Vapnik Chervonenkis (VC) Dimension

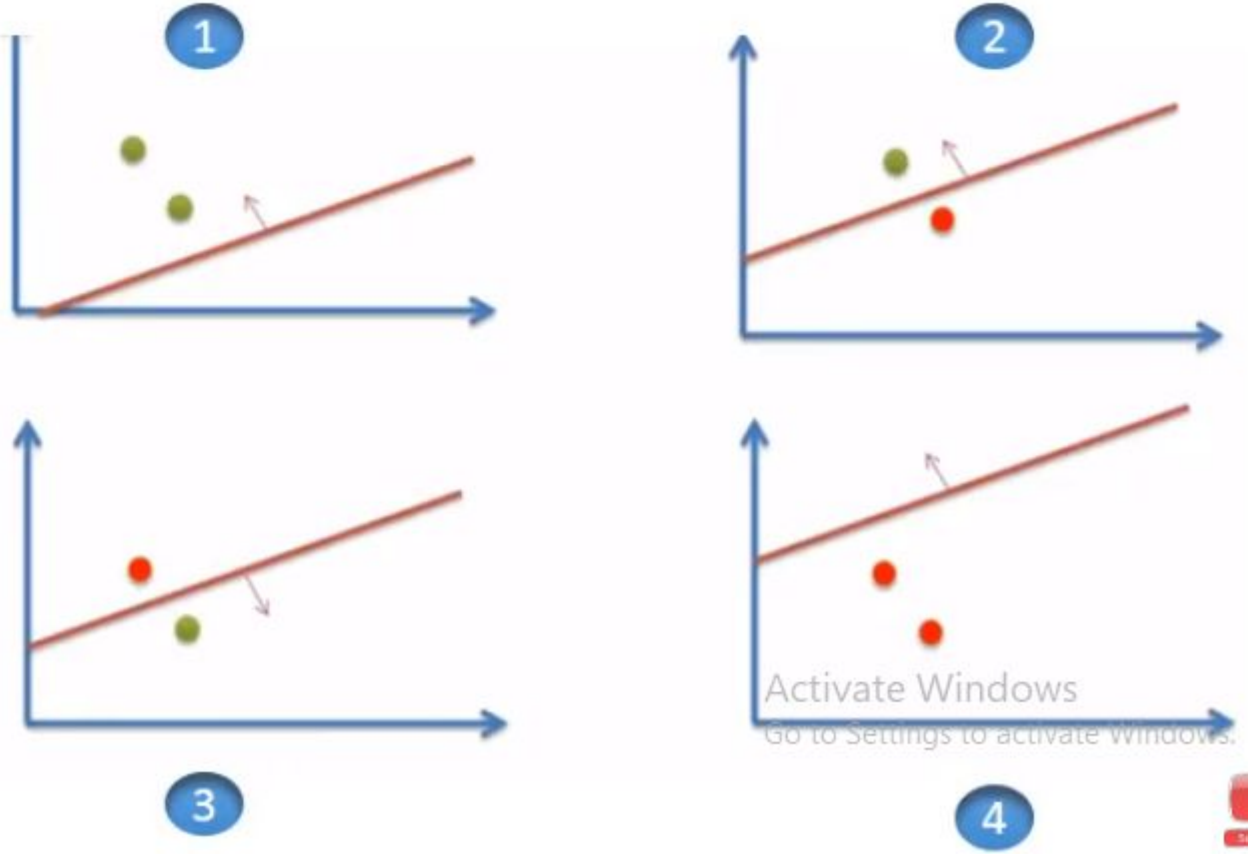
By : Prof. Ansar Sheikh

Vapnik- Chervonenkis (VC) Dimension

- **VC (Vapnik-Chervonenkis) 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 dimension is the capacity of a classification algorithm, and is defined as the **maximum cardinality of the points that the algorithm is able to shatter.**

