Data Science and Artificial Intelligence

Machine Learning

Support Vector Machine

Lecture No. 5

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

- **Topic** Soft margin svm
- **Topic** Svm- advantage disadvantage
- **Topic** Hinge loss in svm
- **Topic** Question
- **Topic**

Turn on Slide map

Topics to be Covered

- Topic — Soft margin svm
- Topic
- Topic
- Topic
- Topic


Nothing will work
unless you do.
Maya Angelou
quotefancy

Kernels

RBF Kernel

$\gamma \Rightarrow$ large $\Rightarrow$ Overfitting

small $\Rightarrow$ Balance fit

$\rightarrow$ hyperparameter

REVISION

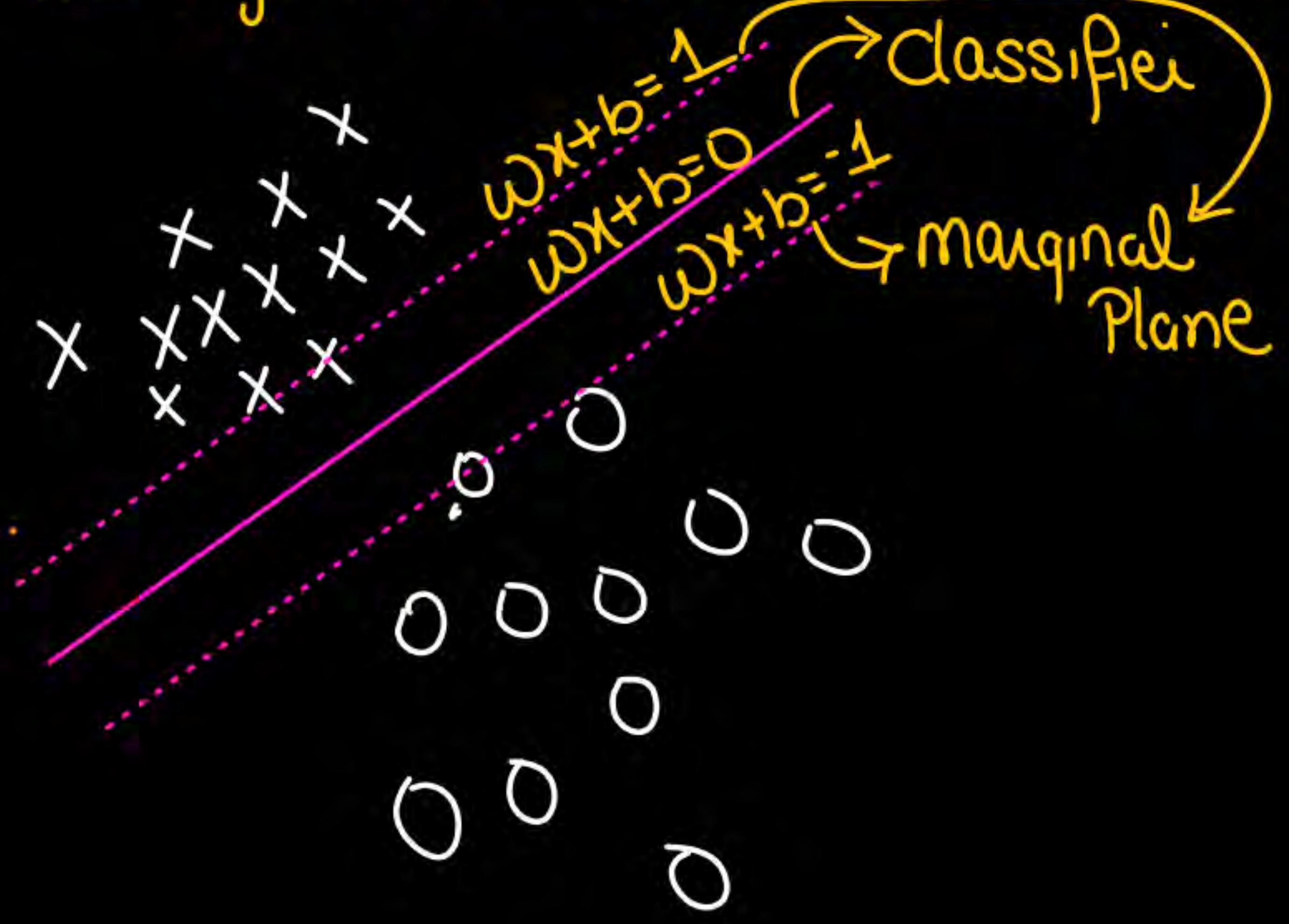
Mercers Theorem

→ K shd be symmetric

→ K matrix shd positive semi definite

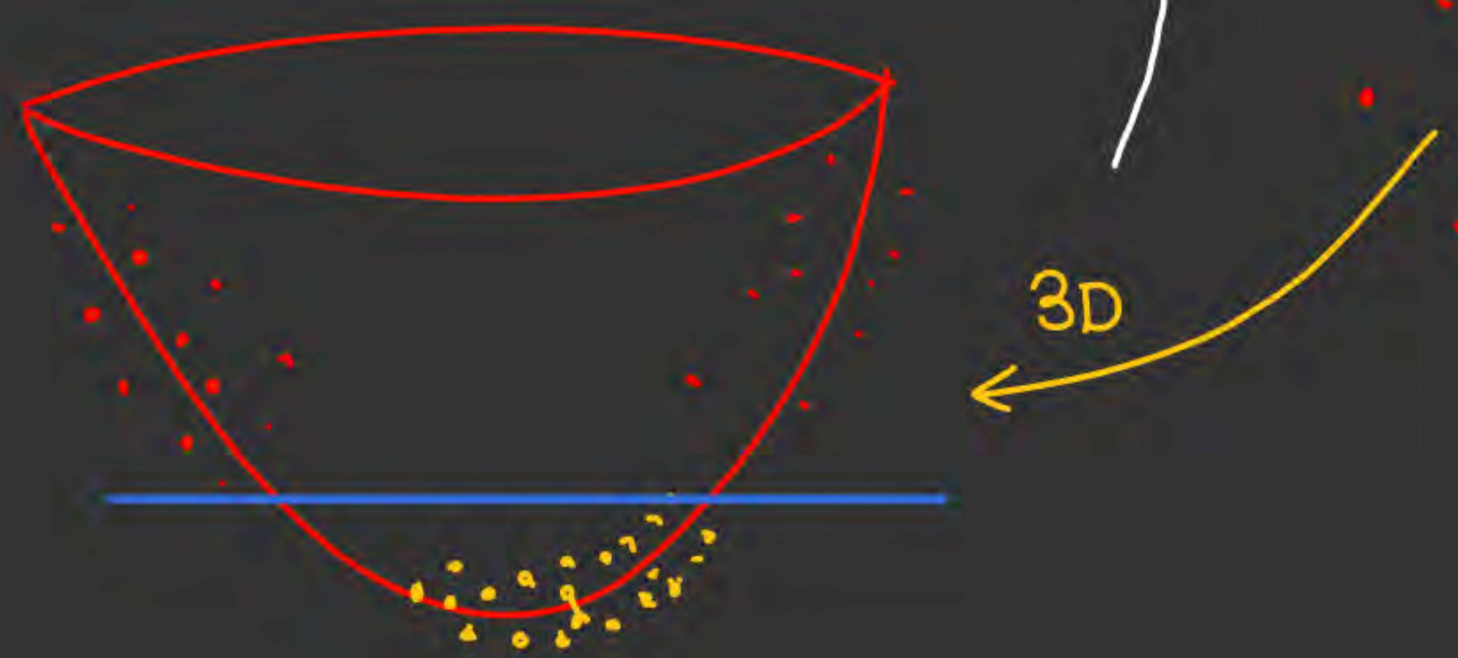
Hard Margin SVM

→ The algo that we used till now was Hard margin SVM.



- So here we assume that data has no noise or overlapping
- But if the data has noise

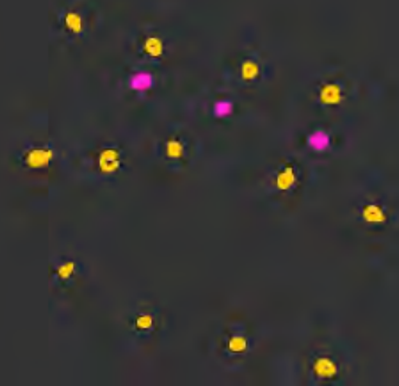
if data has noise i.e
points of 2 classes
are mixed ⇒



