Machine Learning Algorithm

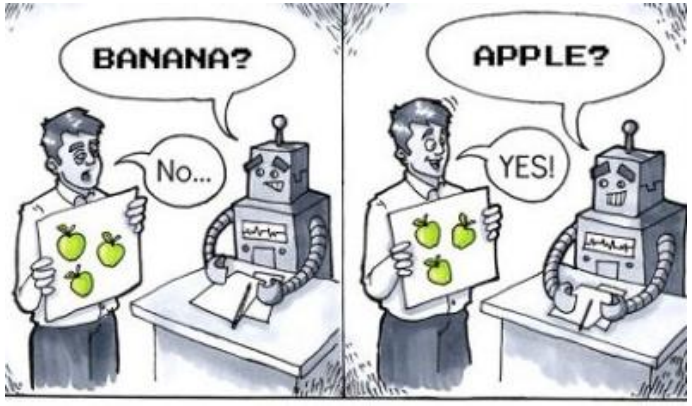
Unit - IV

Dr. Gopal Sakarkar
Department of AI, GHRCE, Nagpur

YouTube Channel on Machine Learning Algorithms
https://tinyurl.com/GopalMachineLearningAlgorithms

Computational Learning Theory

Overview



1. Are there general laws that govern learning?
2. Sample Complexity: How many training examples are needed to learn a successful hypothesis?
3. Computational Complexity: How much computational effort is needed to learn a successful hypothesis?
4. Mistake Bound: How many training examples will the learner misclassify before converging to a successful hypothesis?

- Why learning works
    1. Under what conditions is successful learning possible and impossible?
    2. Under what conditions is a particular learning algorithm assured of learning successfully?
- We need particular setting(models)
    1. Probably approximately correct(PAS)
    2. Mistake bond models

Some Terms

X is the set of all possible instances

C the set of all possible concepts c

where c: X > {0,1}

H is the set of hypotheses considered

by a learner, $H \subseteq C$

L is the learner

D is a probability distribution over X that

generates observed instances

Definition

The true error of hypothesis h, with respect to the target concept c and observation distribution D is the probability that h will misclassify an instance drawn according to D

$$error_D \equiv \underset{x \in D}{P}[c(x) \neq h(x)]$$

In a perfect world, we'd like the true error to be 0

The world isn't perfect

- We typically can't provide every instance for training.
- Since we can't , there is always a chance the examples provided the learner will be misleading
  - "No Free Lunch" theorem
- So we'll go for a weaker thing:
- PROBABLY APPROXIMATELY CORRECT learning

Definition: PAC - learnable

- A concept class C is “PAC learnable” by a hypothesis class H if there exists a learning algorithm L such that..
  - given any target concept c in C,
  - any target distribution D over the possible examples X,
  - and any pair of real numbers 0< e, d <1

- That L takes as input a training set of m examples drawn according to D, where the size of m is bounded above by a polynomial in 1/ε and $1/\delta$
- and outputs an hypothesis h in H about which we can say, with confidence (probability over all possible choices of the training set) greater than 1 − d
- that the error of the hypothesis is less than e.

$$error_D \equiv \underset{x \in D}{P}[c(x) \neq h(x)] \leq \varepsilon$$

For Finite Hypothesis Spaces

- A hypothesis is consistent with the training data if it returns the correct classification for every example presented it.
- A consistent learner returns only hypotheses that are consistent with the training data.
- Given a consistent learner, the number of examples sufficient to assure that any hypothesis will be probably (with probability (1- δ)) approximately (within error ε) correct is...

$$m \geq \frac{1}{\varepsilon}\left(\ln |H| + \ln(1/\delta)\right)$$

Theorem

- If the hypothesis space H is finite, and D is a sequence of m ≥1 independent random examples of some target concept c, then for any $0 \leq \epsilon \leq 1$, the probability that $VS_{H,D}$ contains a hypothesis with error greater than is less than

$$|H|e^{-\epsilon m}$$

- Proof sketch:
  - Prob(1 hyp. w error consistent w/ 1 ex.) <
  - Prob(1 hyp. w error e consistent with m exs.) < $e^{-\epsilon m}$
  - Prob(1 of H hyps. consistent with m exs.) < $|H|e^{-\epsilon m}$

Theorem

Interesting! This bounds the probability that any consistent learner will output a hypothesis h with $error(h) \geq \epsilon$

If we want this probability to be at most δ

$$|H|e^{-\epsilon m} \leq \delta$$

then

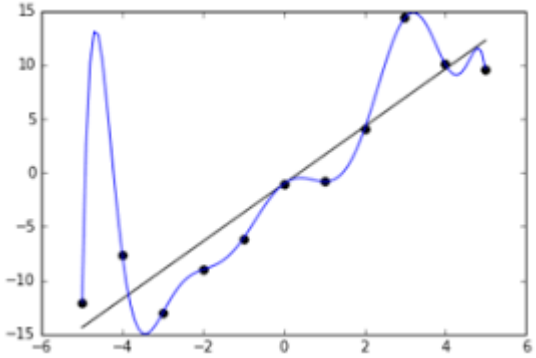
$$m \geq \frac{1}{\epsilon}(\ln |H| + \ln(1/\delta))$$

Theorem

- The PAC Learning framework has 2 disadvants:
  - It can lead to weak bounds
  - Sample Complexity bound cannot be established for infinite hypothesis spaces

- We introduce the VC dimension for dealing with these problems (particularly the second one)

Principal Component Analysis



- Need of PCA:
- It is required to solve the problem of Over-fitting .
- During training phase, when we provide large number of attributes , then our model will get confuse.
- To reduce this problem , PCA is try to reduce over-fitting problem .
- Where ,our model is trying to touch each and every points , thats create
- Over fitting problem.
- PCA is used to convert high dimensionality to low dimensionality.

Principal Component Analysis

- What happens when the given data set has too many variables?

- Here are few possible situations which you might come across:
- You find that most of the variables are correlated on analysis.
- You lose patience and decide to run a model on the whole data. This returns poor accuracy and you feel terrible.
- You become indecisive about what to do
- You start thinking of some strategic method to find few important variables

What is Principal Component Analysis?

