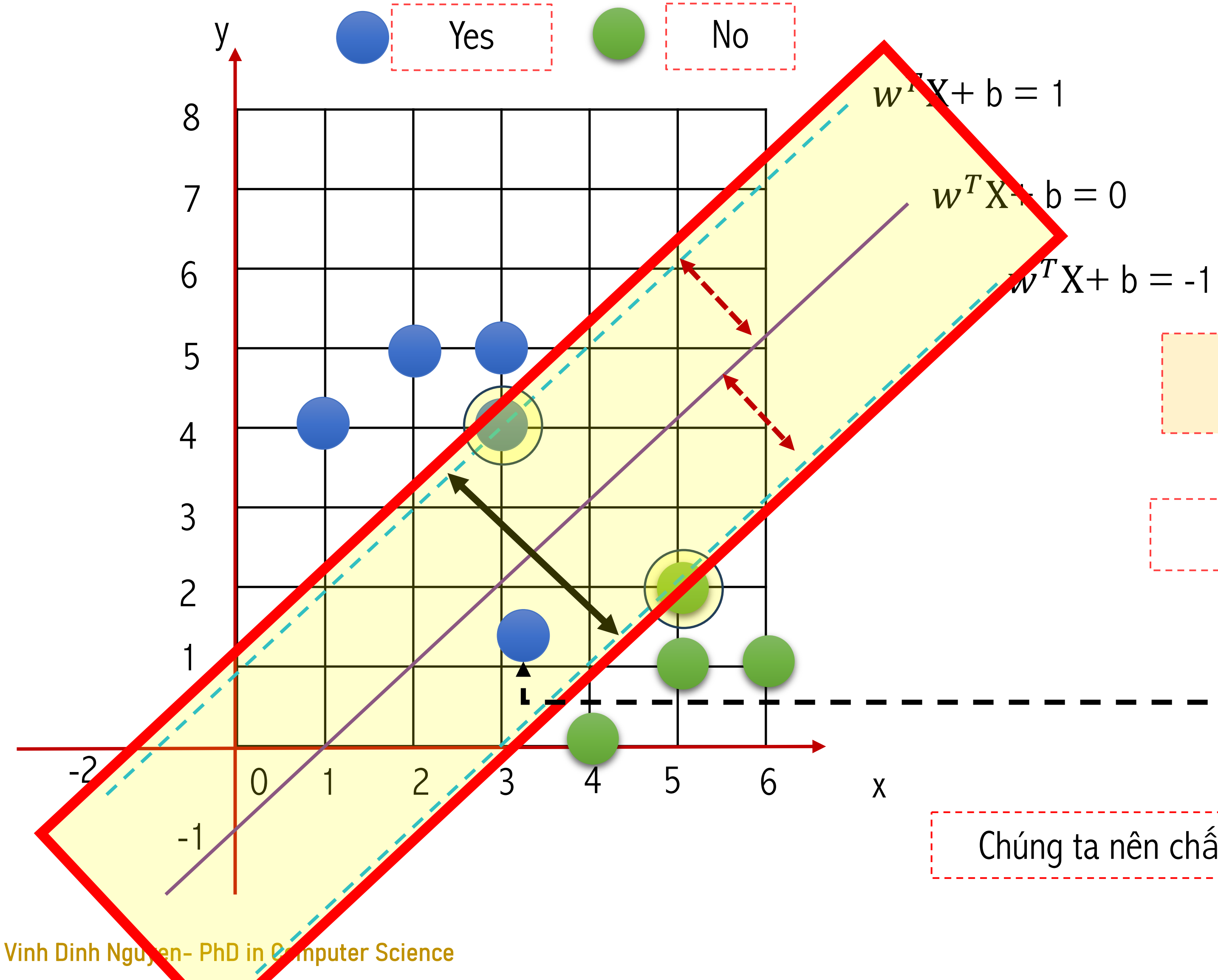
Max $\frac{2}{\|W\|}$ such that $\begin{cases} w^T X + b \geq 1 \; if \; Y = +1 \\ w^T X + b \leq -1 \; if \; Y = -1 \end{cases}$