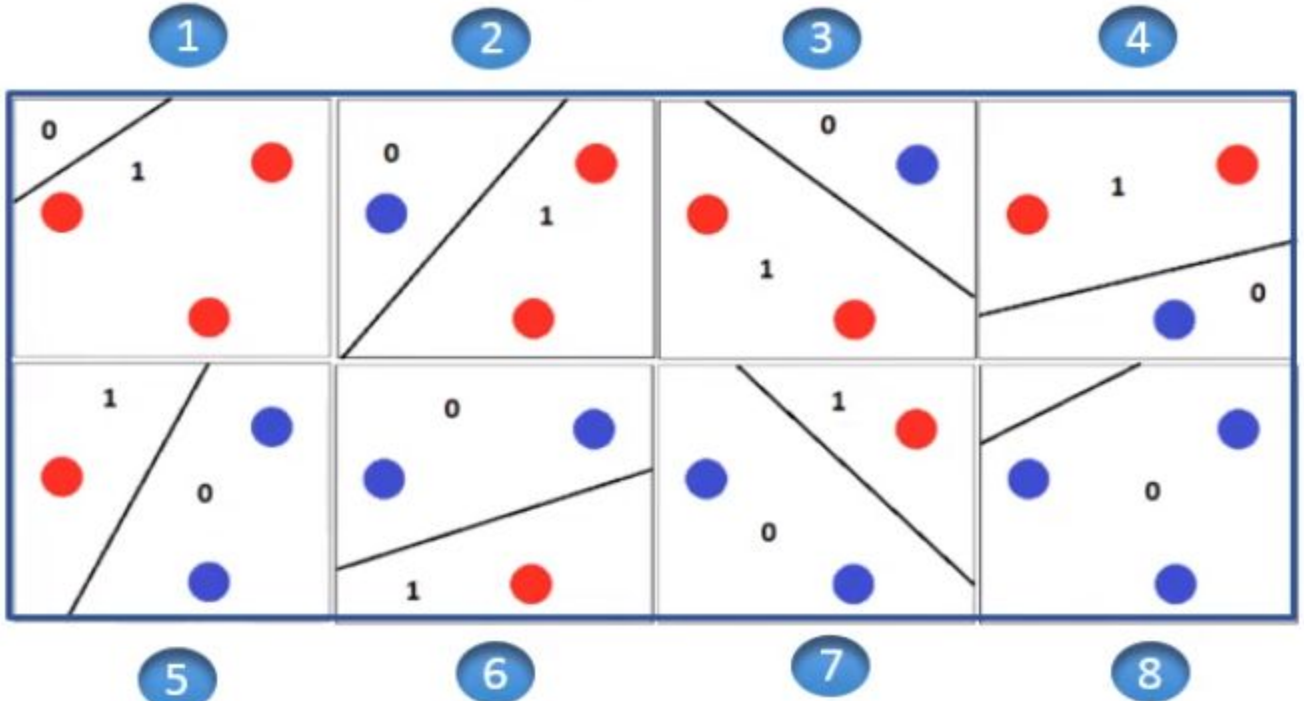
Linear Classifier with two data points

- A binary classifier, first is positive class ‘A’ and another is negative class ‘B’, with two data points.
- The possible combinations of data points are 2^N
- In our case 2^2 , i.e. (++,+-,-+,--)
- In all the cases, the linear classifier can separate the positive and negative data points.



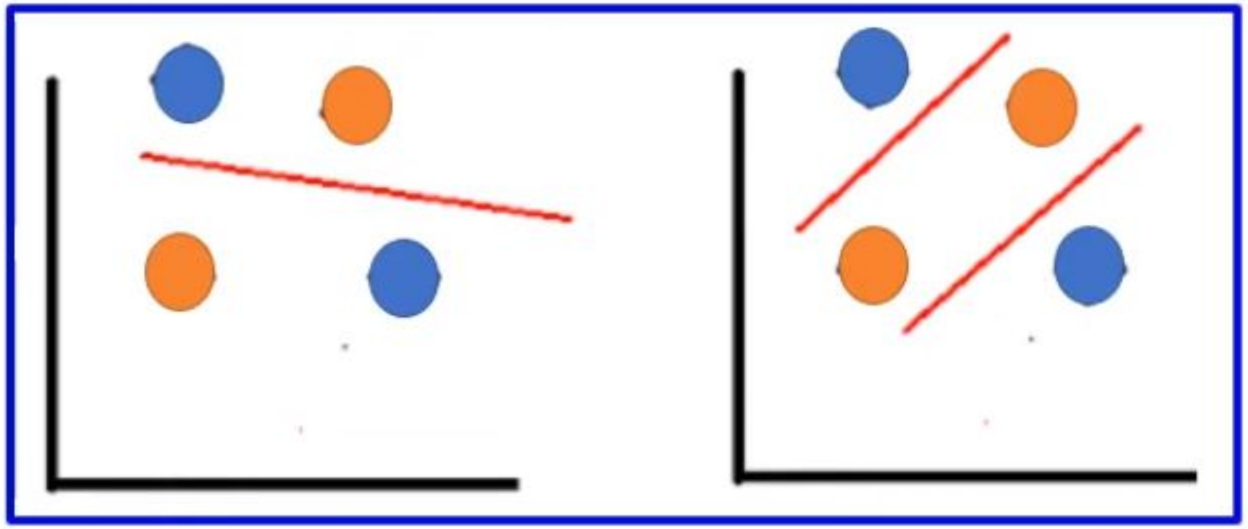
Linear Classifier with three data points