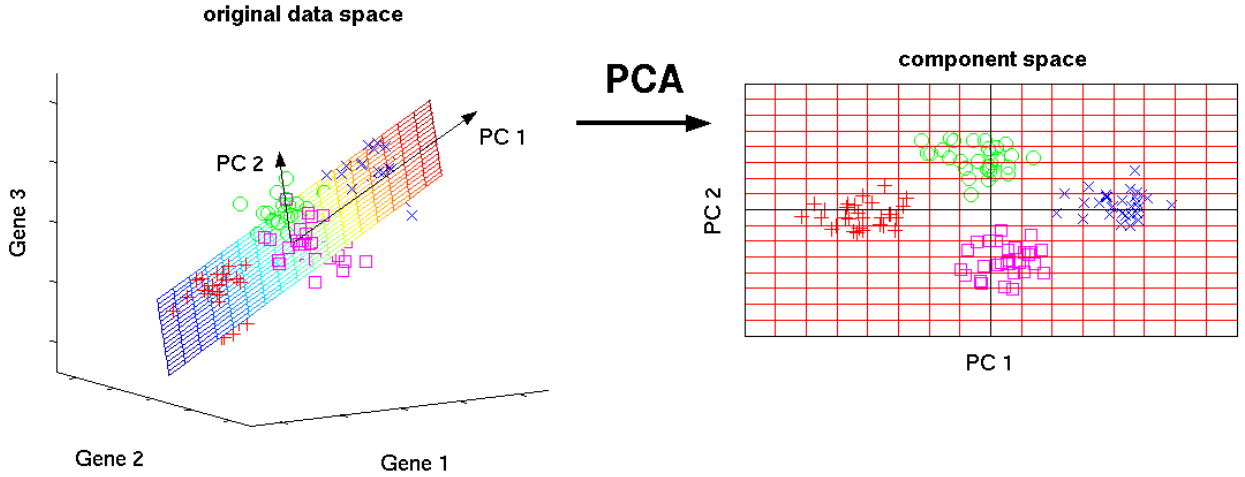
- In simple words, PCA is a method of obtaining important variables (in form of components) from a large set of variables available in a data set.
- It extracts low dimensional set of features by taking a projection of irrelevant dimensions from a high dimensional data set with a motive to capture as much information as possible.
- With fewer variables obtained while minimizing the loss of information, visualization also becomes much more meaningful.
- PCA is more useful when dealing with 3 or higher dimensional data.
- It is always performed on a symmetric correlation or covariance matrix.
- This means the matrix should be numeric and have standardized data

What is Principal Component Analysis?

- Let's say we have a data set of dimension 300 (n) × 50 (p).
- n represents the number of observations and p represents number of predictors.
- Since we have a large p = 50, there can be p(p-1)/2 scatter plots i.e more than 1000 plots possible to analyze the variable relationship.
- Wouldn't is be a tedious job to perform exploratory analysis on this data ?
- In this case, it would be a lucid approach to select a subset of p (p << 50) predictor which captures as much information. Followed by plotting the observation in the resultant low dimensional space.

Example : Principal Component Analysis?

- The im below shows the transformation of a high dimensional data (3 dimension) to low dimensional data (2 dimension) using PCA.
- Not to forget, each resultant dimension is a linear combination of p features



Principal Component Analysis

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

What are principal components ?

- A principal component is a normalized linear combination of the original predictors in a data set. In im above, PC1 and PC2 are the principal components.

- Let's say we have a set of predictors as X^1, X^2...,Xp

- The principal component can be written as:

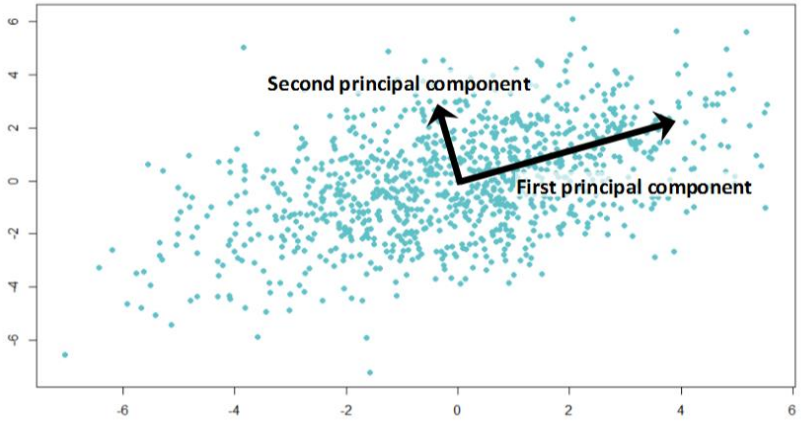
- $Z^1 = \Phi^{11}X^1 + \Phi^{21}X^2 + \Phi^{31}X^3 + ... + \Phi p^1 Xp$
- where,
- Z^1 is first principal component
- Φp^1 is the loading vector comprising of loadings (Φ^1, Φ^2..) of first principal component. The loadings are constrained to a sum of square equals to 1. This is because large magnitude of loadings may lead to large variance. It also defines the direction of the principal component (Z^1) along which data varies the most. It results in a line in p dimensional space which is closest to the n observations. Closeness is measured using aver squared euclidean distance.
- X^1..Xp are normalized predictors. Normalized predictors have mean equals to zero and standard deviation equals to one.

What are principal components ?

- First principal component is a linear combination of original predictor variables which captures the maximum variance in the data set.
- It determines the direction of highest variability in the data. Larger the variability captured in first component, larger the information captured by component.
- No other component can have variability higher than first principal component.
- The first principal component results in a line which is closest to the data i.e. it minimizes the sum of squared distance between a data point and the line.

What are principal components ?

- Second principal component is also a linear combination of original predictors which captures the remaining variance in the data set and is uncorrelated with Z^1. In other words, the correlation between first and second component should is zero. It can be represented as:
- $Z^2 = \Phi^{12}X^1 + \Phi^{22}X^2 + \Phi^{32}X^3 + ... + \Phi p2Xp$
- If the two components are uncorrelated, their directions should be orthogonal (im below). This im is based on a simulated data with 2 predictors. Notice the direction of the components, as expected they are orthogonal. This suggests the correlation b/w these components in zero.



What are principal components ?

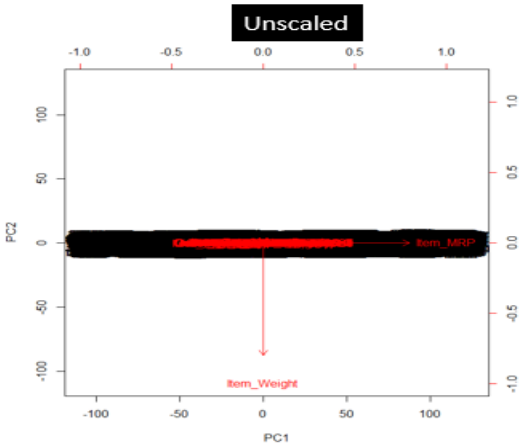
- All succeeding principal component follows a similar concept i.e. they capture the remaining variation without being correlated with the previous component.
- In general, for n × p dimensional data, min(n-1, p) principal component can be constructed.
- The directions of these components are identified in an unsupervised way i.e. the response variable(Y) is not used to determine the component direction.
- Therefore, it is an unsupervised approach.

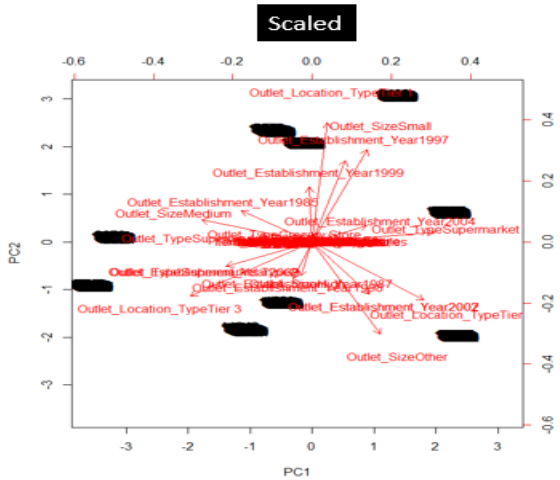
Why is normalization of variables necessary in PCA ?