Max $\frac{2}{\|W\|_2}$ such that Y * $(w^T X + b) \geq 1$

SVM tries to find a hyperplane that maximizes the width of this margin

Constraints: the margin should not span beyond the support vectors

Example

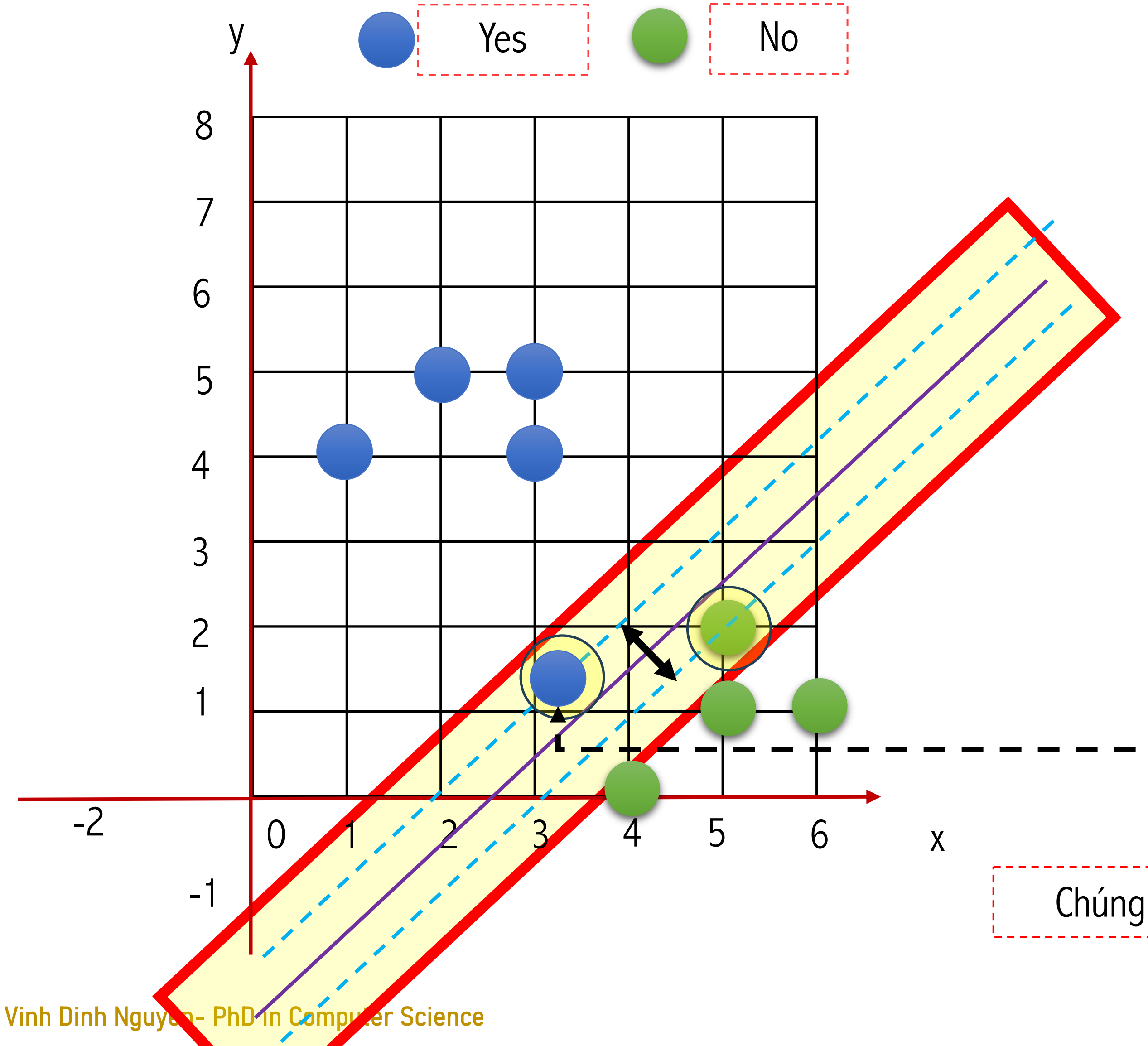


Max $\frac{2}{\|W\|_2}$ such that Y * $(w^TX + b) \geq 1$

What should we do if a green data point is noise?

