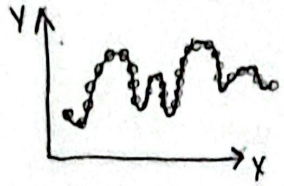
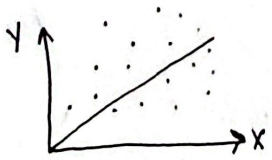
Overfitting and underfitting



- let graph represent f(x) / model or hypothesis; simply o/p of training.
- model as a opp from training phase
- points : datapoints or observation points
- when model tries to cover every datapoints in a x-y plane, then it is called overfitting.
- complex



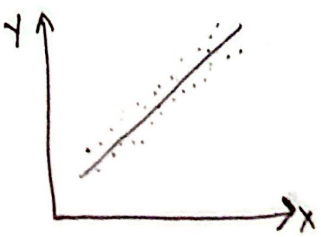
- the line generated doesn't have anything to do with datapoints it is completely independent.
- few datapoints near or on line. So, underfit.

Best fit

* If we train model with less number of features / attributes then model would not be able to distinguish betⁿ ball and orange. with just providing shape feature. - underfitting

* when we provide multiple feature eg: shape, play, eat radius=5. to determine whether the given object is ball or not. it ~~beds~~ leads to overfitting.

Principal Component Analysis

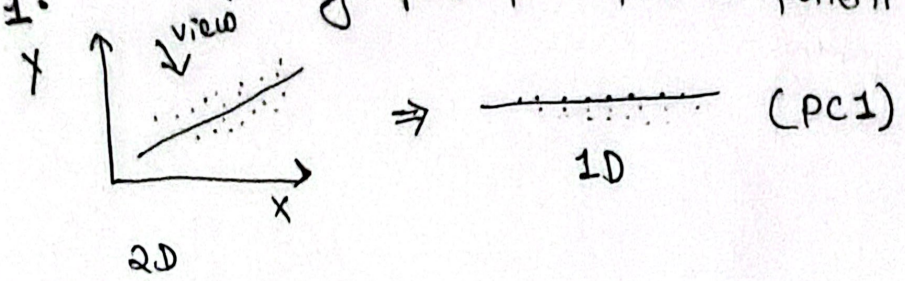


overfitting

- PCA tries to reduce overfitting problem
- To reduce this, PCA tries to convert the high dimentionality to low dimentionality

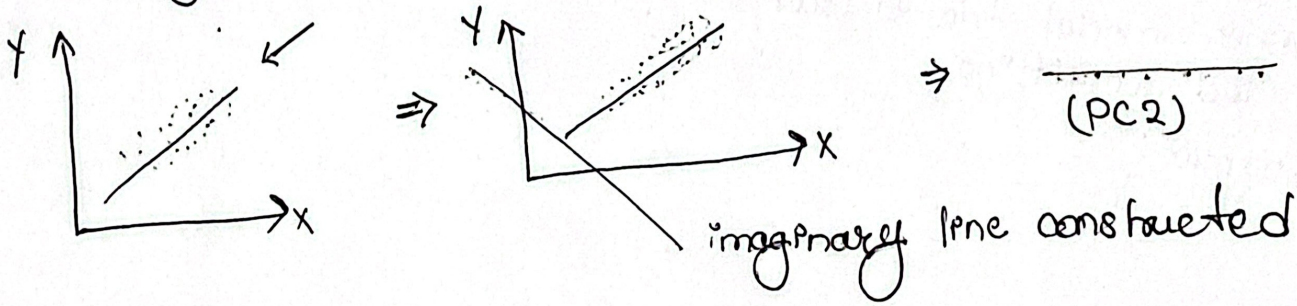
If Overfitting: find principal Component Analysis

1.



when you view this model graph from above, all the datapoints are mapped on the model line. as a result complexity minimized

2. when you view the data points from another direction.