- The principal components are supplied with normalized version of original predictors.
- This is because, the original predictors may have different scales. For example: Imagine a data set with variables' measuring units as gallons, kilometers, light years etc.
- It is definite that the scale of variances in these variables will be large.
- Performing PCA on un-normalized variables will lead to insanely large loadings for variables with high variance.
- In turn, this will lead to dependence of a principal component on the variable with high variance. This is undesirable.

What are principal components ?

- As shown in im below, PCA was run on a data set twice (with unscaled and scaled predictors).
- This data set has ~40 variables.
- You can see, first principal component is dominated by a variable Item_MRP. And, second principal component is dominated by a variable Item_Weight.
- This domination prevails due to high value of variance associated with a variable.
- When the variables are scaled, we get a much better representation of variables in 2D space





What are principal components ?

Step :1

- Find the Mean value of X and Y
- X = 1.81 Y=1.91

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	1.0

What are principal components ?

Step :2 To find the covariance

Covariance

$$\text{covariance}(X,Y) = \frac{\sum_{i=1}^{n} (\bar{X}_i - X)(\bar{Y}_i - Y)}{(n-1)}$$

- So, if you had a 3-dimensional data set (x,y,z), then you could measure the covariance between the x and y dimensions, the y and z dimensions, and the x and z dimensions. Measuring the covariance between x and x , or y and y , or z and z would give you the variance of the x , y and z dimensions respectively.

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	1.0

What are principal components ?

Step :2 To find the covariance

Covariance

- What is the interpretation of covariance calculations?

  e.g.: 2 dimensional data set

  x: number of hours studied for a subject

  y: marks obtained in that subject

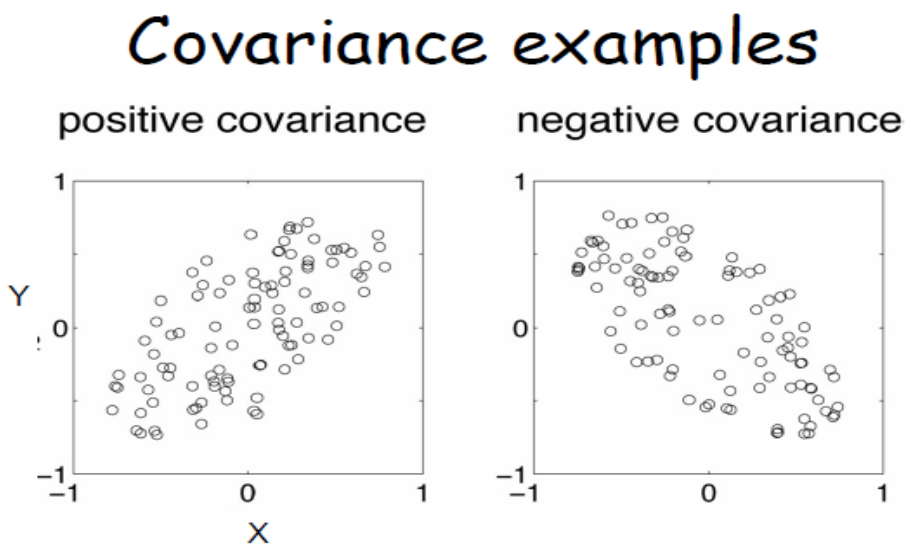
  covariance value is say: 104.53

  what does this value mean?

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	1.0

What are principal components ?

Step :2 To find the covariance



X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	1.0

What are principal components ?

Step :2 To find the covariance

Covariance

- Exact value is not as important as it's sign.
- A positive value of covariance indicates both dimensions increase or decrease together e.g. as the number of hours studied increases, the marks in that subject increase.
- A negative value indicates while one increases the other decreases, or vice-versa e.g. active social life at PSU vs performance in CS dept.
- If covariance is zero: the two dimensions are independent of each other e.g. heights of students vs the marks obtained in a subject

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	1.0

What are principal components ?

Step :2 To find the covariance

It is the mean value of the product of the deviations of two varieties from their respective means.

$$Cov(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-y)}{N-1}$$

Before to find the covariance , we have to find out **Covariance Matrix** :

A **covariance matrix** is a square **matrix** giving the **covariance** between each pair of elements of a given random vector.

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	1.0

What are principal components ?

X = 1.81 Y=1.91

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	1.0

Step :2 Find out Covariance Matrix :

$$\text{Cov}(A) = \begin{bmatrix} \text{Cov}(X,X) & \text{Cov}(Y,X) \\ \text{Cov}(X,Y) & \text{Cov}(Y,Y) \end{bmatrix}$$

Cov(X,X)=

X	$x-\bar{x}$	$(x-\bar{x})(x-\bar{x})$
2·5	0·69	0·476
0·5	−1·31	1·7161
Sum = 5.5490		

Now, to find the Cov(X,X)
5.5490 / N-1
i.e 5.5490 /9
=0.6165

Do this for all **10** data entry points of X column.

What are principal components ?

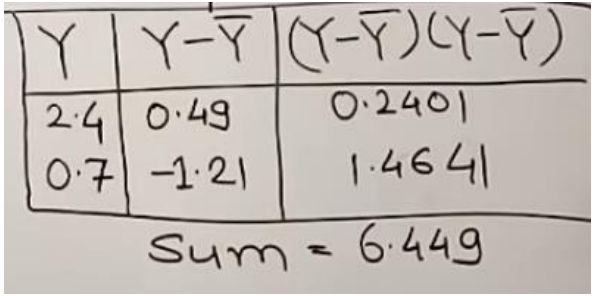
X = 1.81 Y=1.91

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	1.0

Step :2 Find out Covariance Matrix :

$$\text{Cov}(A) = \begin{bmatrix} \text{Cov}(X,X) & \text{Cov}(Y,X) \\ \text{Cov}(X,Y) & \text{Cov}(Y,Y) \end{bmatrix}$$

Cov(Y,Y)=



Y	$Y-\bar{Y}$	$(Y-\bar{Y})(Y-\bar{Y})$
2.4	0.49	0.2401
0.7	-1.21	1.4641
	Sum = 6.449	

Now, to find the Cov(Y,Y)
6.449 / N-1
i.e 6.449 /9
=0.7165

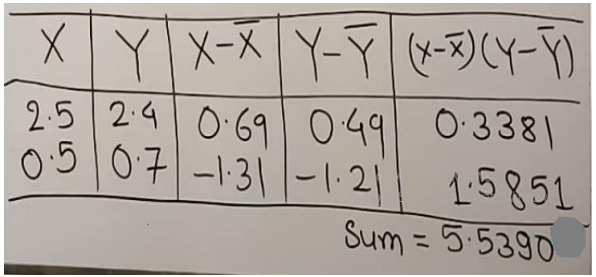
Do this for all **10** data entry points of X column.

What are principal components ?