Chúng ta nên chấp nhận như là miss-classification hay là thay đổi hyperplane

Example

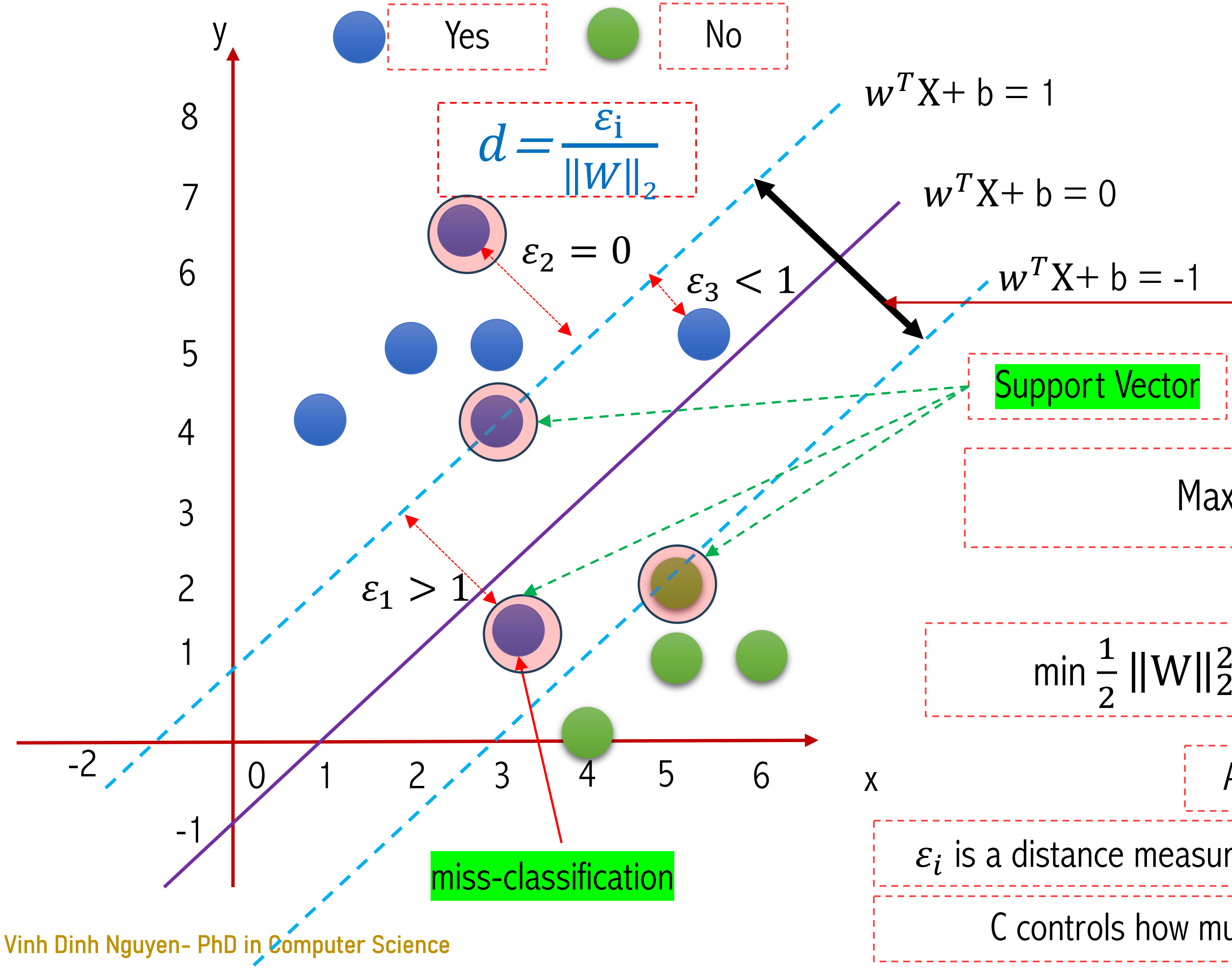


$$\text{Max } \frac{2}{\|W\|_2} \text{ such that } Y * (w^T X + b) \geq 1$$

What should we do if a green data point is noise?

Chúng ta nên chấp này như là miss-classification hay là thay đổi hyperplane