$$e^{-\gamma \|distance\|^2}$$

distance is measure
of similarity

noisy data



Kernel working

→ find similar points → map into higher dimension
Keep them close

dissimilar " → far away mapping

⇛ So Kernels cannot remove noise
from data

To handle noisy data we use soft margin svm

So here we use a slack variable

Hard margin svm

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

Constraint untunable fix

all points away from marginal Plane

Soft margin

$y_i(\omega x_i+b) \geq 1-\xi_i$

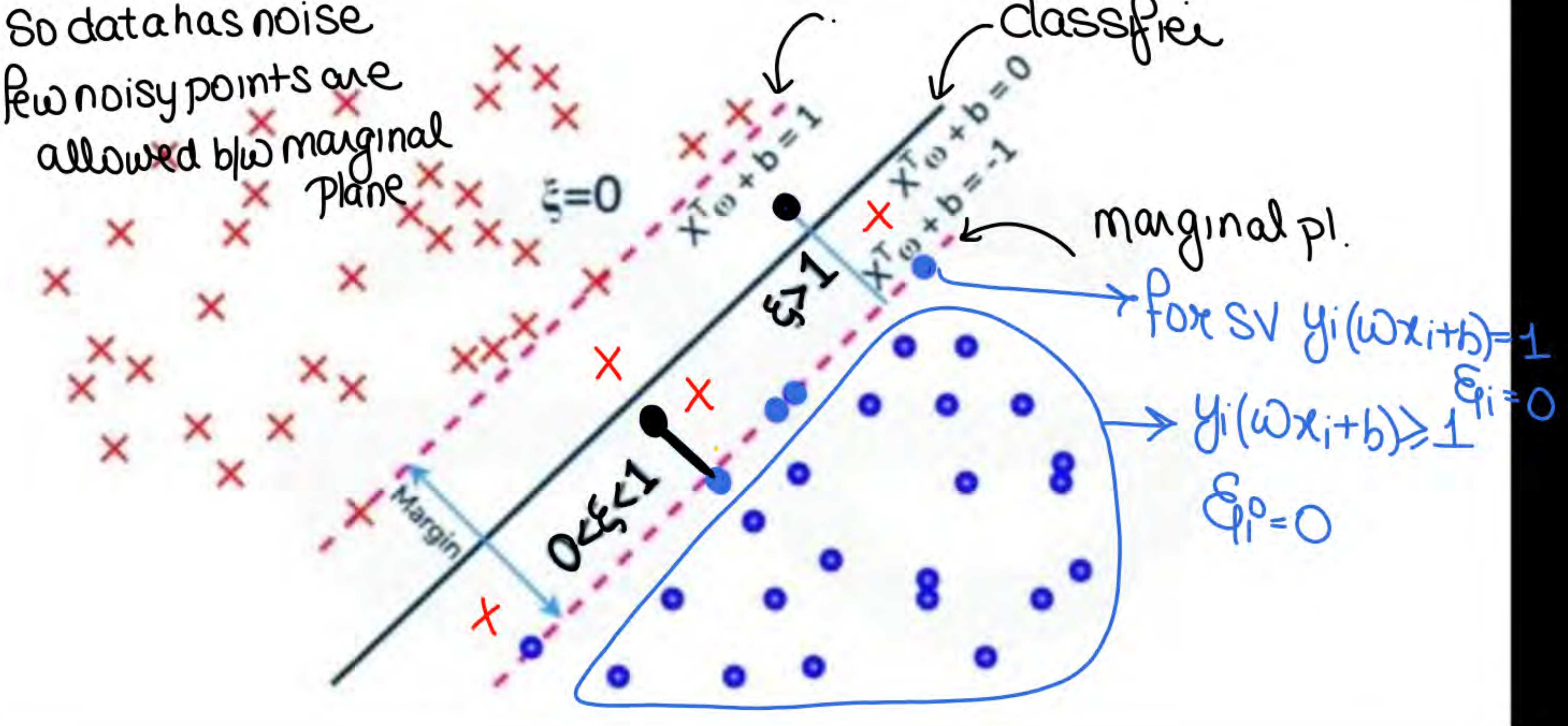
slack variable