Step :2 Find out Covariance Matrix :

$$\text{Cov}(A) = \begin{bmatrix} \text{Cov}(X,X) & \text{Cov}(Y,X) \\ \text{Cov}(X,Y) & \text{Cov}(Y,Y) \end{bmatrix}$$

Cov(X,Y)=



X	Y	X-$\bar{X}$	Y-$\bar{Y}$	(X-$\bar{X}$)(Y-$\bar{Y}$)
2.5	2.4	0.69	0.49	0.3381
0.5	0.7	-1.31	-1.21	1.5851
				Sum = 5.5390

Now, to find the Cov(X,Y) and Cov(Y,X)
5.5390 / N-1
i.e 5.5390 /9
=0.6154

Do this for all **10** data entry points of X column.

X = 1.81 Y=1.91

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	1.0

What are principal components ?

X = 1.81 Y=1.91

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	1.0

Step :2 Find out Covariance Matrix :

$$Cov(A) = \begin{bmatrix} Cov(X,X) & Cov(Y,X) \\ Cov(X,Y) & Cov(Y,Y) \end{bmatrix}$$

$$Cov(X,Y) = \begin{bmatrix} 182.2 & 182.2 \\ 182.2 & 182.2 \end{bmatrix}$$

Now, to find eigenvalue

$$C - \lambda * I = 0$$

Where, λ is eigenvalue
I identity matrix
C is covariance matrix

What are principal components ?

X = 1.81 Y=1.91

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	1.0

Now, to find Eigen- Value

$$\text{Cov}(X,Y)= \begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix} - \lambda * \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 0$$

$$\begin{bmatrix} 0.6165 - \lambda & 0.6154 \\ 0.6154 & 0.7165 - \lambda \end{bmatrix} = \lambda^2 - 1333\,\lambda + 0.0630 = 0$$

What are principal components ?

X = 1.81 Y=1.91

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	1.0

Now, to find **Eigen- Value**

$$\lambda^2 - 1333\,\lambda + 0.0630 = 0$$

We have to generate 2 eigenvalues

$\lambda_1 =$ **0.0490** **0**

$\lambda_2 =$ **1.2840** **364.4**

What are principal components ?

X = 1.81 Y=1.91

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	1.0

Now, to find Eigen- Vector

$$C * V = \lambda * V$$

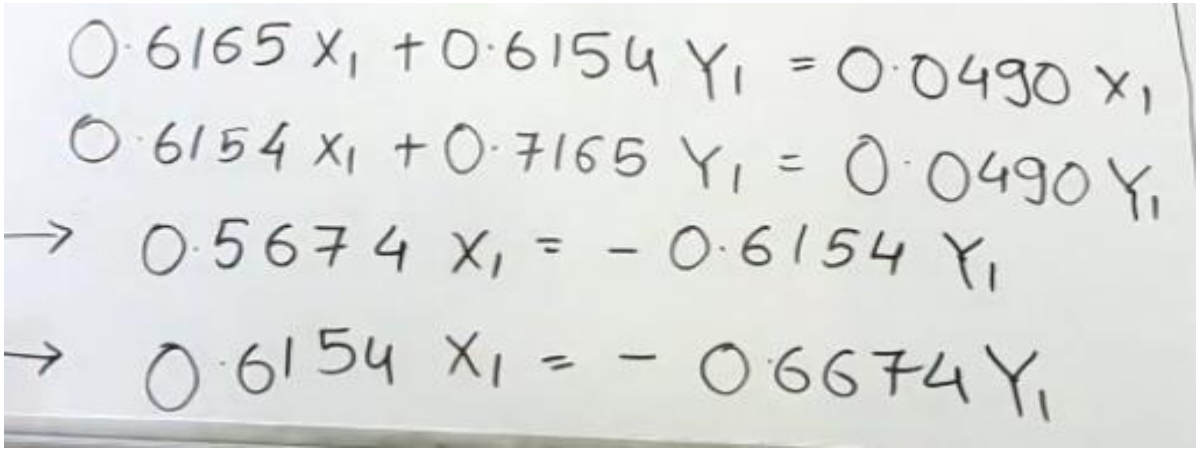
Where, C is covariance matrix
V is Eigen Vector
λ is Eigen Value

Now, to find Eigen- Vevtor1, consider λ_1

$$\begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix} \begin{bmatrix} X1 \\ Y1 \end{bmatrix} = \mathbf{0.0490} \begin{bmatrix} X1 \\ Y1 \end{bmatrix}$$

What are principal components ?

$$\begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix} \begin{bmatrix} X1 \\ Y1 \end{bmatrix} = \mathbf{0.0490} \begin{bmatrix} X1 \\ Y1 \end{bmatrix}$$

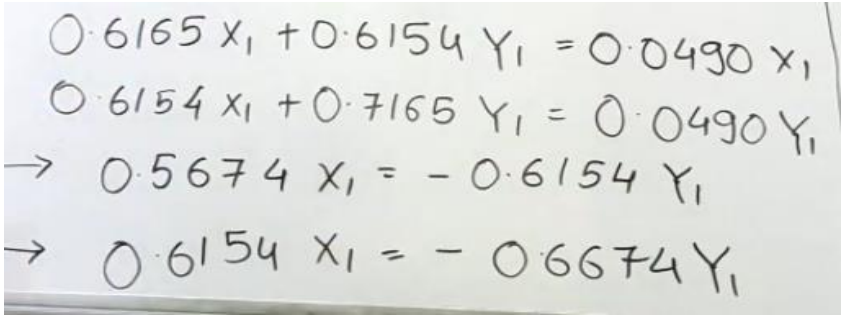


X = 1.81 Y=1.91

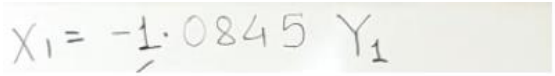
X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	1.0

What are principal components ?

Now, to find **Eigen- Vector**



X1= - 0.6154 / 0.5674 Y1



Let us consider $Y_1 = 1$

$$\begin{bmatrix} X_1 \\ Y_1 \end{bmatrix} \quad \begin{bmatrix} -1.0845 \\ 1 \end{bmatrix}$$

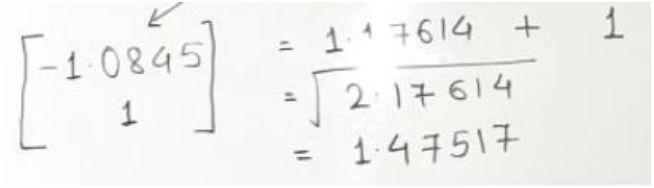
X = 1.81 Y=1.91

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	1.0

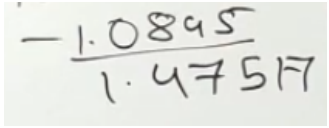
What are principal components ?

