Data Science and Artificial Intelligence

Machine Learning

Support Vector Machine

Lecture No. 4

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

- Topic — SVM's
- Topic — Primal, dual, Kernels.
- Topic
- Topic
- Topic

Topics to be Covered

- **Topic** Svm
- **Topic** Kernels
- **Topic** Theorem on Kernels
- **Topic** Soft margin svm
- **Topic** Questions


NEVER STOP
BELIEVING
IN HOPE
BECAUSE
MIRACLES
HAPPEN
EVERY DAY
blurmark.com

SVM (algorithm)

Things to Remember

1. Prob $\min \frac{1}{2}\|\omega\|^2$

st $y_i(\omega x_i + b) \geq 1$

$\omega = [\omega_1, \omega_2 - - \omega_D]$

2. Primal $\min_{\omega, b} \max_{\lambda_i} \frac{1}{2}\|\omega\|^2 + \sum_{i=1}^{N} \lambda_i \left(1 - y_i(\omega x_i + b)\right)$

$\lambda_i \geq 0$

SVM (algorithm)

3. $\omega = \sum \lambda_i y_i x_i$

$\sum \lambda_i y_i = 0$

4. dual $\max\limits_{\lambda_i} \left\{ \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{j=1}^{N} \sum_{i=1}^{N} \lambda_i \lambda_j y_i y_j x_i x_j^T \right\}$

$\lambda_i \geq 0$

$\sum \lambda_i y_i = 0$

REVISION

SVM (algorithm)

5. $\underbrace{x_i\, x_j^T}_{\text{inner product}} \Rightarrow (x_i^1\ x_i^2 - - - x_i^D) \begin{bmatrix} x_j^1 \\ x_j^2 \\ x_j^3 \\ \vdots \\ x_j^D \end{bmatrix} \Rightarrow \{x_i^1 x_j^1 + x_i^2 x_j^2 + x_i^3 x_j^3 + - - - - x_i^D x_j^D\}$

6. higher dimension $\left\{ \begin{array}{l} x_i \longrightarrow \phi(x_i) \\ x_j \longrightarrow \phi(x_j) \end{array} \right\}$

SVM (algorithm)

7. dual max $\sum_{i=1}^{N}\lambda_i - \frac{1}{2}\sum\sum \lambda_i \lambda_j y_i y_j \underbrace{\phi(x_i)\phi(x_j)^T}_{K(x_i,x_j)}$

8. actual conversion of points into higher dimension is not needed

$$\cancel{\begin{array}{l} x_i \rightarrow \phi(x_i) \\ x_j \rightarrow \phi(x_j) \end{array}}$$

$$\phi(x_i)\phi(x_j)^T \Rightarrow K(x_i, x_j)$$

Why the kernels are successful in classifications

Kernels in Support Vector Machines (SVMs) are functions that **calculate** the similarity **of** pairs of data points in a high-dimensional space. They **allow** SVMs to **discover** complex, non-linear patterns in data by implicitly **relating the** input data **to** a higher-dimensional feature space where the data **can** be linearly **extracted**..

Backend

Here are some most commonly used kernel functions in SVMs:

The linear kernel can be defined as:

$$K(x, y) = x \cdot y$$

inner product

normal dual eq.

Here are some most commonly used kernel functions in SVMs:

One definition of the polynomial kernel is:

Where x and y are the input feature vectors, c is a constant term, and d is the degree of the polynomial, K(x, y) = ~~(x. y + c)d~~. The constant term is added to, and the dot product of the input vectors elevated to the degree of the polynomial.

$(x.y+c)^d$

$\langle x.y \rangle^2 = K(x,y) \qquad 3 \longrightarrow 9$

- more the Non linearity of Kernel fxn ⇒ more is expansion in dimension

SVMs

Here are some most commonly used kernel functions in SVMs:

The Gaussian kernel can be defined as: most used kernel fxn

RBF Kernel

Radial Basis fxn.

K(x, y) = exp(-gamma * ||x - y||^2)

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$$

$\Downarrow$

distance b/w 2 points

$$\left[(x_i^1 - x_j^1)^2 + (x_i^2 - x_j^2)^2 + \cdots + (x_i^D - x_j^D)^2\right]$$

So Kernel fxn try to find Similarity b/w points.

So this Kernel fxn

$$K(x_i, x_j) = e^{-\gamma(\text{distance b/w points})^2}$$

- So Value of Kernel fxn ⇒ large for Close points
- Very small when distance b/w points is large

→ Since the close points are similar to each other ⇒ Kernal Value large

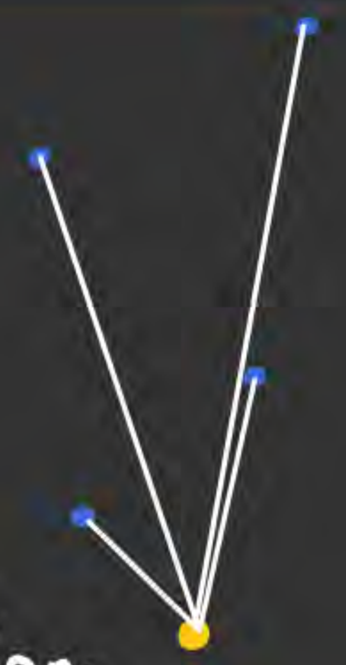
→ points far from each other dissimilar ⇒ Kernel value Small ⇒ 0

So Kernel fxn ⇒ $e^{-\gamma(\text{distance b/w points})^2}$

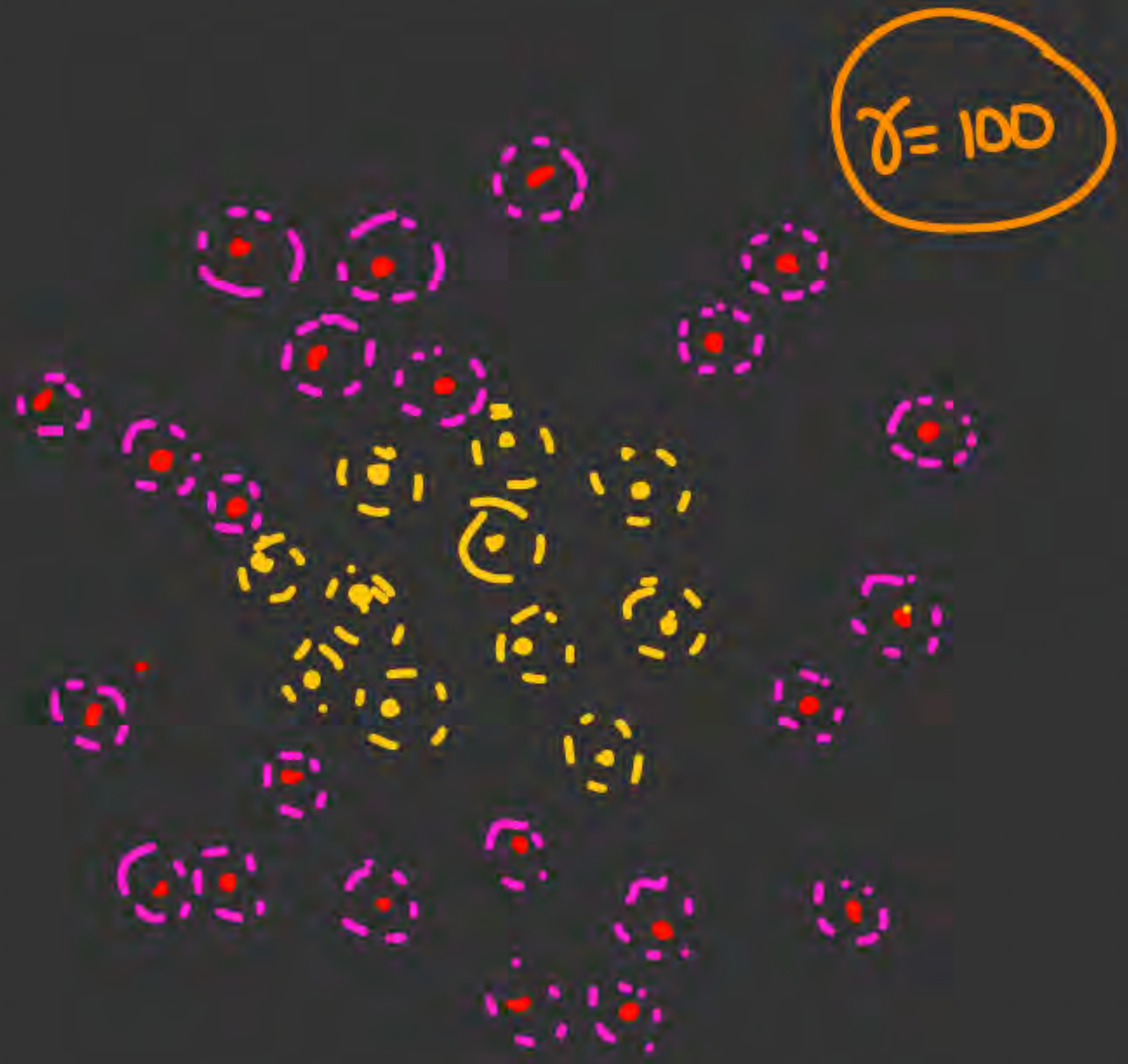
RBF

(γ is a hyper Parameter)

$(\gamma = .01)$, So if γ is small then the neighbourhood of any Point is wide spread.