- Binary classification with three data points (in 2D space)
- The 3 points can take either class A (+) or class B (-) which gives us 2^3 (=8) possible combinations (or learning problems).
- a line can shatter 3 points (in general position).



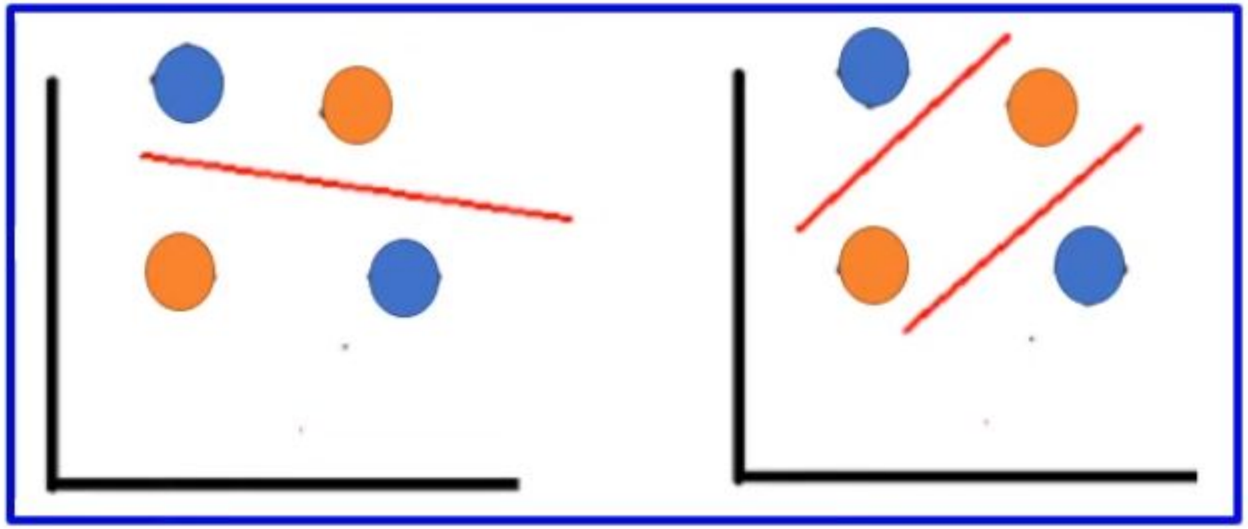
Linear Classifier with four data points

- Now, for the case of 4 points, we can have maximum of 2^4 (=16) possible combinations.
- In Figure that the line was **unable to shatter the two classes**.
- So, we can say that the linear classifier can shatter at most 3 points.



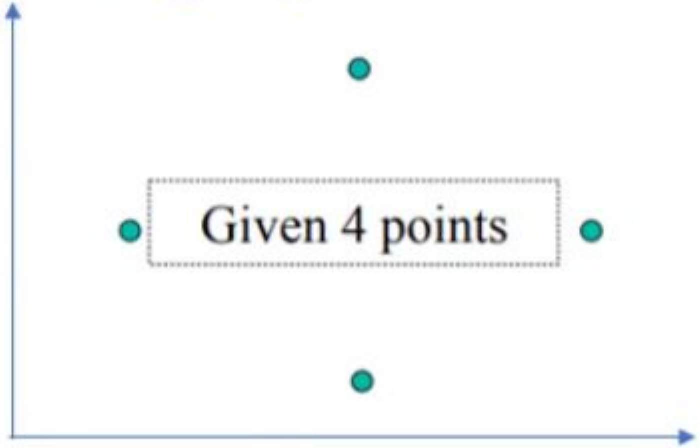
Linear Classifier with four data points

- Now, for the case of 4 points, we can have maximum of 2^4 (=16) possible combinations.
- In Figure that the line was **unable to shatter the two classes**.
- So, we can say that the linear classifier can shatter at most 3 points.