Now, to find **Eigen- Vector**

Now , do the $X^2 + Y^2$ and then $\sqrt{X^2 + Y^2}$



Now do the



$$\begin{bmatrix} X1 \\ Y1 \end{bmatrix} = \begin{bmatrix} -0.7351 \\ 0.6778 \end{bmatrix}$$

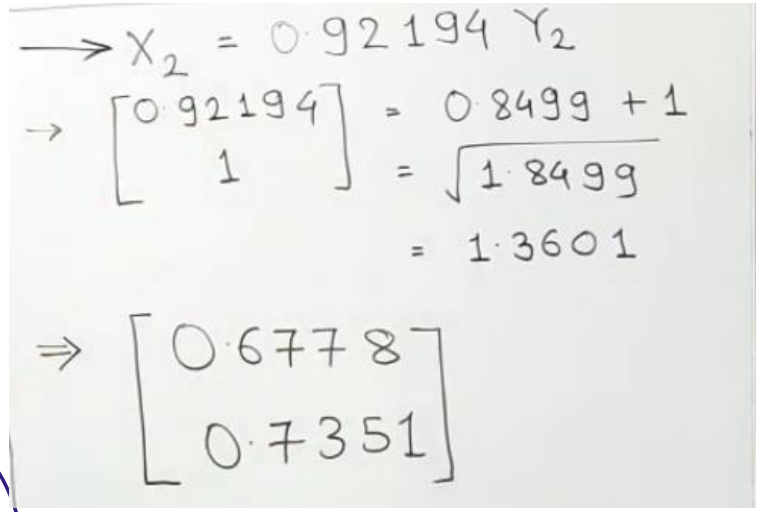
X = 1.81 Y=1.91

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	1.0

What are principal components ?

Now, to find Eigen-Vector , consider λ_2

$$\begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix} \begin{bmatrix} X2 \\ Y2 \end{bmatrix} = 1.2840 \begin{bmatrix} X2 \\ Y2 \end{bmatrix}$$



Let us consider Y2 =1

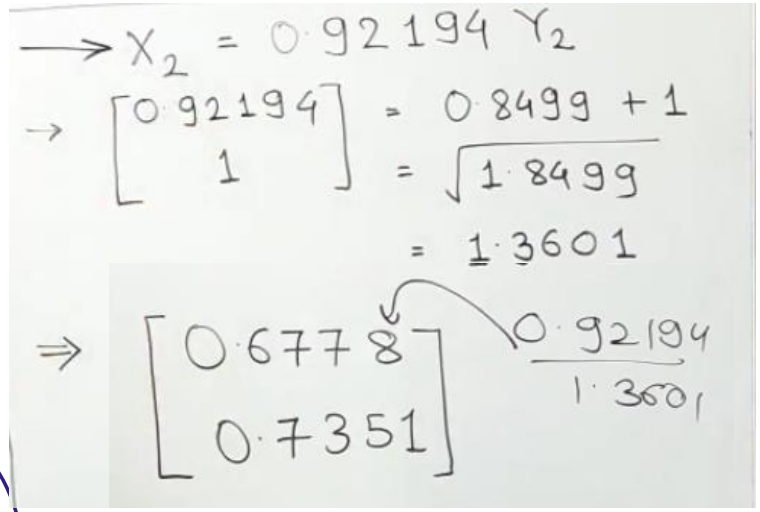
X = 1.81 Y=1.91

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	1.0

What are principal components ?

Now, to find Eigen-Vector , consider λ_2

$$\begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix} \begin{bmatrix} X2 \\ Y2 \end{bmatrix} = 1.2840 \begin{bmatrix} X2 \\ Y2 \end{bmatrix}$$



Let us consider Y2 =1

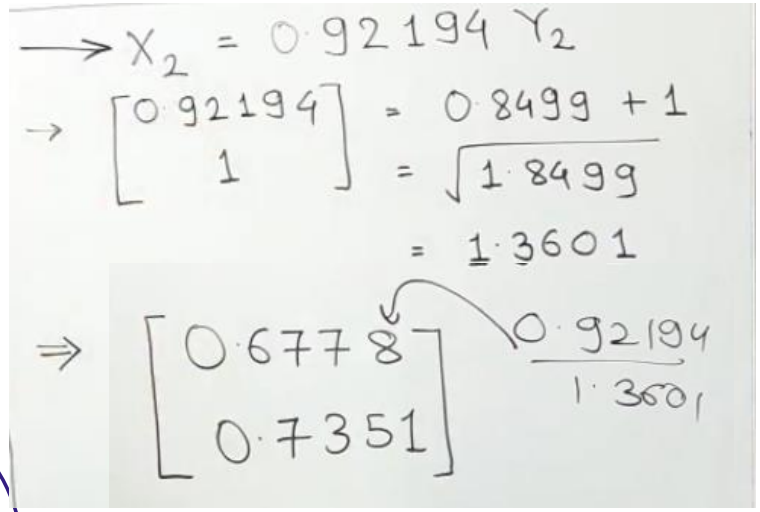
X = 1.81 Y=1.91

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	1.0

What are principal components ?

Now, to find Eigen-Vector , consider λ_2

$$\begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix} \begin{bmatrix} X2 \\ Y2 \end{bmatrix} = 1.2840 \begin{bmatrix} X2 \\ Y2 \end{bmatrix}$$



Let us consider Y2 =1

This is your final second **Eigen- Vector**

X = 1.81 Y=1.91

X	Y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	1.0

Principal Component Analysis

Summary of PCA

01

Find the Mean value of Attributes (X and Y)

02

Find out Covariance Matrix

03

Now, to find Eigen- Value

04

Finally , to find Eigen- Vector

The VC-Dimension

- Definition: A set of instances S is shattered by hypothesis space H iff for every dichotomy of S there exists some hypothesis in H consistent with this dichotomy.

- Definition: The Vapnik-Chervonenkis dimension, VC(H), of hypothesis space H defined over instance space X is the size of the largest finite subset of X shattered by H. If arbitrarily large finite sets of X can be shattered by H, then VC(H)=□

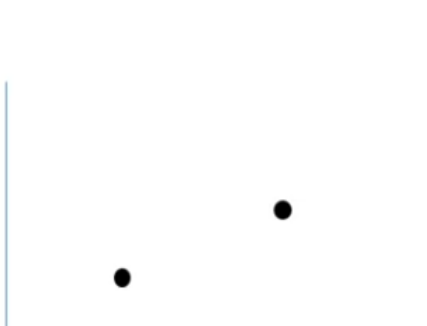
- The Vapnik–Chervonenkis (VC) dimension is a measure of the capacity (complexity, expressive power, richness, or flexibility) of a space of functions that can be learned by a statistical classification algorithm.