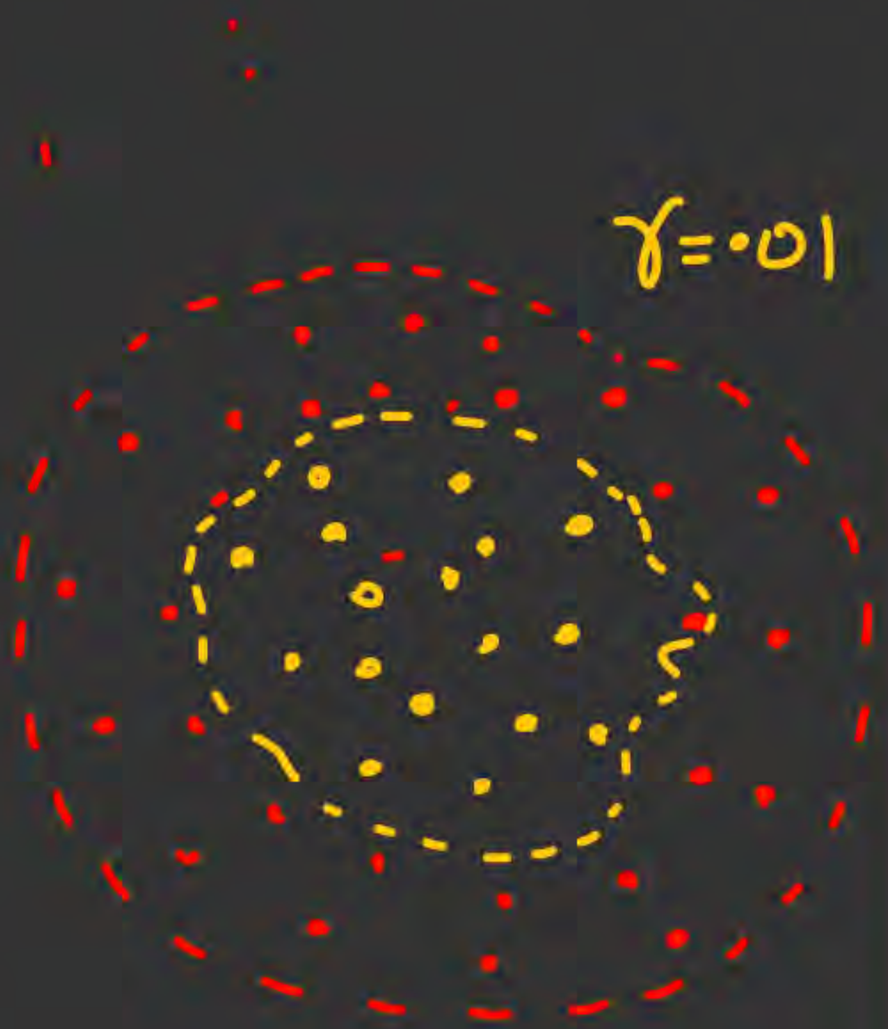
$\gamma = 100$, if γ is large then neighbourhood of any point is compact.



- If γ is v.v. small $\Rightarrow$ underfit

- If γ is large $\Rightarrow$
  - * overfitting
  - * Complexity high
  - * Bias = 0
  - * Variance = high

- If γ is small
  ↳ Perfect fit.



- the order of Non linearity in exp term $e^{-x} = \infty$

$$e^{-x} = \left(1 - x + x^2/2! - \frac{x^3}{3!} - - - -\right)$$

- So the RBF Kernel Can Convert the feature space into ∞ dimension Space.

Here are some most commonly used kernel functions in SVMs:

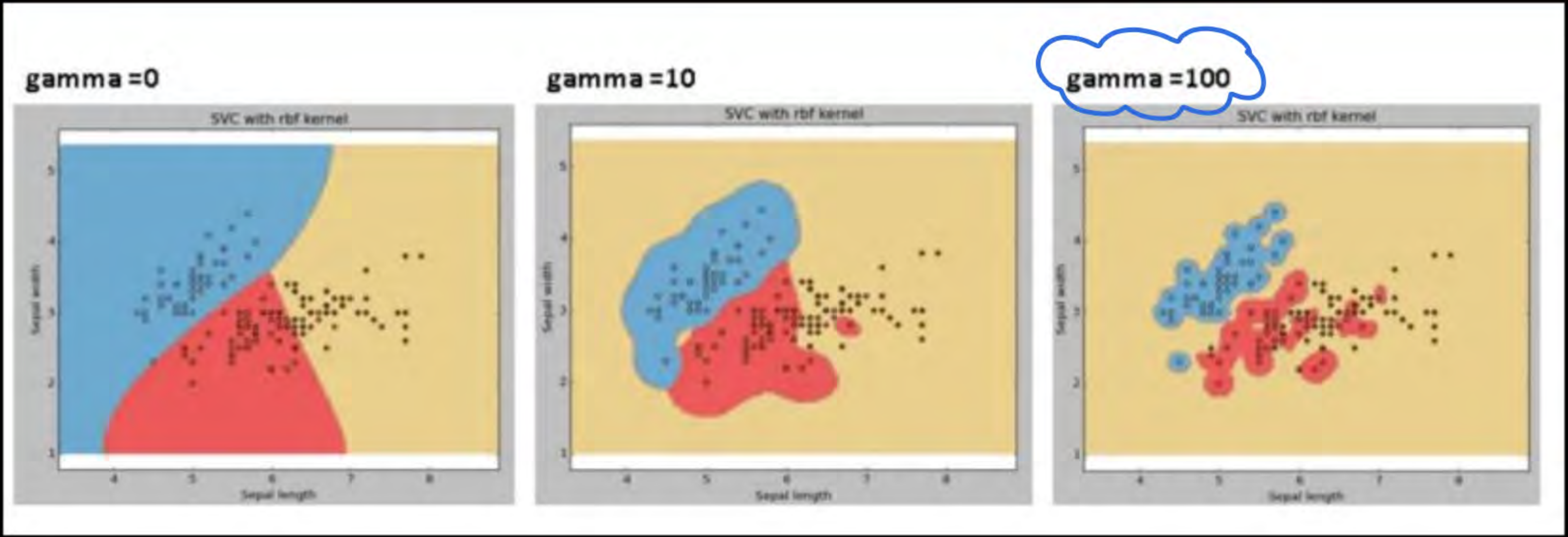
The Laplacian kernel can be defined as:

K(x, y) = exp(-gamma * ||x - y||)

⇒ This is also similar to RBF.
⇒ Non linearity = ∞ order.