$\Rightarrow \xi_i \geq 0$

* ξ_i is distance of point from marginal plane

Soft Margin SVM

- So data has noise
- few noisy points are allowed b/w marginal plane



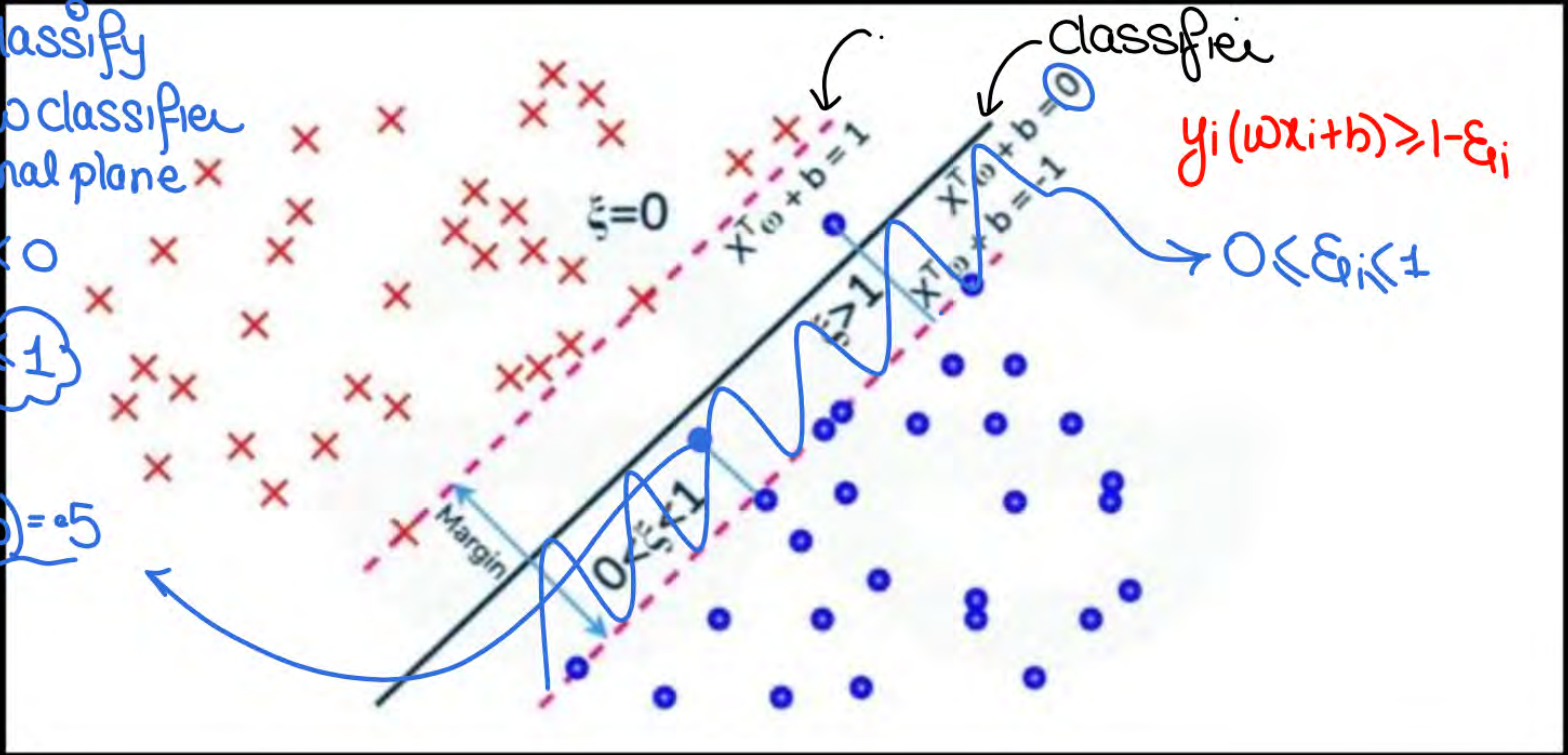
Soft Margin SVM

Correctly classify
but it is b/w classifier
and marginal plane

$-1 < \omega x_i + b < 0$

$y_i(\omega x_i + b) \leq 1$

$y_i(\omega x_i + b) = .5$



$y_i(\omega x_i + b) \geq 1 - \xi_i$

$0 \leq \xi_i \leq 1$

SVMs

Soft Margin SVM

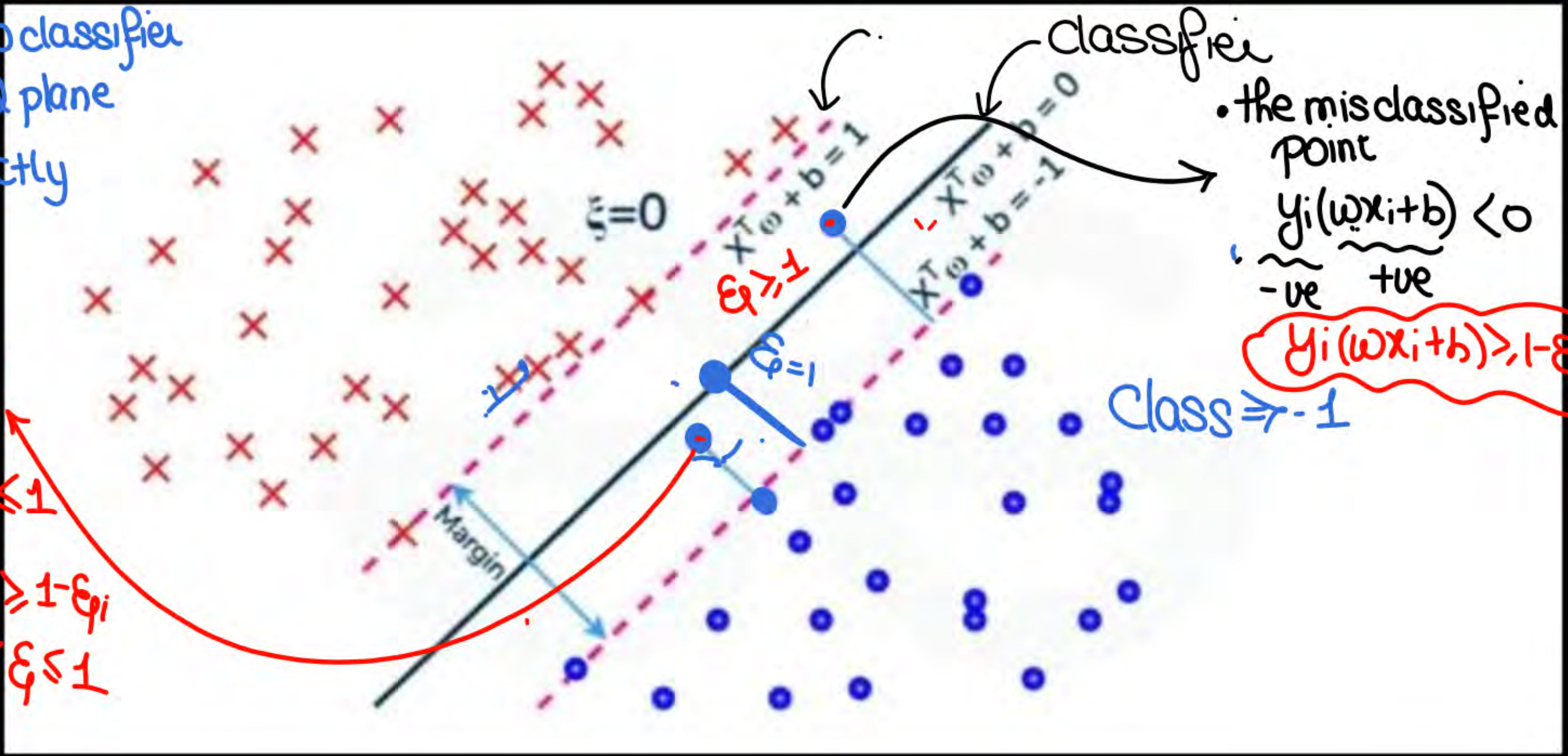
- any point b/w classifier and marginal plane but Correctly Classify

