Supervised learning -

Vapnik - Chervonenkis (VC) Dimension →

→ VC dimension is a measure of the capacity or complexity of a space of functions that can be learned by a classification algorithm (more specifically, hypothesis)

→ The basic definition of VC is the capacity of a classification algorithm, and is defined as the maximum cardinality of points that the algorithm is able to shatter.

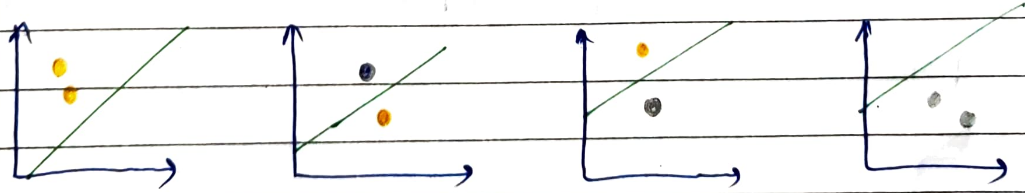
** Shattering is the ability of a model to classify a set of points perfectly.

Linear classifier with two data points →

→ A binary classifier, first is positive class 'A' and another is negative class 'B'.

→ Possible combinations of data points are 2^N.

In our case → 2^2 {++, +-, -+, --}

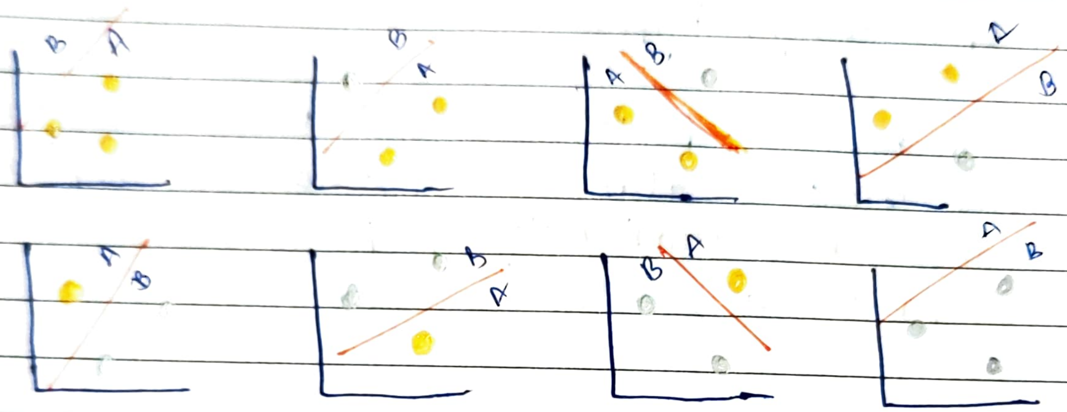


In all the cases, the linear classifier can seperate the positive and negative data points.

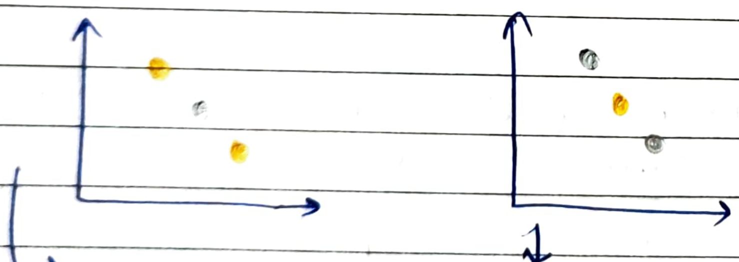
linear classifier with 3 data points.

→ Binary classification with 3 data points (in 2D space). These 3 data points can take either class A or B,

$2^3 = 8$ possible combinations.



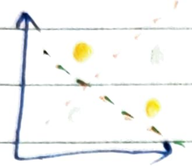
→ A line can shatter 3 points (in general position).



↳ If the data points are placed like this, using linear classifier we cannot classify the data points.

Linear classifier with 4 data points –

$2^4 = 16$ combinations.




Here, line is unable to shatter the two classes.

So, we can say that the linear classifier can shatter at most 3 points.

Rectangle Classifier →

To overcome the drawback of linear classifier, we will move to rectangle classifier.

→ In case of 4 data points, Rectangle classifier can easily shatter in all possible ways.