The VC-Dimension

Total data points =2

Classification
Class : A, true, 1,yes [green]
Class: B,false,0,no [red]

The VC-Dimension

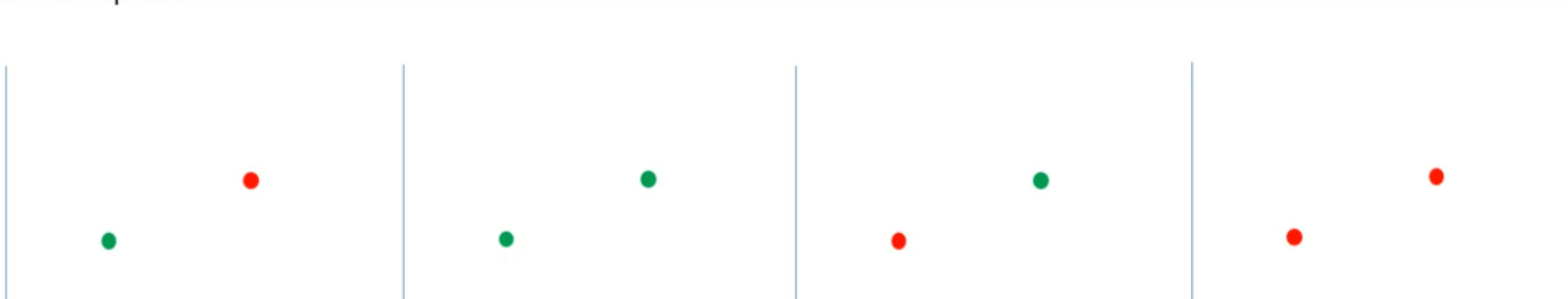


Total data points =2

Classification
Class : A, true, 1,yes [green]
Class: B,false,0,no [red]

The VC-Dimension

Total data points =2



Two numbers can be classified in four different ways.

The VC-Dimension

Total data points =3

Three numbers can be classified in 8 different ways.

The VC-Dimension

Total data points =N

Two numbers can be classified in 2^N different ways.

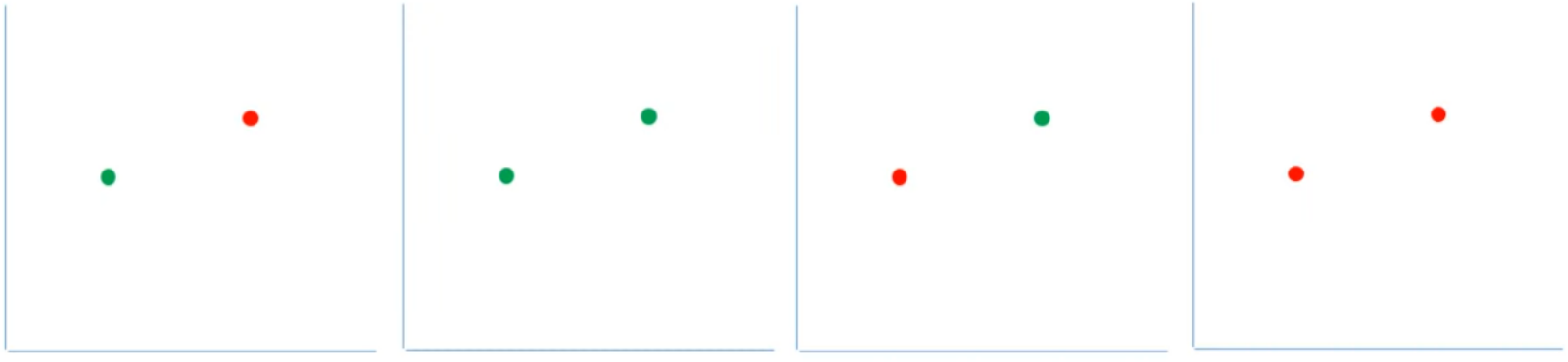
The VC-Dimension

Shattering

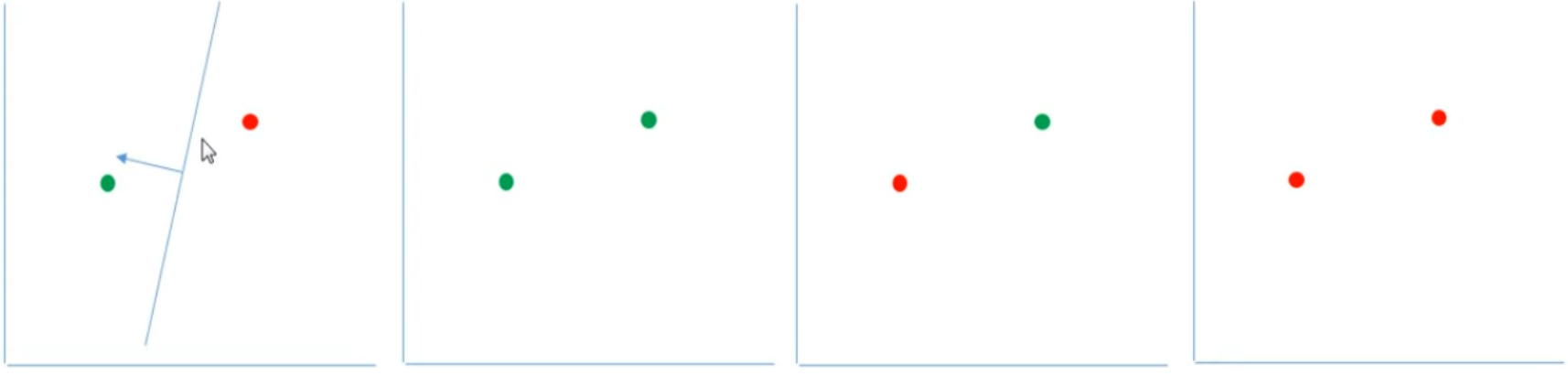
A hypothesis class H can shatter N data points for which we can find a hypothesis $h \in H$ that separates the positive examples from the negative for every problem, then we say *H shatters N* points.

The VC-Dimension

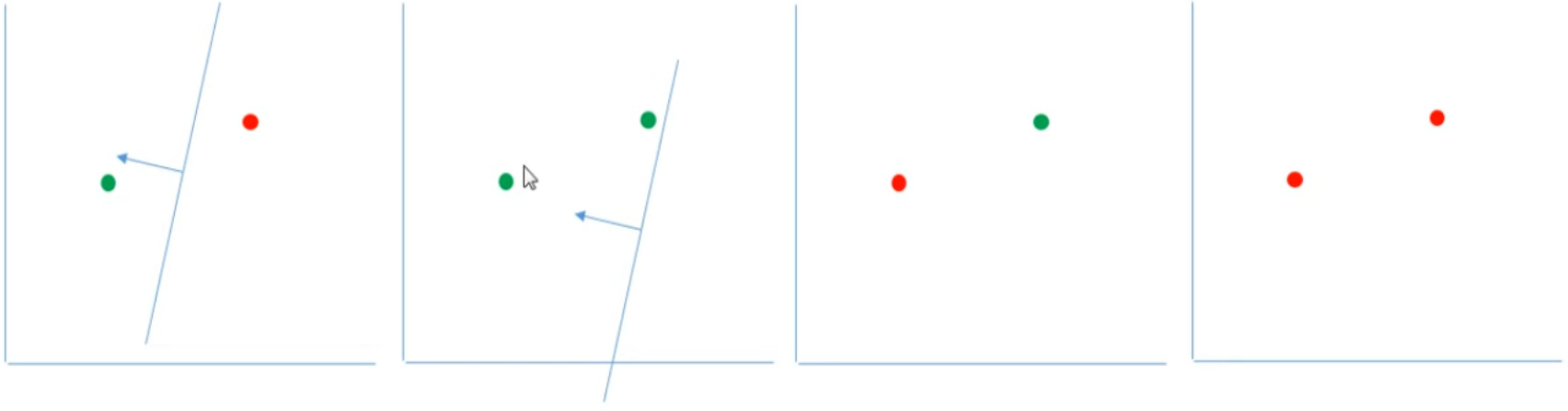
Can a line hypothesis class can shatter 2 data points?