$-1 < (\omega x_i + b) < 0$

$\Rightarrow y_i(\omega x_i + b) \leq 1$

$y_i(\omega x_i + b) \geq 1 - \xi_i$

$0 \leq \xi \leq 1$



Soft Margin SVM

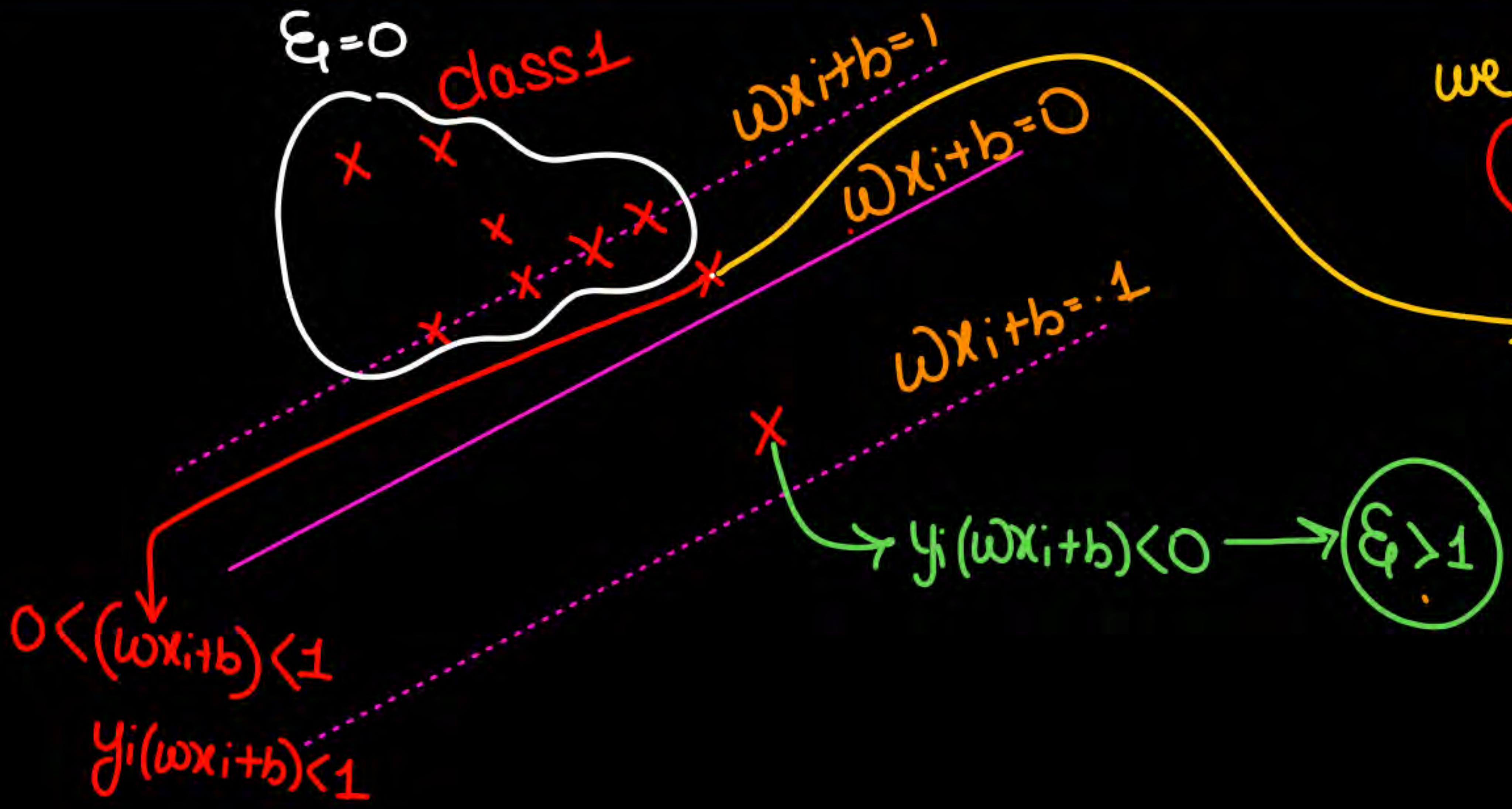
- So the ϵ_i (slack variable)
  - → $\epsilon_i = 0$ SV and points away from marginal plane
  - → $0 < \epsilon_i < 1$ points correctly classify but b/w classifier and marginal plane
  - $\epsilon_i > 1$ " wrongly classified by classifier.

Here we introduce slack variable ϵ_i.

- if confidence score < 1, it means that classifier did not classify the point correctly and incurring a linear penalty of ϵ_i
- If $0<\epsilon_i<1$ it means the point is correctly classified but lies between the hyperplane and margin plane
- If $\epsilon_i>1$ if means the point is on wrong side of hyperplane
- C is a regularization parameter that balances the trade-off between maximizing the margin and minimizing classification errors.

SVMs

Soft Margin SVM



we want

$y_i(wx_i + b) \geq 1 - \epsilon_i$

for all points

$0 < \epsilon_i < 1$

Soft Margin SVM

Problem ⇒ $\min \frac{||\omega||^2}{2} + C\sum_{i=1}^{N} \xi_i$ → So C has to be tuned

C is a hyperparameter

St. $y_i(\omega x_i + b) \geq 1 - \xi_i$
and $\xi_i > 0$

Why we have kept ξ_i's in min
↳ becoz we want to minimize the misclassification

⇒ (C = 0 ⇒ failed algo)

⇒ C = v.v. large ⇒ Similar to Hard margin svm

⇒ C = v.v. small ⇒ ξ can be large misclassification

Skip

So lagrangian $\Rightarrow$

$$\min_{\omega, \varepsilon_i, b} \max_{\mu_i \lambda_i,\ \mu_i, \lambda_i > 0} \overbrace{\frac{1}{2}(\omega \omega^T) + C\sum_{i=1}^{N} \varepsilon_i + \sum_{i=1}^{N} \lambda_i \left(1 - \varepsilon_i - y_i(\omega x_i + b)\right) - \sum_{i=1}^{N} \mu_i \varepsilon_i}^{\mathcal{L}}$$

KKT

1) $\frac{\partial L}{\partial \omega} \Rightarrow \omega - \sum_{i=1}^{N} \lambda_i y_i x_i = 0$, $\underbrace{\omega = \sum \lambda_i y_i x_i}$

2) $\frac{\partial L}{\partial b} \Rightarrow \underbrace{\sum_{i=1}^{N} \lambda_i y_i = 0}$

3) $\frac{\partial L}{\partial \varepsilon_i} \Rightarrow C - \lambda_i - \mu_i = 0$, $\underbrace{(\mu_i + \lambda_i = C)}$