Rectangle Classifier

- In Four data point set, The **rectangle classifier can shattered in all possible ways**
- Given such 4 points, we assign them the {+,-} labels, in all possible ways.
- For each labeling it must exist a rectangle which produces such assignment, i.e. such classification
- Our classifier: **inside the rectangle positive and outside negative examples,** respectively

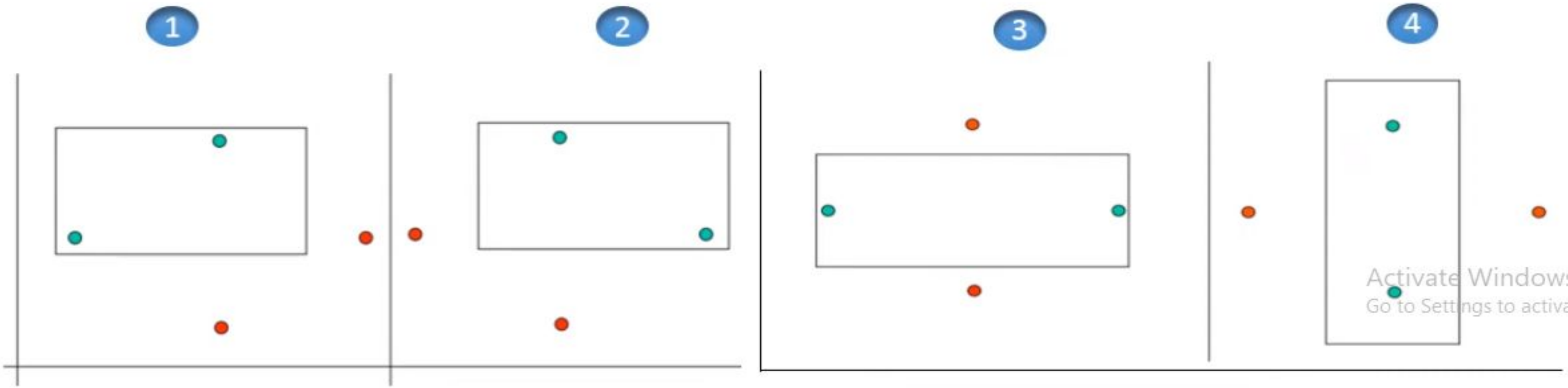


Rectangles Classifier…

- Given 4 points (linearly independent), we have the following assignments:
- a) All points are "+" ⇒ use a rectangle that includes them
- b) All points are "-" ⇒ use a empty rectangle
- c) 3 points "-" and 1 "+" ⇒ use a rectangle centered on the "+" points
- d) 3 points "+" and one "-" ⇒ we can always find a rectangle which exclude the "-" points

Rectangles Classifier…

- e) 2 points "+" and 2 points "-" ⇒ we can define a rectangle which includes the 2 "+" and excludes the 2 "-".



Rectangles Classifier with five data points

- For any 5-point set, we can define a rectangle which has the most extern points as vertices
- If we assign to such vertices the “+” label and to the internal point the “-” label, there will not be any rectangle which reproduces such assignment

Vapnik- Chervonenkis dimension (VC dim).

- A dataset containing N points.
- These N points can be labeled in 2^N ways as positive and negative
- A hypothesis $h \in H$ that separates the positive examples from the negative, then we say **H *shatters* N points.**
- **The maximum number 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.**

Text Book Example : Rectangle can shatter four points.

- An axis-aligned rectangle can shatter four points in two dimensions.
- Then *VC(*H*)*, when H is the hypothesis class of axis-aligned rectangles in two dimensions.
- **Rectangle can separate the positive and negative examples for all possible labeling.**
- Only rectangles covering two points, with all possible shatter are shown in the diagram.