$$K(x,y) = e^{-\gamma ||x-y||}$$

RBF or Gaussian Kernel



RBF or Gaussian Kernel

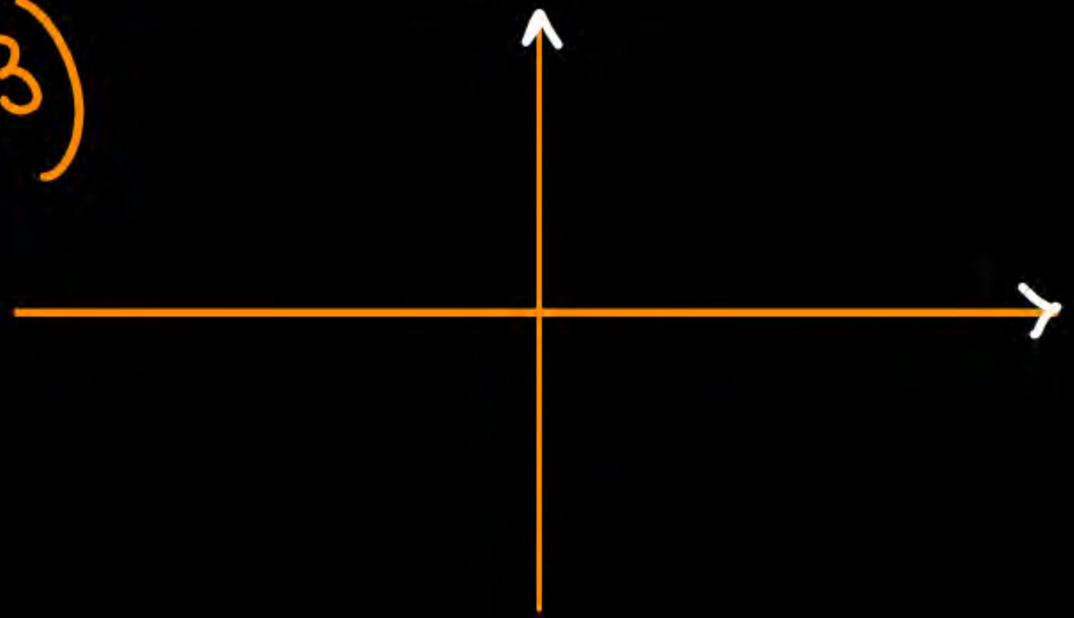
γ importance ✓

Sigmoid Kernel

It is defined as

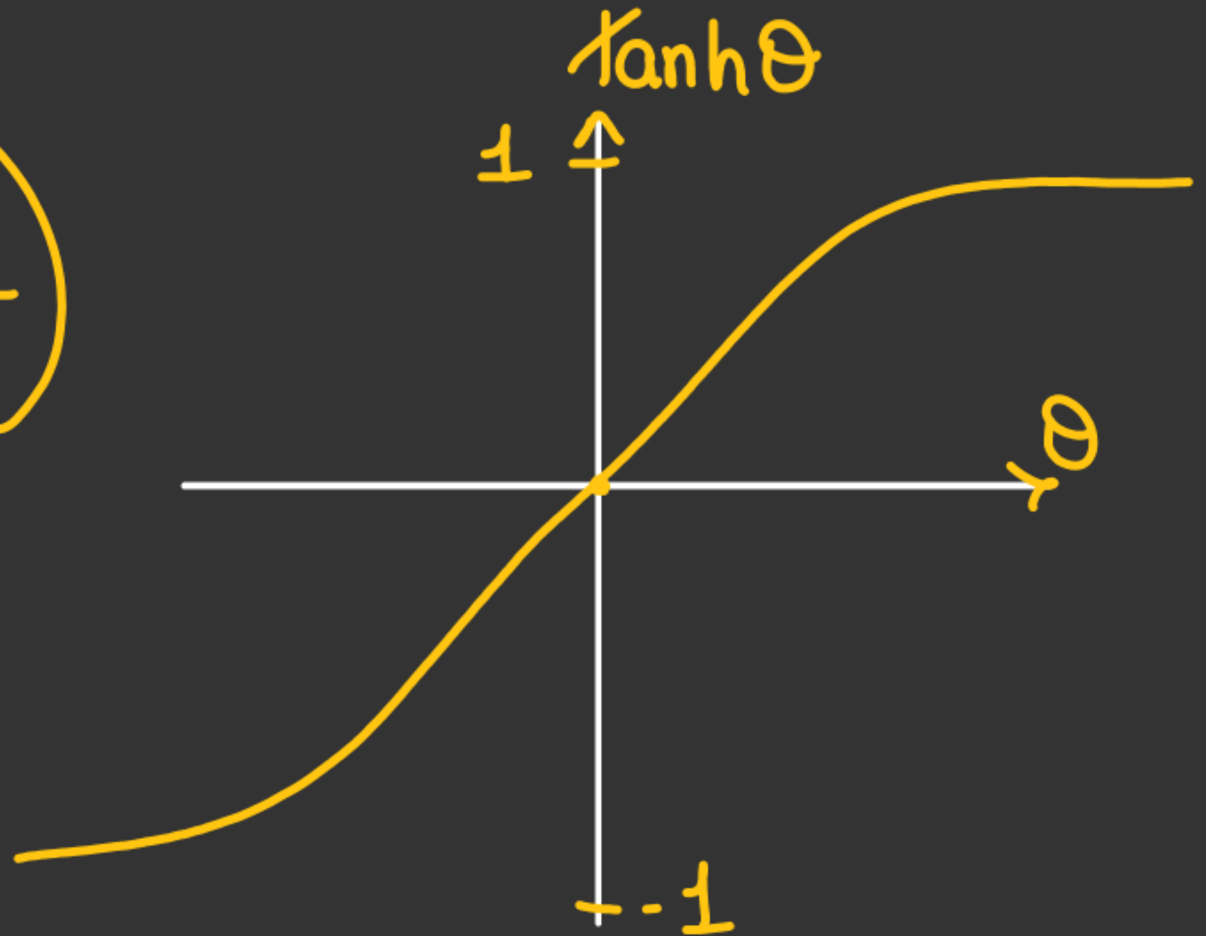
$K(x, y) = tanh(alpha * x * y + beta)$, *where x and y are the input vectors, alpha and beta are parameters, and tanh is the hyperbolic tangent function.*

$$K(x,y) = \tanh(\alpha(x \cdot y) + \beta)$$



$$\tanh(\theta) = \left\{ \frac{e^{\theta} - e^{-\theta}}{e^{\theta} + e^{-\theta}} \right\} \Rightarrow \left(\frac{e^{2\theta} - 1}{e^{2\theta} + 1} \right)$$

$$\Rightarrow \quad K(x, y) = \tanh(\alpha x \cdot y + \beta)$$



Linear kernels:

- Suitable for high-dimensional data or linearly separable data.
- Computes the dot product of input vectors, efficient for large feature sets.
- Simple and often used as a baseline for comparison.

RBF kernel:

- Default choice for non-linear problems in SVMs.
- Captures complex relationships without prior knowledge of data.
- Sensitive to hyperparameter tuning, especially gamma.

not much imp, only read it.

Polynomial kernels:

- Effective for problems with polynomial patterns.
- Commonly used in computer vision and image recognition.
- Degree parameter controls the complexity of the polynomial.

Sigmoid kernel:

- Useful for neural network applications.
- Appropriate when data distribution resembles a sigmoid.
- Requires careful tuning of parameters for best performance.

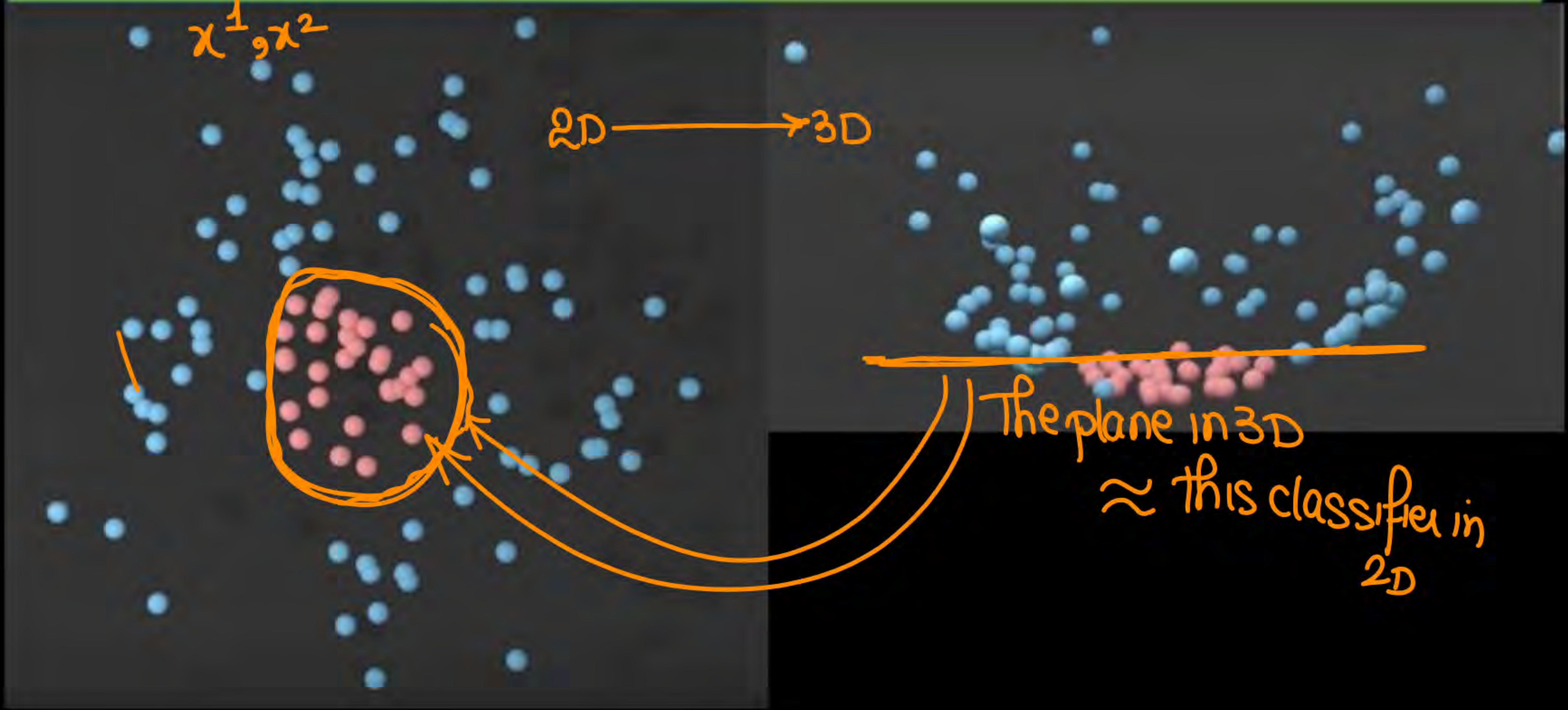
How the kernels do the transformation

Mathematical definition: $K(x, y) = <f(x), f(y)>$. *Here K is the kernel function, x, y are n dimensional inputs. f is a map from n-dimension to m-dimension space.* $< x,y>$ *denotes the dot product. usually m is much larger than n.*