Example



$$d = \frac{|2|}{\sqrt{A^2+B^2}} = \frac{2}{\|W\|_2}$$

Slack variable in SVM

Max $\frac{2}{\|W\|_2}$ such that Y * (w^TX+ b) ≥ 1

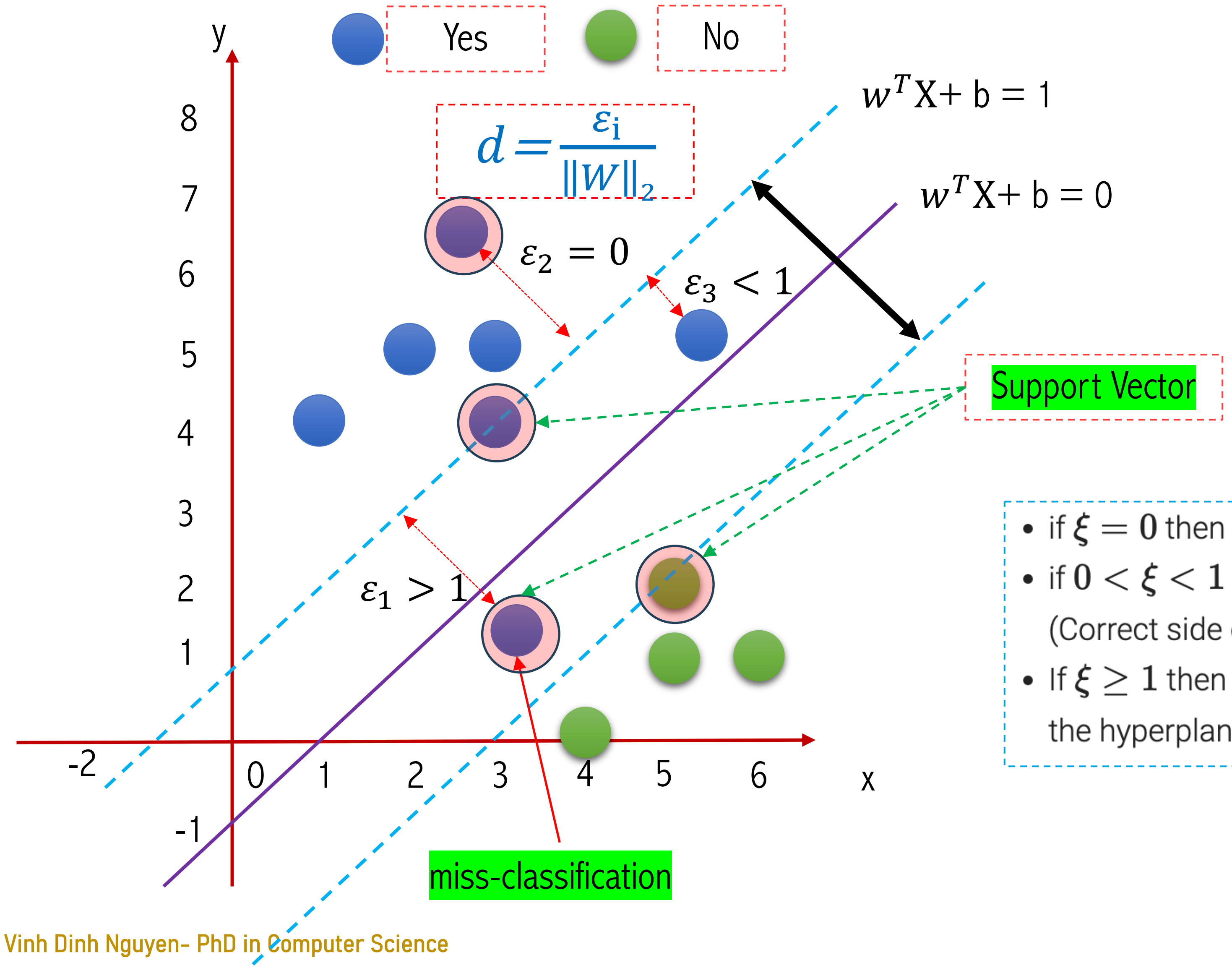
min $\frac{1}{2}\|W\|_2^2 + \boxed{C \sum_i^N \varepsilon_i}$ such that Y * (w^TX+ b) $\geq 1 - \boxed{\varepsilon_i}$