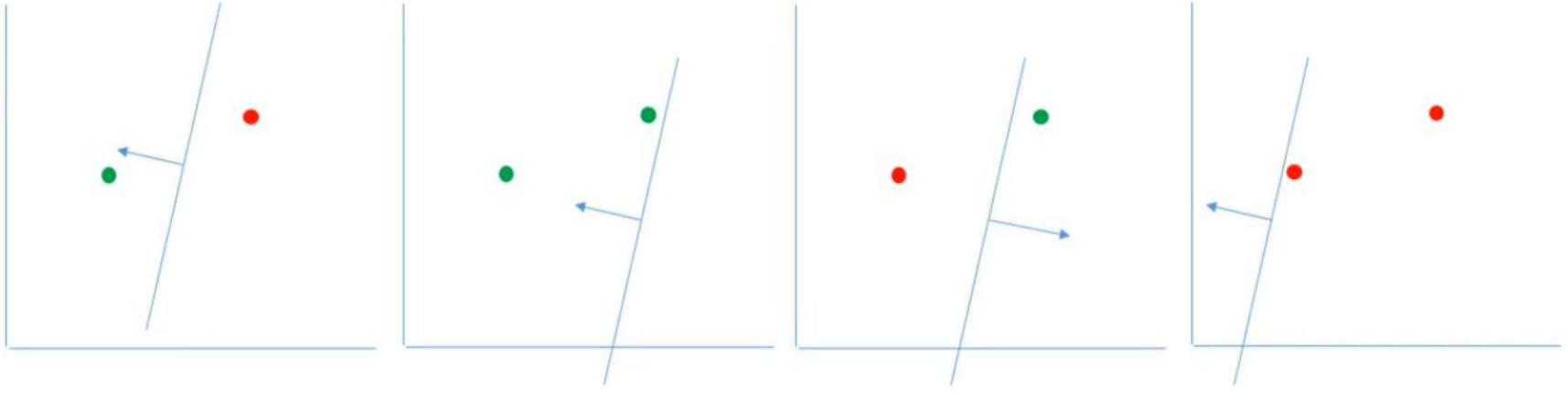
The VC-Dimension



The VC-Dimension

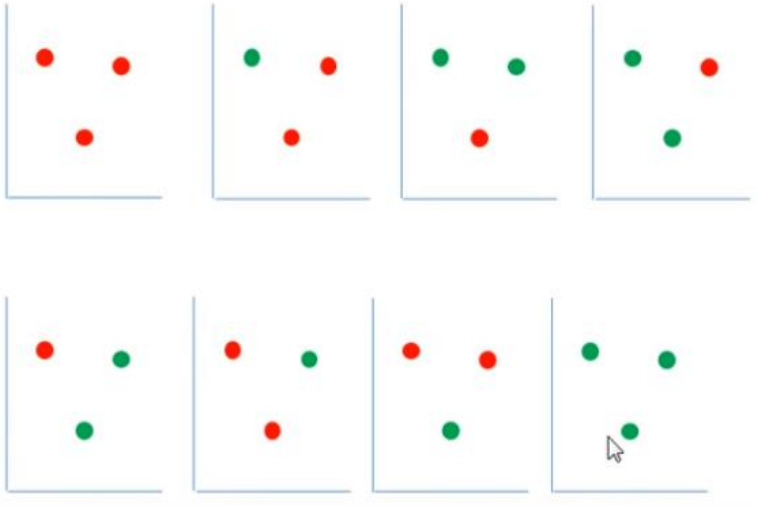


The VC-Dimension



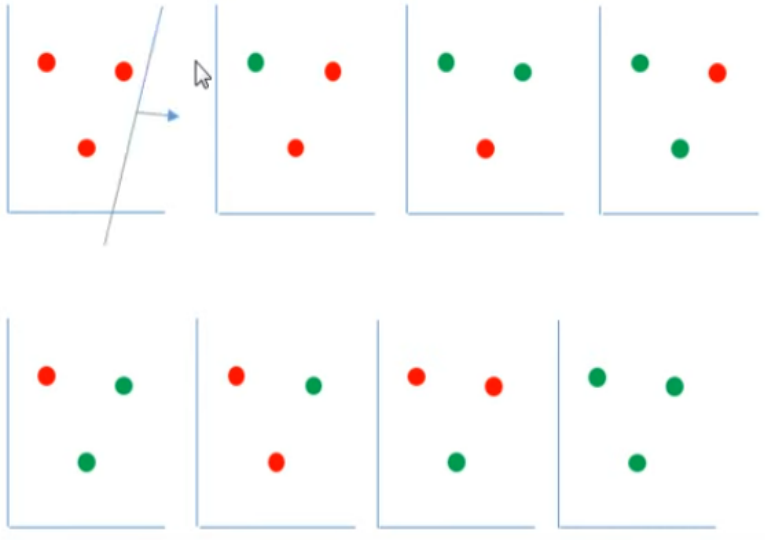
The VC-Dimension

Can a line hypothesis class shatter 3 data points?



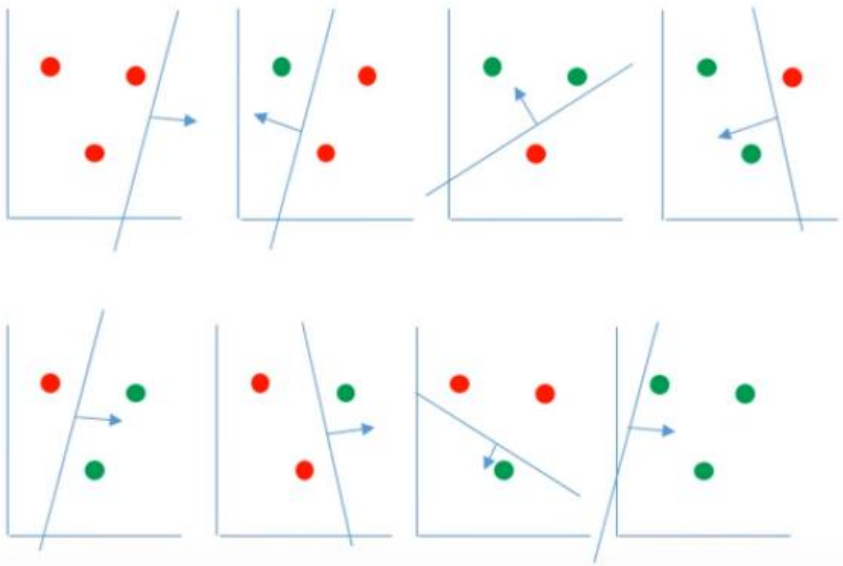
The VC-Dimension

Can a line hypothesis class shatter 3 data points?