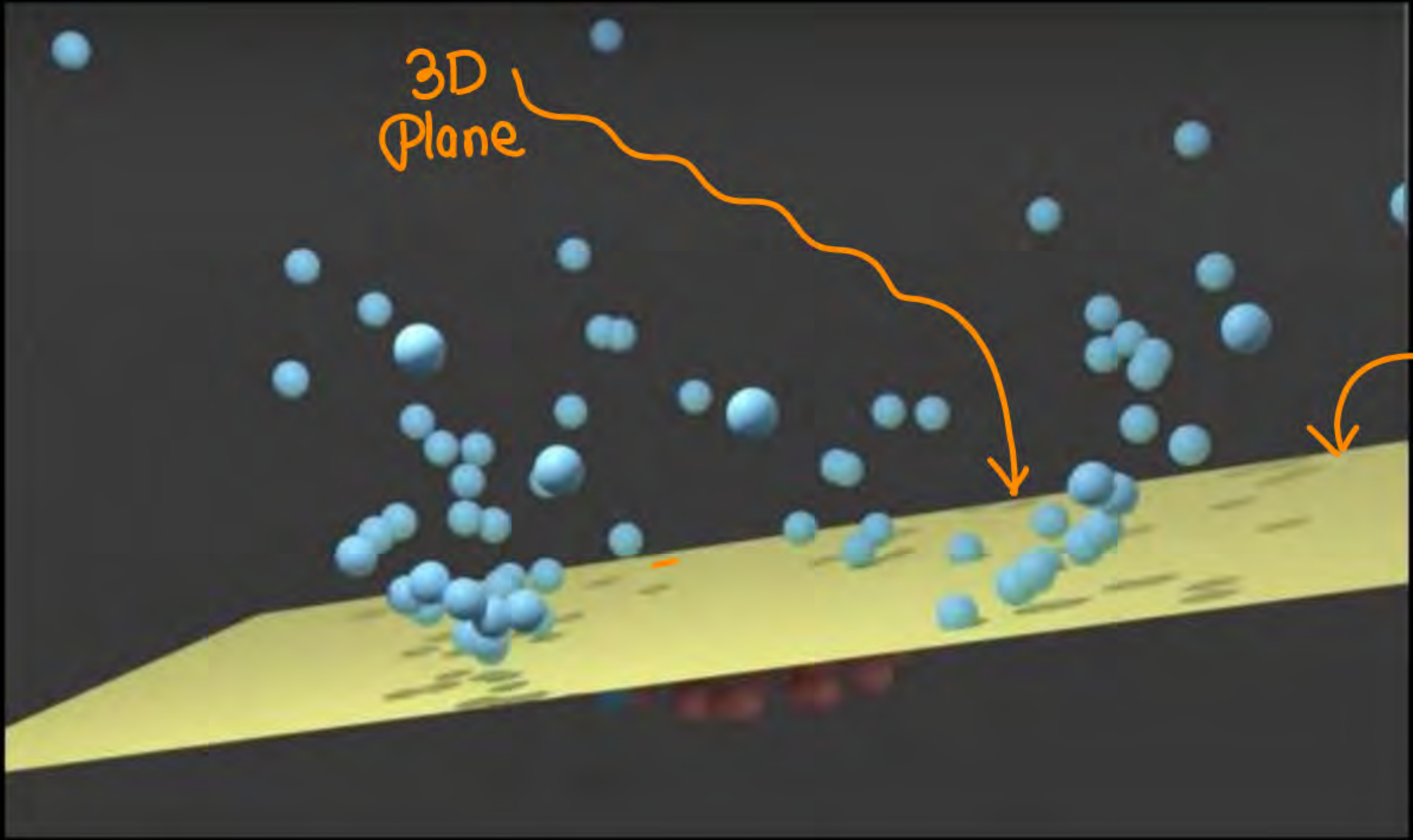
SVMs

D dimension $(wx + b)$ → D Parameters

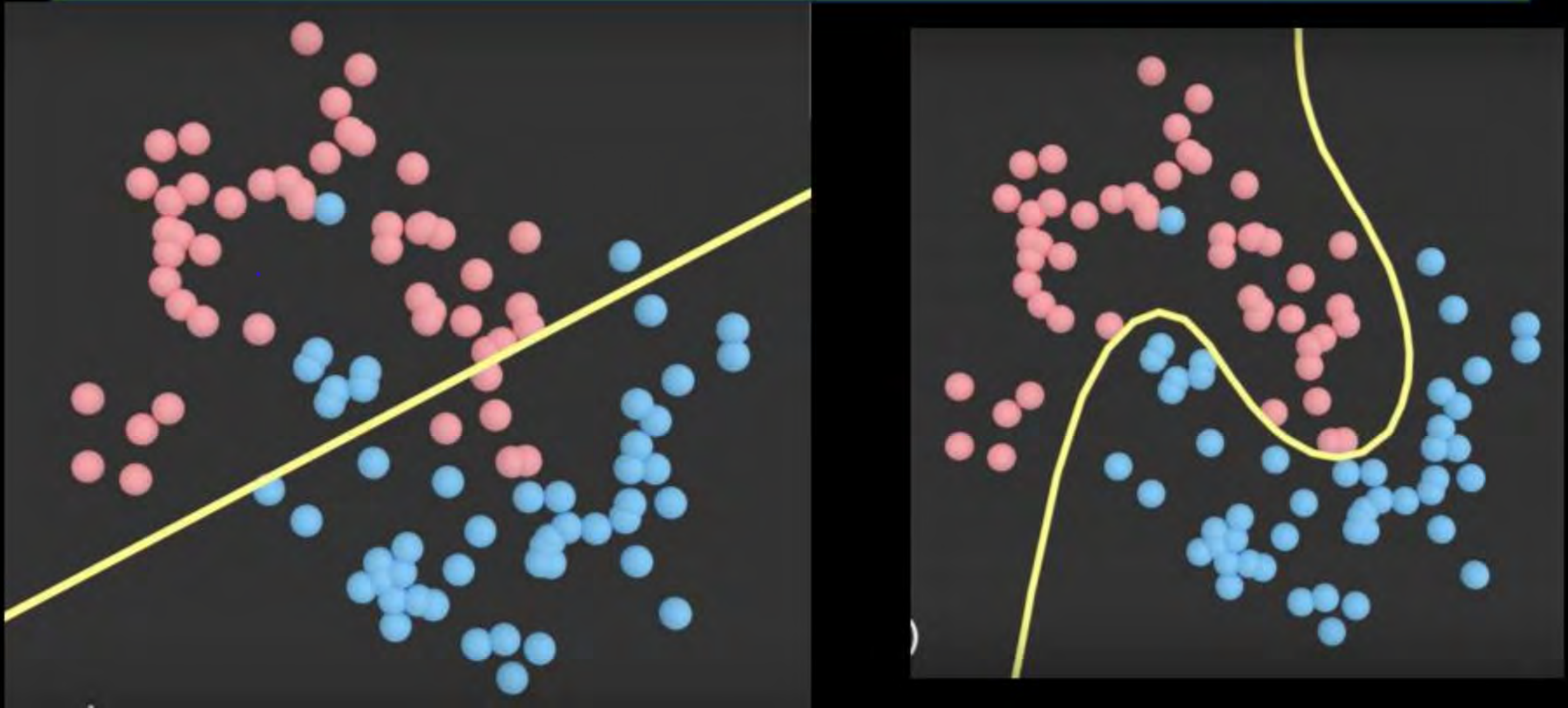
Kernels in SVMs



Kernels in SVMs



Linear Kernel



Mercers Theorem

Mercer's theorem

Only those fxn can be used as kernels if they satisfy →

A function KK is a kernel function if it satisfies two conditions:

$K(x,y) = K(y,x)$

- **Condition 1** - The function KK must be symmetric.
- **Condition 2** - The function must be positive semi-definite.

we have 10 datapoints

$\Rightarrow$ we have 10x10 matrix

$\Rightarrow$ So kernel matrix shd be positive semi definite $\Rightarrow$

$\Rightarrow$ i.e det of matrix ≥ 0
eigen value of matrix ≥ 0

$$\begin{bmatrix} K(x_1, x_1) & K(x_1, x_2) & \cdots & K(x_1, x_{10}) \\ K(x_2, x_1) & K(x_2, x_2) & \cdots & K(x_2, x_{10}) \\ \vdots & & & \\ K(x_{10}, x_1) & \cdots & & K(x_{10}, x_{10}) \end{bmatrix}$$

The Kernel trick will always work ??

NO, this wont work on noisy data...

Practice

uestion: You are training a linear SVM classifier with a binary classification problem. The SVM decision boundary is defined as 3x - 2y - 4 = 0. You want to classify a new data point with coordinates (5, 7). What is the signed distance of the data point to the decision boundary?

$3x - 2y - 4 = 0$

$$\frac{3(5) - 2(7) - 4}{\sqrt{3^2 + 2^2}} \Rightarrow \left(\frac{15 - 14 - 4}{\sqrt{9+4}}\right) \Rightarrow \left(-3/\sqrt{13}\right)$$

Practice

Question: You are training a support vector machine with a polynomial kernel. The kernel function is defined as K(x, y) = (x · y + 1)^2. You want to calculate K(3, 4). What is the value of K(3, 4)?

Single dimension

$= (3 \cdot 4 + 1)^2$

$\Rightarrow (13)^2$

$\Rightarrow 169$

a) 25

b) 121

c) 144

d) 169

Practice

Question: You are using an RBF kernel in an SVM. The width parameter (gamma) is set to 0.1. What is the effect of increasing gamma on the SVM decision boundary?

a) It results in a more flexible (complex) decision boundary. γ inc

b) It results in a less flexible (simpler) decision boundary.

c) It does not affect the decision boundary.

d) The effect on the decision boundary depends on the value of C.

Practice

Question: What is the primary objective of a Support Vector Machine (SVM)?

a) Minimize the number of support vectors.

b) Maximize the margin between classes.

c) Minimize the number of features.

d) Maximize the number of support vectors.

SVMs

Practice

Question: In SVM, what is the role of a kernel function?

a) It determines the regularization parameter.

b) It transforms data into a higher-dimensional space.

c) It computes the margin between classes.

d) It minimizes the number of support vectors.

Practice

Question: In a binary SVM classification, how are data points represented on the correct side of the decision boundary (positive class) with respect to the margin?