OUR CLASSIFIER ⇒ Inside the rectangle → positive & outside the rect → neg. examples

given four points (linearly independent), we have the following assignments:

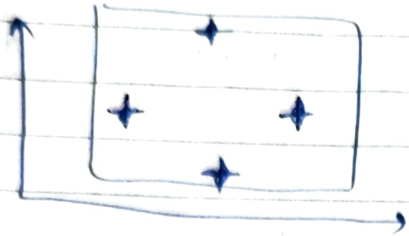
a) All points are positive '+' ⇒ use a rectangle that includes them.

b) All points are negative '−' ⇒ use a empty rectangle.

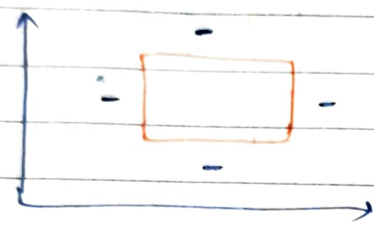
c) 3 points '−' and 1 '+' ⇒ use a rectangle centered on positive points.

d) 3 points '+' and 1 '−' ⇒ we can always find a rectangle which excludes the '−' points.

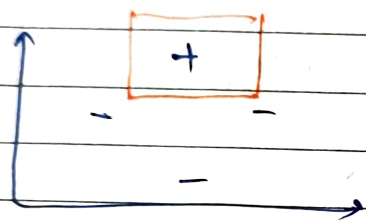
a)



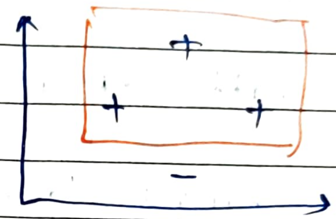
b)



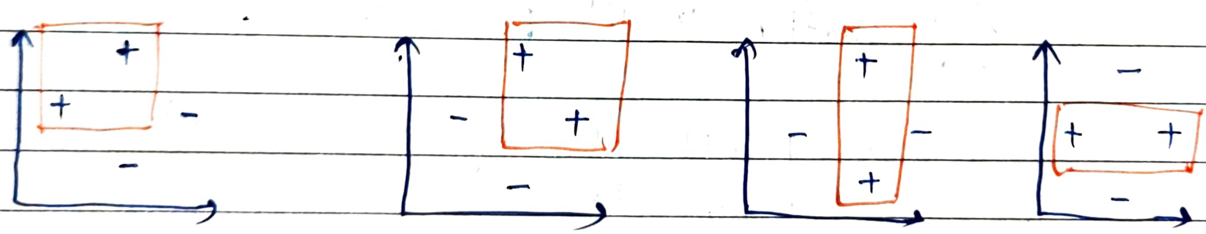
c)



d)



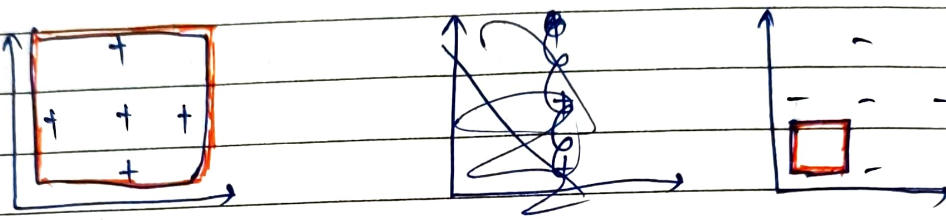
e) 2 points '+' and 2 points '−' ⇒ define a rectangle which includes 2 '+' points & excludes 2 '−' points.

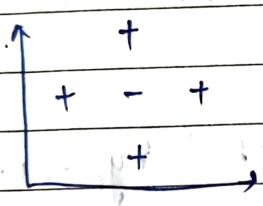


Rectangle classifier with 5 data points :

$2^5 = 32$ combinations.

If all data points are positive or All data points are negative, we can easily classify.





$\Rightarrow$ In this case, using a rectangular classifier we can not classify it.

Vapnik - Cheervonenkis dimensions (VC dimension)

- A dataset containing N points.
- These N points can be labeled in 2^N ways as positive or negative.

→ A hypothesis $h \in H$ that seperates the positive examples from the negative, then we say H shatters N points.

→ The maximum no. of points that can be shattered by H is called the Vapnik-Chervonenkis (VC) dimension of H, is denoted as VC(H), and measures the capacity of H.