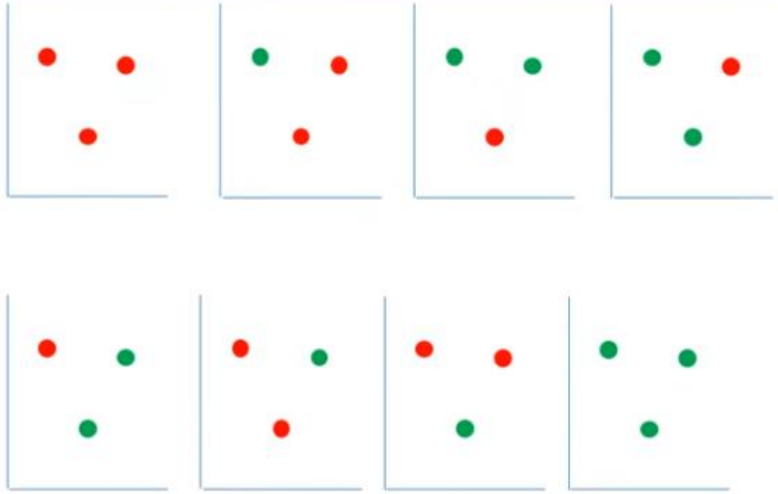
The VC-Dimension

Can a line hypothesis class shatter 3 data points?



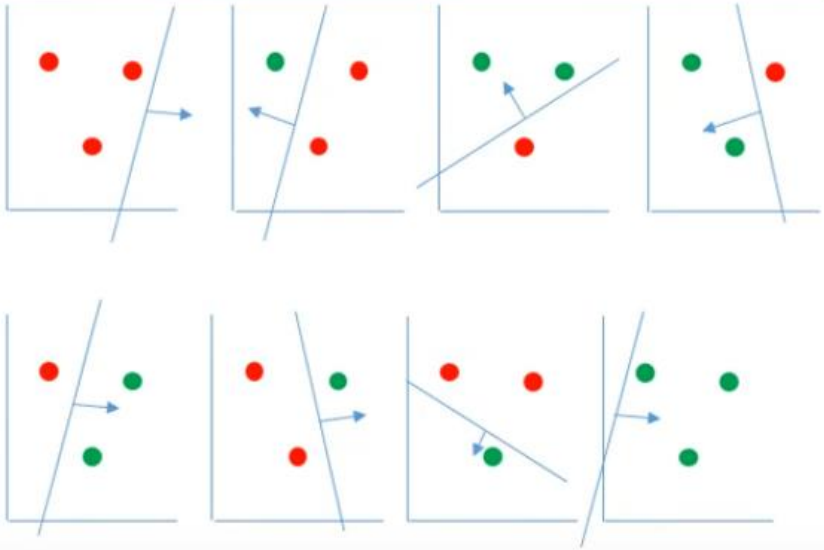
The VC-Dimension

Note : When we say H shatter N data points than it does not mean that H can shatter every N data points. If you can find even a single dataset of N data points for which H can shatter them.



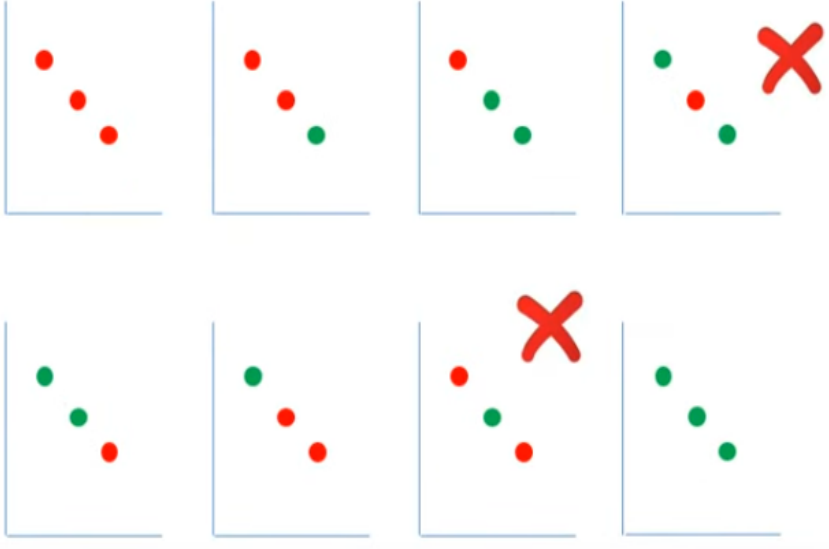
The VC-Dimension

Note : When we say H shatter N data points than it does not mean that H can shatter every N data points. If you can find even a single dataset of N data points for which H can shatter them.



The VC-Dimension

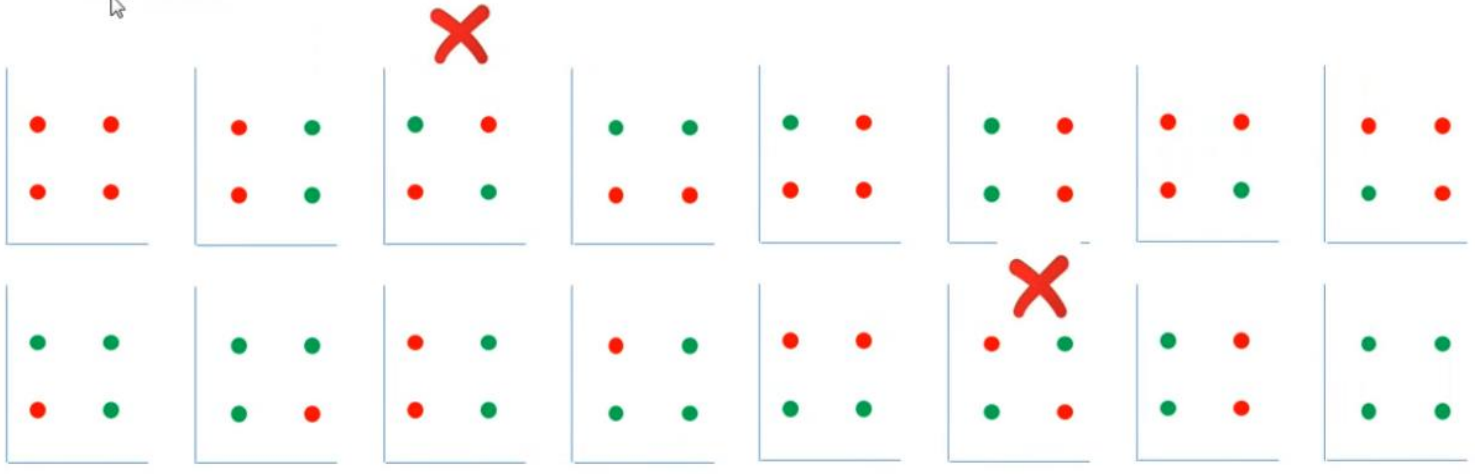
Can a line hypothesis class shatter these 3 data points?



If you can find even a single dataset of N data points for which H can classify positive items from negative for each problems.

The VC-Dimension

Total data points =4



We cant find a dataset of 4 points which can be shattered by line class.

The VC-Dimension

VC dimension of a hypothesis class H is the maximum number of data points which can be shattered by H.

VC dimension of line class is 3.

Thank You!