Bogus
Bekare

a) Support vectors

b) Misclassified points

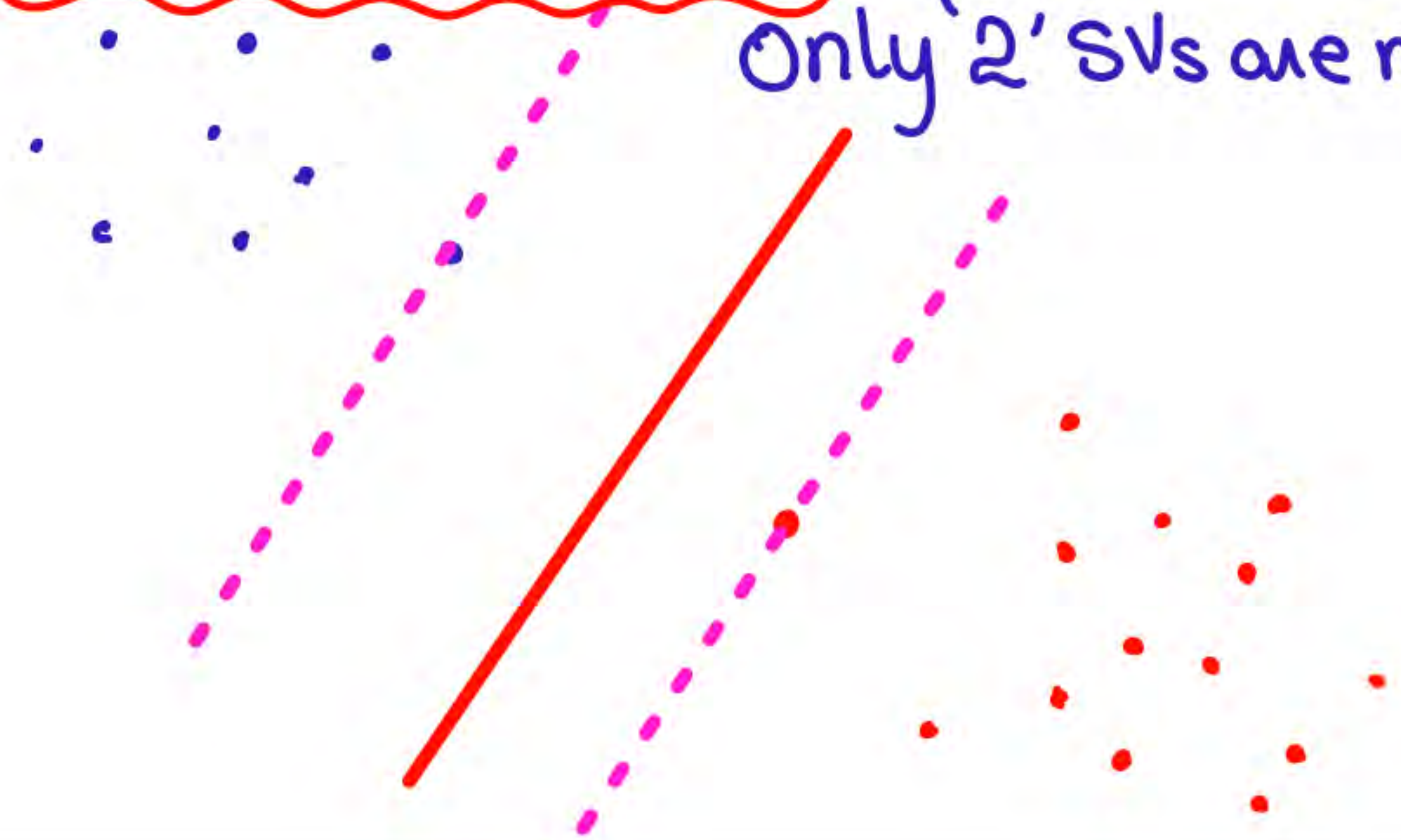
c) Negative values

d) Positive values

Practice

Question: If you have a linearly separable dataset with 100 data points, how many support vectors will an ideal SVM model have?

Only '2' SVs are needed.

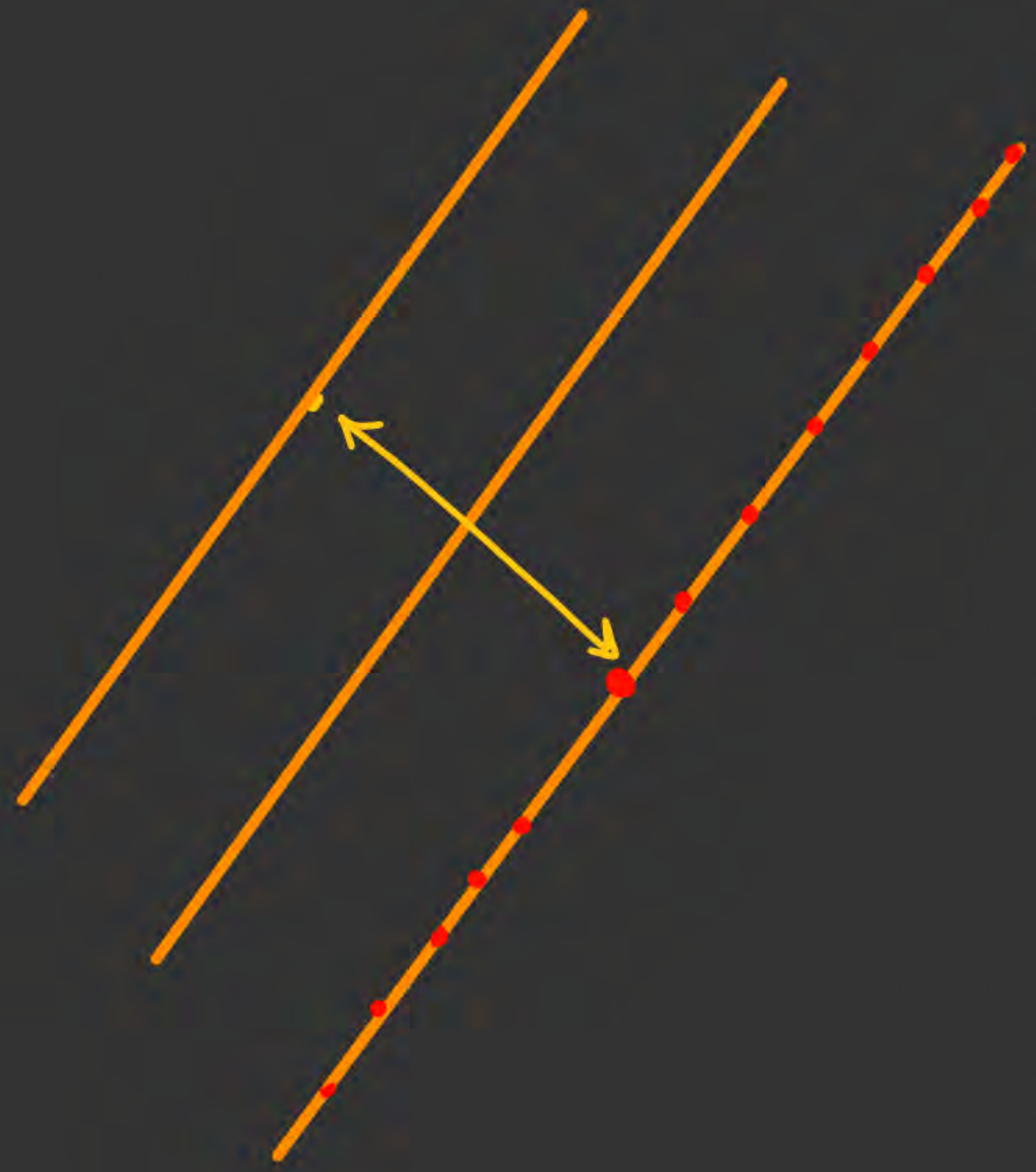


a) 100

b) 50

c) 10

d) 2



Practice

Question: In SVM, the term "support vectors" refers to:

~~a)~~ Data points used to train the model. ← training data

~~b)~~ Data points located far from the decision boundary.