4) $1-\xi_i - y_i(\omega x_i + b) \leq 0$

5) $\lambda_i\left(1-\xi_i - y_i(\omega x_i + b)\right) = 0$

6) $\xi_i \geq 0$

7) $\underline{\mu_i}\ \underline{\xi_i} = 0$

Soft Margin SVM

So for points away from marginal plane

⇒ • $\xi_i = 0$

⇒ • $\lambda_i = 0$

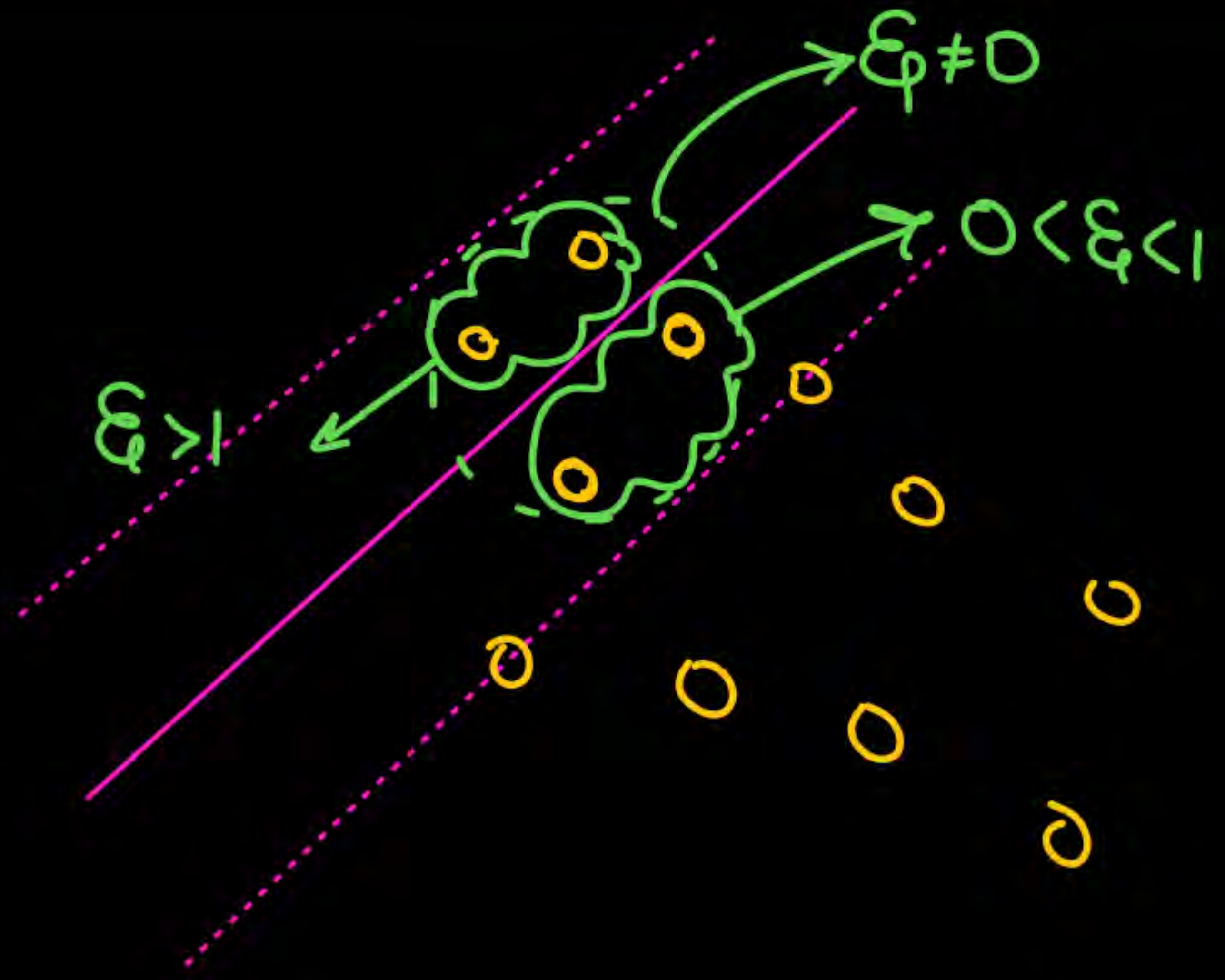
⇒ • $\mu_i = C$ ✓

So for Support Vectors ⇒ $\lambda_i \neq 0$

$\xi_i = 0$

$\mu_i \neq 0$

Soft Margin SVM



$\mu = 0$

$\lambda \neq 0$ bcoz

$\lambda_i \left(1 - \xi_i - \underbrace{y_i(w x_i + b)}_{(1-\xi_i)}\right) = 0$

$\underbrace{\hspace{3em}}_{\text{zero}}$

$\lambda_i \neq 0$

$\Rightarrow \lambda_i = C$

- For points far from marginal plane

$$\left\{\begin{array}{l} \varepsilon_i = 0 \\ \lambda_i = 0 \\ \mu_i = C \end{array}\right.$$

Correctly Classified

- Points which are b/w marginal plane, and misclassified points

$\lambda_i \neq 0$

$\varepsilon_i \neq 0$

$\mu_i = 0$

$\lambda_i = C$

- SV'S $\lambda_i \neq 0$, $\varepsilon_i = 0$

$\mu_i \neq 0$

$\mu_i + \lambda = C$

SVM for Regression

What is the cost function
In SV classification...

In support vector
classification we have
created a division line ...

Advantages of svm

non linear data can be classified

Outlier donot effect working

Robust to noise but data shd not be completely overlapping

Advantages of SVM

- Handling high-dimensional data: SVMs are effective in handling high-dimensional data, which is common in many applications such as image and text classification.
- Handling small datasets: SVMs can perform well with small datasets, as they only require a small number of support vectors to define the boundary.
- Modeling non-linear decision boundaries: SVMs can model non-linear decision boundaries by using the kernel trick, which maps the data into a higher-dimensional space where the data becomes linearly separable.
- Robustness to noise: SVMs are robust to noise in the data, as the decision boundary is determined by the support vectors, which are the closest data points to the boundary.

Advantages of SVM

- Sparse solution: SVMs have sparse solutions, which means that they only use a subset of the training data to make predictions. This makes the algorithm more efficient and less prone to overfitting.
- Regularization: SVMs can be regularized, which means that the algorithm can be modified to avoid overfitting.

Disadvantages of SVM

- Computationally expensive: SVMs can be computationally expensive for large datasets, as the algorithm requires solving a quadratic optimization problem.
- Choice of kernel: The choice of kernel can greatly affect the performance of an SVM, and it can be difficult to determine the best kernel for a given dataset.
- Sensitivity to the choice of parameters: SVMs can be sensitive to the choice of parameters, such as the regularization parameter, and it can be difficult to determine the optimal parameter values for a given dataset.
- Memory-intensive: SVMs can be memory-intensive, as the algorithm requires storing the kernel matrix, which can be large for large datasets.
- Limited to two-class problems: SVMs are primarily used for two-class problems, although multi-class problems can be solved by using one-versus-one or one-versus-all strategies.