Allow for miss classification

ε_i is a distance measure of the data points from their corresponding blue line.

C controls how much weight should set on the misclassification data

Example



- if $\xi = 0$ then the corresponding point ξ is on the margin or further away.
- if $0 < \xi < 1$ then the point ξ is within the margin and classified correctly (Correct side of the hyperplane).
- If $\xi \geq 1$ then the point is misclassified and present at the wrong side of the hyperplane.

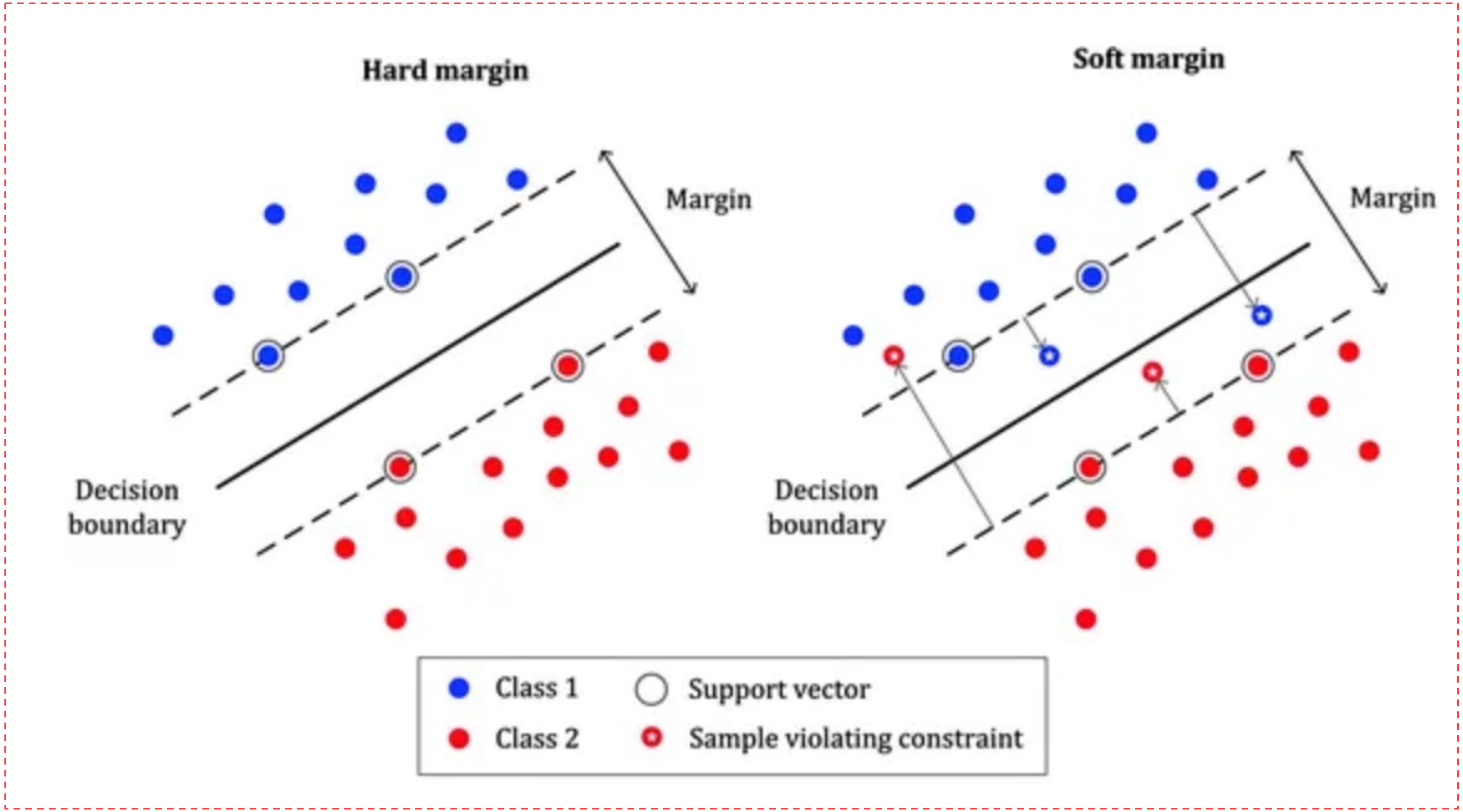
Dicussion

$$\min \frac{1}{2} ||W||^2 + C \sum_i^N \varepsilon_i \text{ such that } Y * (w^T X + b) \geq 1 - \varepsilon_i \qquad \varepsilon_i \geq 0$$

How about the case C is small?
Soft Margin Classifier

How about the case C is large?
Hard Margin Classifier

Hard Margin vs Soft Margin



Further Study

Primal Problem

$$\min \frac{1}{2} ||W||_2^2 + C \sum_i^N \varepsilon_i \text{ such that } Y * (w^T X + b) \geq 1 - \varepsilon_i$$

$$\varepsilon_i \geq 0$$

Another form:

$$\text{Objective Function}: \min_{\beta, b, \xi_i} \left\{ \frac{||\beta^2||}{2} + C \sum_{i=1}^{n} \xi_i^2 \right\}$$

$$\text{s.t Linear Constraint}: y_i(\beta^T x_i + b) \geq 1 - \xi_i$$

The **Lagrangian** can be defined as below. Notice we only need one **Lagrange Multiplier** due to the dropped constraint.

$$L = \frac{||\beta^2||}{2} + C \sum_{i=1}^{n} \xi_i^2 - \sum_{i=1}^{n} \alpha_i (y_i(\beta^T x_i + b) - 1 + \xi_i)$$

The HyperParameter C is also called as *Regularization Constant*.

If k = 1 , then the loss is named as Hinge Loss and if k =2 then its called Quadratic Loss

Kernel trick migh apply here

$$\phi(x_i)^T \phi(x_j)$$

The Dual Objective can be written as,

$$\max_{\alpha} L_{dual} = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \left(\boxed{x_i^T x_j} + \frac{1}{2C} \delta_{ij} \right)$$

$$\text{s.t constraint}: \alpha_i \geq 0, \forall i \in D, \text{ and } \sum_{i=1}^{n} \alpha_i y_i = 0$$

However, this way we won't be able use the objective function to solve for non-linear cases. Hence, we will find an equivalent problem named **Dual Problem** and solve that using **Lagrange Multipliers.**


Thank you