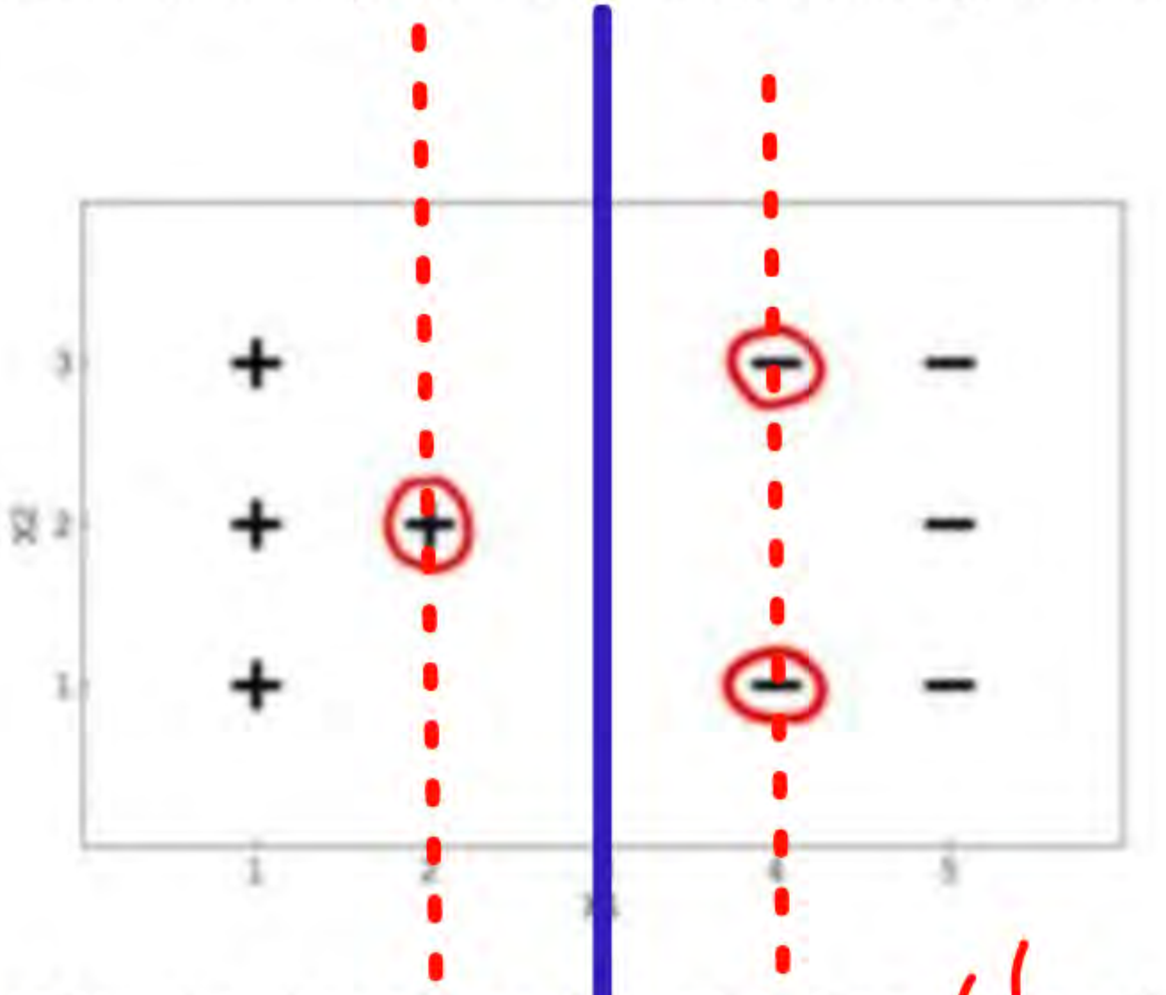
~~c)~~ Data points on the correct side of the decision boundary.

d) Data points closest to the decision boundary.

SVMs

Practice

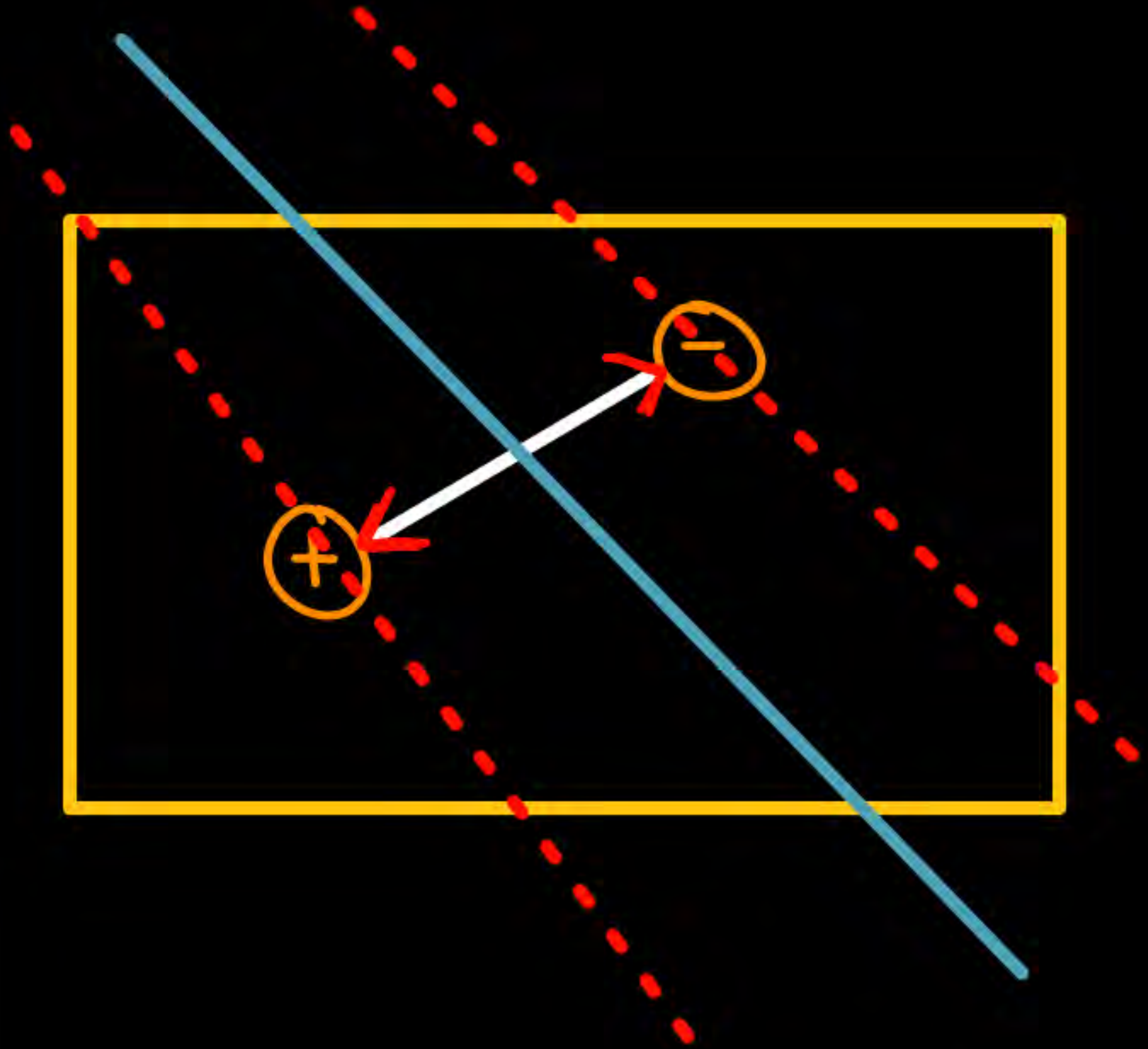
Suppose you are using a Linear SVM classifier with 2 class classification problem. Consider the following data in which the points circled red represent support vectors.

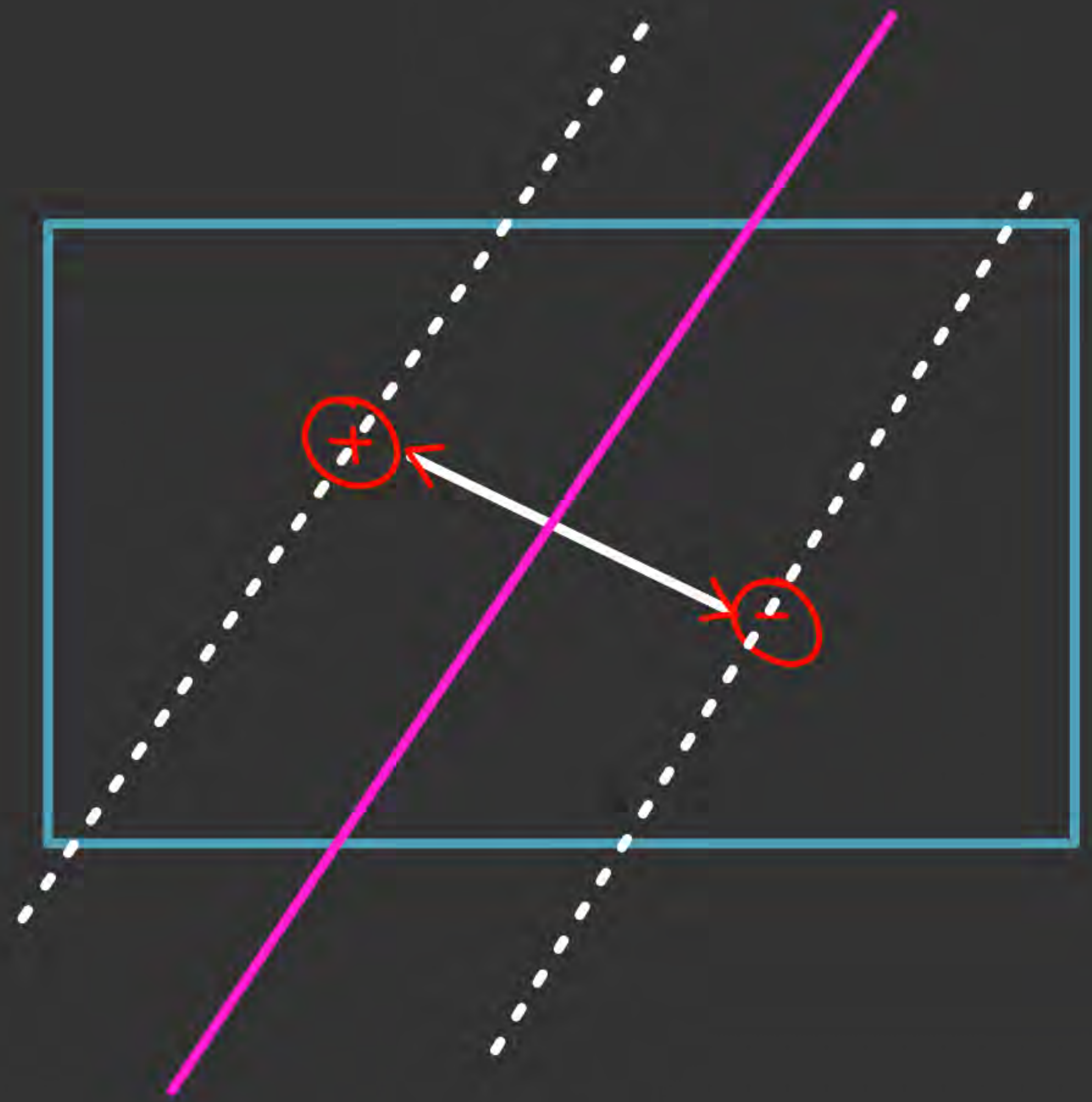


Will the decision boundary change if any of the red points are removed?