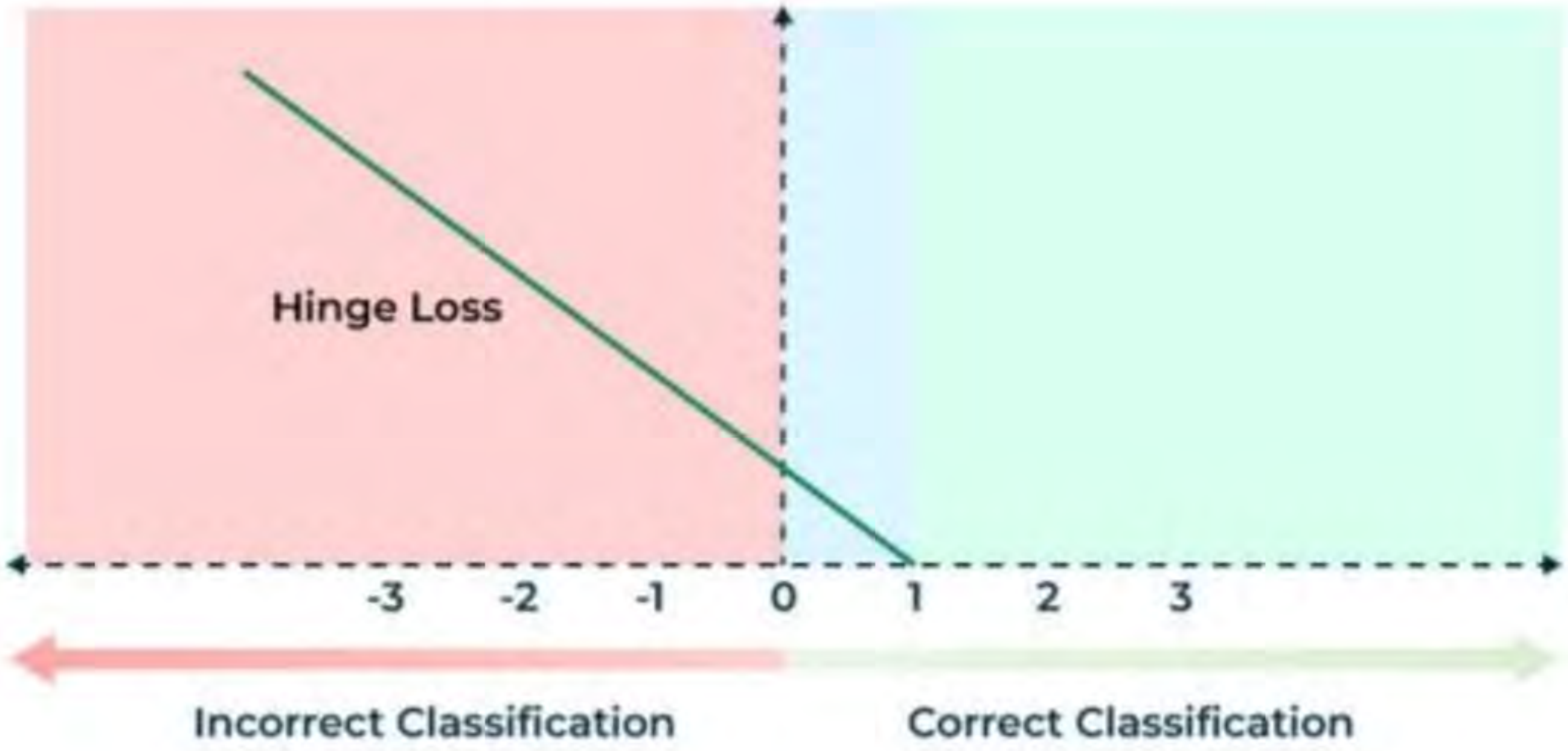
Disadvantages of SVM

- Not suitable for large datasets with many features: SVMs can be very slow and can consume a lot of memory when the dataset has many features.
- Not suitable for datasets with missing values: SVMs requires complete datasets, with no missing values, it can not handle missing values.

Hinge Loss in SVMs

Mathematically, Hinge loss for a data point can be represented as :

$$L(y, f(x)) = max(0, 1 - y * f(x))$$



Hinge Loss

Hinge Loss in SVMs

- If we look at the mathematical formulation the ***hinge loss is effectively present in the constraints*** of a hard margin. This ensures that the decision boundary (the hyperplane) is positioned in such a way that it maximizes the margin without allowing any data points to be within or on the wrong side of the margin.
- Here the hinge loss component, is part of the objective function itself through slack variable.

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

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

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

Practice

Choose the correct option regarding classification using SVM for two classes

Statement i : While designing an SVM for two classes, the equation $y_i(a^t x_i + b) \geq 1$ is used to choose the separating plane using the training vectors.

Statement ii : During inference, for an unknown vector x_j , if $y_j\left(a^t x_j + b\right) \geq 0$, then the vector can be assigned class 1.

Statement iii : During inference, for an unknown vector x_j , if $\left(a^t x_j + b\right) > 0$, then the vector can be assigned class 1.

a. Only Statement i is true
b. Both Statements i and iii are true
c. Both Statements i and ii are true
d. Both Statements ii and iii are true

Practice

QUESTION 7:

Suppose we have the below set of points with their respective classes as shown in the table. Answer the following question based on the table.

X	Y	Class Label
1	0	+1
-1	0	-1
2	1	+1
-1	-1	-1
2	0	+1

What will happen to maximum margin if we remove the point (-1,0) from the training set?

a. Maximum margin will decrease
b. Maximum margin will increase
c. Maximum margin will remain same
d. Can not decide

Practice

Suppose we have the below set of points with their respective classes as shown in the table. Answer the following question based on the table.