a. Yes

b. No





Practice

The soft margin SVM is more preferred than the hard-margin svm when:

1. The data is linearly separable
2. The data is noisy and contains overlapping point

Practice

After training an SVM, we can discard all examples which are not support vectors and can still classify new examples?

a. True
b. False

Practice

In the linearly non-separable case, what effect does the C parameter have on the SVM mode.

a. it determines how many data points lie within the margin
b. it is a count of the number of data points which do not lie on their respective side of the hyperplane
c. it allows us to trade-off the number of misclassified points in the training data and the size of the margin
d. it counts the support vectors

Practice

Q Suppose that we use a RBF kernel with appropriate parameters to perform classification on a particular two class data set where the data is not linearly separable. In this scenario

a. the decision boundary in the transformed feature space is non-linear
b. the decision boundary in the transformed feature space is linear
c. the decision boundary in the original feature space is linear
d. the decision boundary in the original feature space is non-linear

Practice

1) Support vector machine may be termed as:

A. Maximum aprori classifier
B. Maximum margin classifier
C. Minimum apriori classifier
D. Minimum margin classifier

Practice

3) If the hyperplane $W^T X+b=0$ correctly classifies all the training points (X_i, y_i), where $y_i=\{+1, -1\}$, then:

- A. $||W-1|| = 2$
- B. $X= 1$
- C. $W^T X_i+b \geq 0$ for all i
- D. $y_i(W^T X_i+b) \geq 0$ for all i

Practice

2) In a hard margin support vector machine:

- A. No training instances lie inside the margin
- B. All the training instances lie inside the margin
- C. Only few training instances lie inside the margin
- D. None of the above

Practice

4) The constraint in the primal optimization problem solved to obtain the hard margin optimal separating hyperplane is:

- A. $y_i(W^TX_i+b) \geq 1$ for all i
- B. $y_i(W^TX_i+b) \leq 1$ for all i
- C. $(W^TX_i+b) \geq 1$ for all i
- D. $(W^TX_i+b) \leq 1$ for all i

Practice

10) In a hard margin SVM $W^T X+b=0$, suppose X_j's are the support vectors and α_j's the corresponding Lagrange multipliers, then which of the following statements are correct:

- A. $W = \sum a_j y_j X_j$
- B. $\sum a_j y_j = 0$
- C. Either A or B
- D. Both A and B

SVMs

Practice

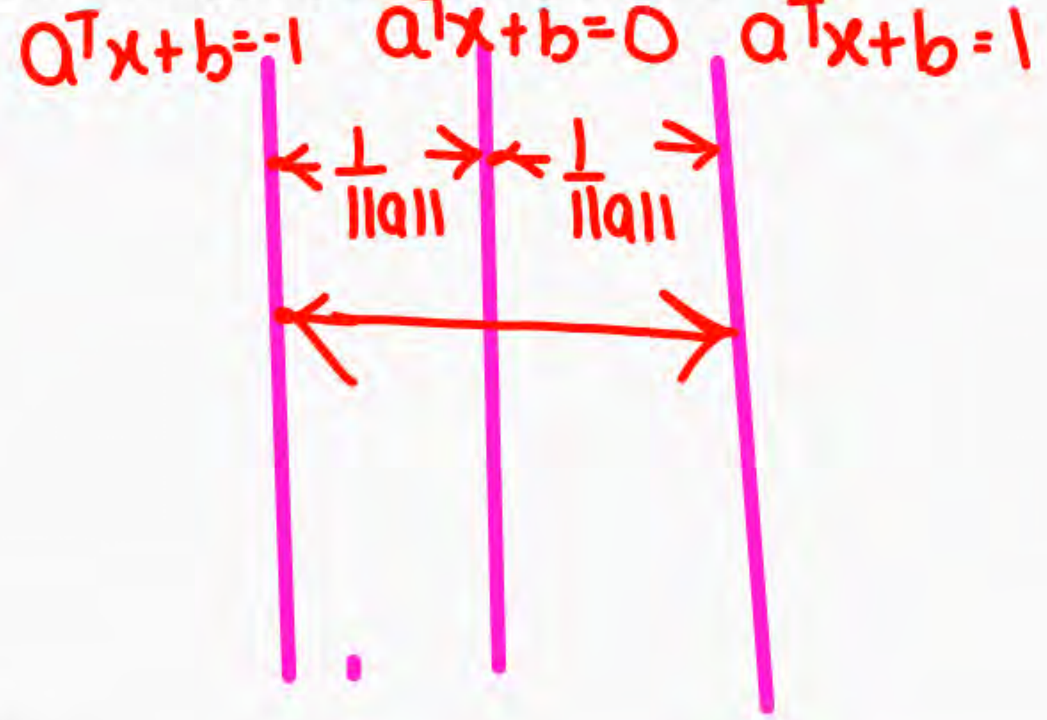
3) In a support vector machine (SVM) for classification of the points $\bar{\mathbf{x}}$, let the hyperplanes be given as

$$\bar{\mathbf{a}}^T\bar{\mathbf{x}} + \bar{\mathbf{b}} \geq 1$$

$$\bar{\mathbf{a}}^T\bar{\mathbf{x}} + \bar{\mathbf{b}} \leq -1$$

The distance between the hyperplanes is given as

- $\frac{2}{\|\bar{\mathbf{a}}\|}$
- $\|\bar{\mathbf{a}}\|$
- $\frac{2b}{\|\bar{\mathbf{a}}\|}$
- $2(b+1)\|\bar{\mathbf{a}}\|$



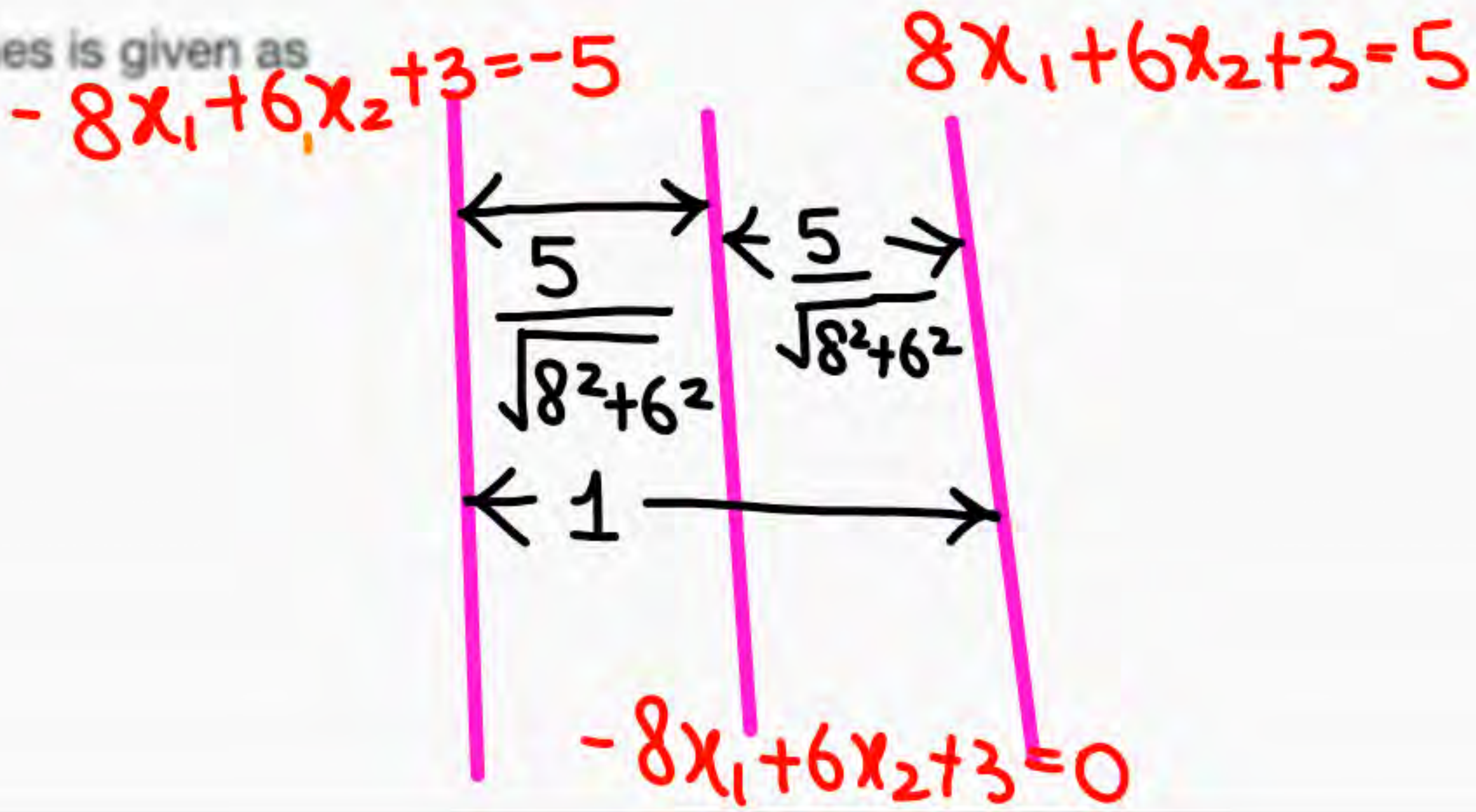
Practice

4) In a support vector machine (SVM) for classification of the points $\bar{\mathbf{x}}$, let the hyperplanes be given as

$$-8x_1 + 6x_2 + 3 \geq 5$$
$$-8x_1 + 6x_2 + 3 \leq -5$$

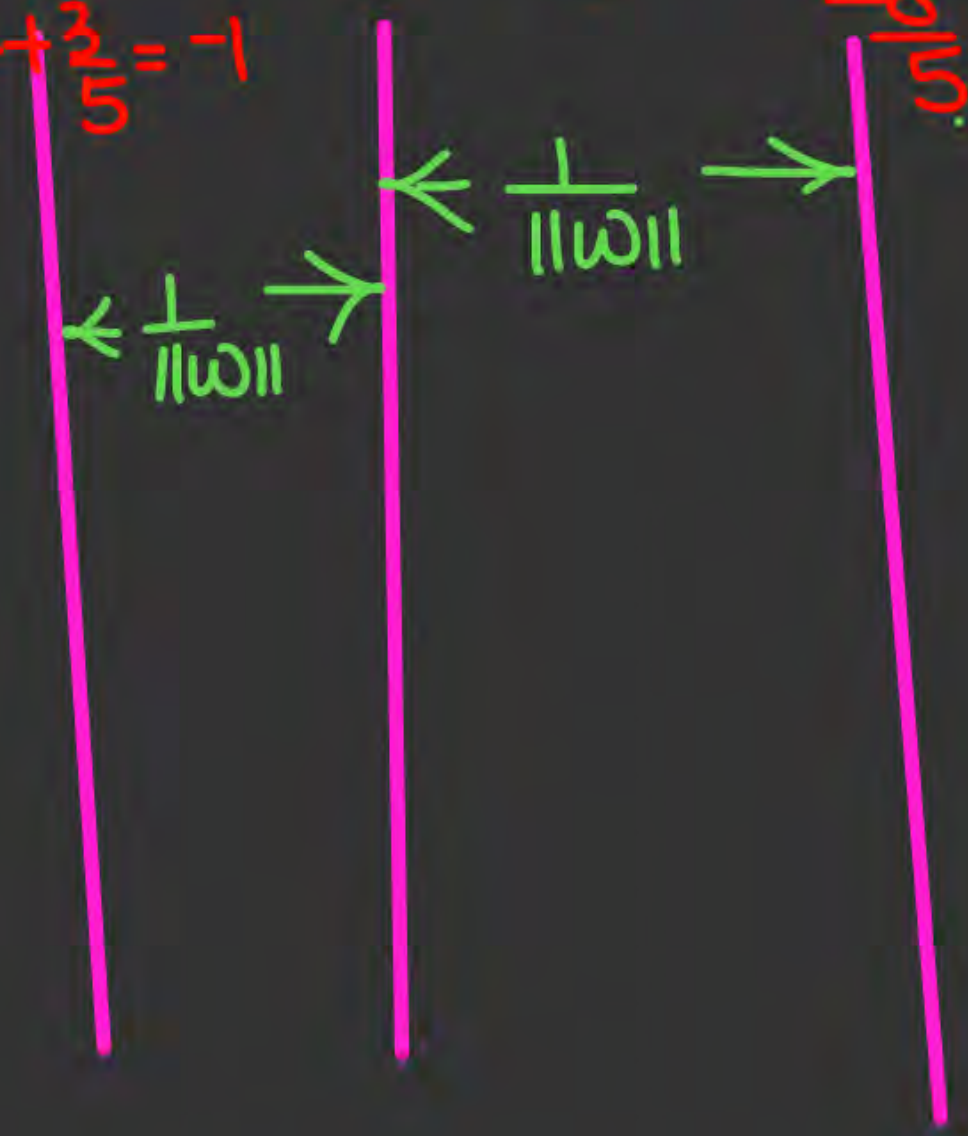
The distance between the hyperplanes is given as

- $\frac{2}{10}$
- 5
- 1 ✓
- 10



$$-\frac{8}{5}x_1 + \frac{6}{5}x_2 + 3/5 = 0$$

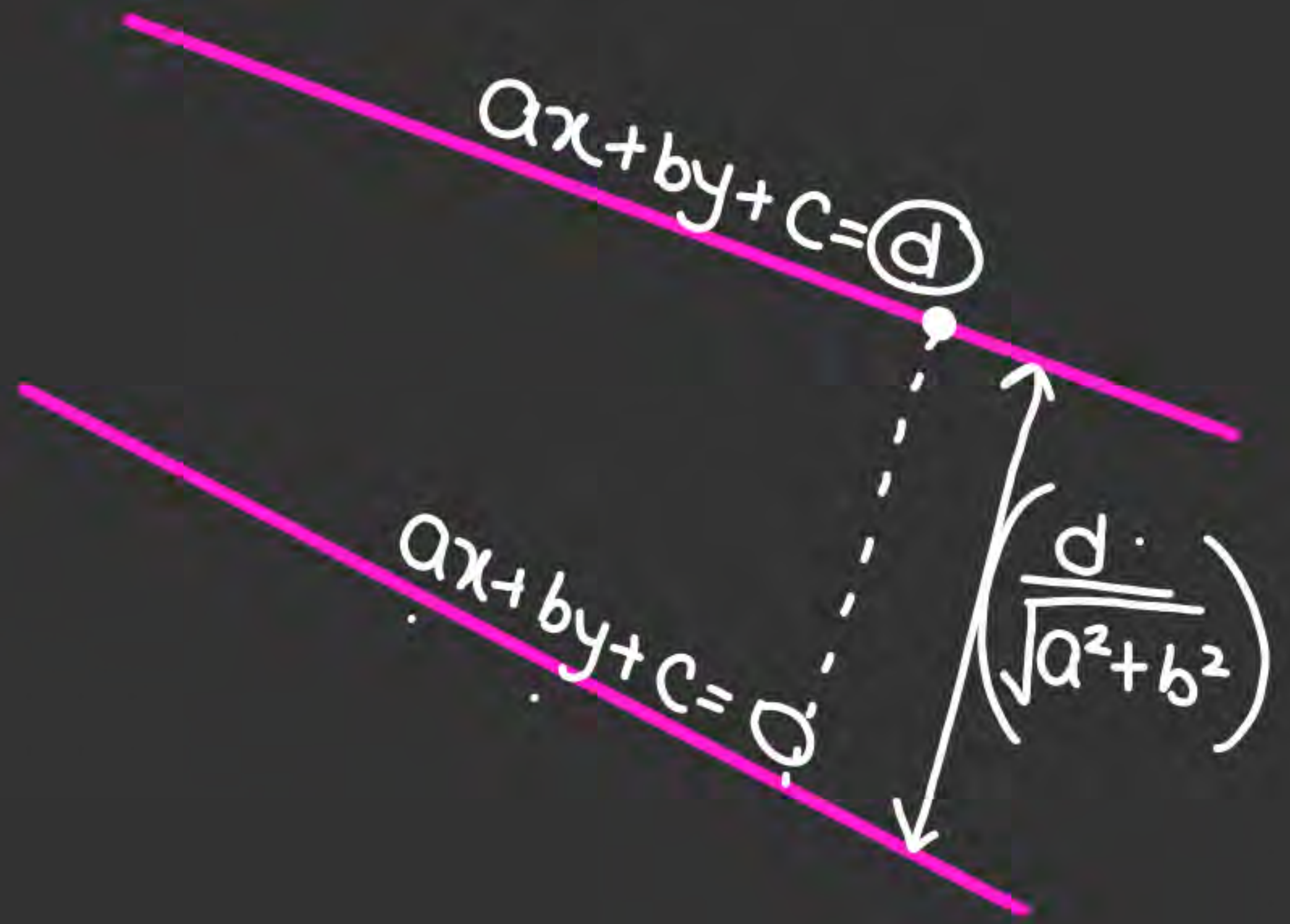
$$-\frac{8}{5}x_1 + \frac{6x_2}{5} + \frac{3}{5} = -1 \qquad -\frac{8}{5}x_1 + \frac{6}{5}x_2 + \frac{3}{5} = 1$$



$$\text{dist} = \frac{2}{\|w\|}$$

$$= \frac{2}{\sqrt{\frac{8^2}{5^2} + \frac{6^2}{5^2}}}$$

$$= \frac{2}{\sqrt{100/25}} = 2/2 = 1$$


$ax+by+c=d$
$ax+by+c=0$
$\left(\frac{d}{\sqrt{a^2+b^2}}\right)$

Practice

SVM is a supervised Machine Learning can be used for Options :

- Regression
- Classification
- both a or b
- None of These

Practice

Closest Point to the hyperplane are support vectors

- True
- False
- Unpredictable
- None of these

Practice

In SVM, the dimension of the hyperplane depends upon which one?

- ○ the number of features
- ○ the number of samples
- ○ the number of target variables
- ○ All of the above

THANK - YOU