X	Y	Class Label
1	0	+1
-1	0	-1
2	1	+1
-1	-1	-1
2	0	+1

What can be a possible decision boundary of the SVM for the given points?

a. $y = 0$
b. $x = 0$
c. $x = y$
d. $x + y = 1$

Practice

Suppose we have the below set of points with their respective classes as shown in the table. Answer the following question based on the table.

X	Y	Class Label
1	0	+1
-1	0	-1
2	1	+1
-1	-1	-1
2	0	+1

Find the decision boundary of the SVM trained on these points and choose which of the following statements are true based on the decision boundary.

i) The point (-1,-2) is classified as -1

ii) The point (1,-2) is classified as -1

iii) The point (-1,-2) is classified as +1

iv) The point (1,-2) is classified as +1

Practice

Which one of the following is a valid representation of hinge loss (of margin = 1) for a two-class problem?

y = class label (+1 or -1).

p = predicted (not normalized to denote any probability) value for a class.?

a. $L(y, p) = \max(0, 1 - yp)$
b. $L(y, p) = \min(0, 1 - yp)$
c. $L(y, p) = \max(0, 1 + yp)$
d. None of the above

(MCQ)

#Q. Consider the problem of finding an optimal hyperplane for non-seperable patterns, we introduce a new set of variables, $\{\xi_i\}_{i=1}^{N}$ into the definition of the 2 points separating hyperplane as $d_i(w^T x_i + b) > 1 - \xi_i$. Choose the correct statements from the options given below.

A The slack variable ξ_i can take both positive and negative values.

B For $0 < \xi_i \leq 1$ the data point falls inside the region of separation, but on the correct side of the decision surface.

C For $\xi_i > 1$ the data point falls on the wrong side of the separating hyperplane.

D For support vectors ξ_i will be always zero.

(MCQ)

#Q. For the nonseparable case, we minimize the cost function defined as

$$L = \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{N}\xi_i$$

(True/False) The optimal value of C is obtained by minimizing the cost function with respect to C.

A True

B False

(MCQ)

#Q. In continuation with question 2, consider the following statements:

(a) The parameter C can be chosen using cross validation approach.

(b) When C is assigned a small value, the training samples are considered to be noisy, and less emphasis should therefore be placed on it.

(c) The optimization problem for linearly separable patterns can be considered as a special case of optimization problem for nonseparable patterns, by setting $\xi_I = 0$ for points all i.

d) When C is assigned a large value, the implication is that the designer of the SVM has high confidence in the quality of the training samples.

Which of the above statements are correct?

A Only a and c

B Only b and d

C Only a, b and c

D a, b, c and d

(MCQ)

#Q. If we are using a kernel function k to evaluate the inner products in a feature space with feature map 4, the associated Gram matrix G has entries $G_{1_j} = k(x_i, x_j) = \phi(x_1)^T \phi(x_j)$. Then the kernel matrix G is

A Positive definite.

B Negative definite.

C Positive semi-definite.

D Negative semi-definite.

(MCQ)

#Q. In the linearly non-separable case, what effect does the C parameter have on the SVM model?

A it determines the count of support vectors

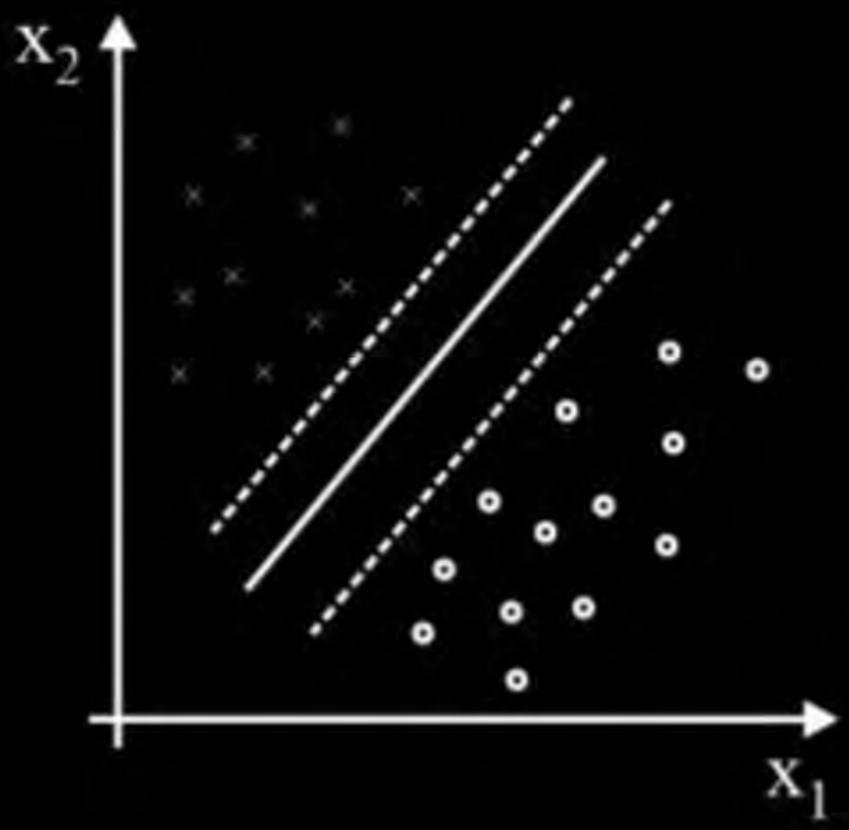
B it is a count of the number of data points which do not lie on their respective side of the hyperplane

C it determines how many data points lie within the margin

D it allows us to trade-off the number of misclassified points in the training data and the size of the margin

(MCQ)

#Q. What is the leave-one-out cross-validation error estimate for maximum margin separation in the following figure?



A 0

B 2

C 3

D 6

THANK - YOU