Number of principal components $\leq$ the number of attributes given in a data

priority: PC1 > PC2 > PC3 > PC4 .....

principal components should have orthagonal properties. in between two principal components there should exist orthagonal properties, that is, PC1 & PC2 should be independent of eachother.

Numerical

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	0.9
$\bar{X}$	$\bar{Y}$
= 1.81	1.91

Attributes = 2 ; No. of PCA ≤ no. of attributes.

1. find the mean of attribute x & y.
2. find covariance matrix with 2 attributes

$$C = \begin{bmatrix} cov(x,x) & cov(x,y) \\ cov(y,x) & cov(y,y) \end{bmatrix}$$

$$Cov(x,y) = \sum_{i=1}^{n=10} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

ⓐ Covariance of (x,x)

$$Cov(x,x) = \sum_{i=1}^{10} \frac{(x-\bar{x})(x-\bar{x})}{n-1}$$

i) $i = 1 \cdots 10$

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

$\Rightarrow Cov(x,x) = \frac{5.5490}{9} = 0.6165$

ⓑ Covariance of (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

$Cov(y,y) = \frac{6.449}{9} = 0.7165$

ⓒ Covariance of (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
2.2	2.9	0.39	0.99	0.3861

$Cov(x,y) = \frac{5.5390}{9} = 0.6154$

3. $$C = \begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix}$$

4. Find Eigen value : total amount of variance that can be explained by a given principal component.

$C - \lambda I = 0$ ∴ I - identity matrix

λ - Eigen value to be found

$$\begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 0$$

$$\begin{bmatrix} 0.6165 - \lambda & 0.6154 \\ 0.6154 & 0.7165 - \lambda \end{bmatrix} = 0 \quad \text{(take determinant)}$$

$$\lambda^2 - 1.333\lambda + 0.0630 = 0$$

$\lambda_1 = 0.0490$

$\lambda_2 = 1.2840$

For each eigon value, one eigon ~~ma~~ vector should be found.

$CV = \lambda V$ V - Eigon vector

$$\begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix} * \begin{bmatrix} x_1 \\ Y_1 \end{bmatrix} = 0.0490 \begin{bmatrix} x_1 \\ Y_1 \end{bmatrix} \quad \text{For } \lambda 1$$

$0.6165x_1 + 0.6154Y_1 = 0.0490x_1$

$0.6154x_1 + 0.7165Y_1 = 0.049Y_1$

$\Rightarrow 0.5674x_1 = -0.6154Y_1$

$\Rightarrow 0.6154x_1 = -0.6674Y_1$

$X_1 = -1.0845\, Y_1$

$$\begin{bmatrix} -1.0845 \\ 1 \end{bmatrix} = 1.17614 + 1$$

$$= 2.27614$$

Squaring root

$$\sqrt{2.27614} = 1.47517$$

Now,

$$\frac{-1.0845}{1.47517} = -0.7351$$

Similarly, $\frac{1}{1.47517} = 0.6778$

So,

For $\lambda_1 = \begin{bmatrix} -0.7351 \\ 0.6778 \end{bmatrix}$

Similarly for λ_2:

$$CV = \lambda V$$

$$\begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix} \begin{bmatrix} X_2 \\ Y_2 \end{bmatrix} = \begin{bmatrix} X_2 \\ Y_2 \end{bmatrix} * 1.2840$$

$$\begin{bmatrix} 0.6165X_2 + 0.6154Y_2 = 1.2840X_2 \\ 0.6154X_2 + 0.7165Y_2 = 1.2840Y_2 \end{bmatrix}$$

$-0.6675X_2 + 0.6154Y_2 = 0$

CaseI $0.6154Y_2 = 0.6675X_2$

Similarly,

CaseII $0.7165Y_2 = 0.6686X_2$

CaseI

$X_2 = 0.92194\, Y_2$

$$\begin{bmatrix} 0.92194 \\ 1 \end{bmatrix}$$

$= (0.92194)^2 + (1)^2$

$= 0.8499 + 1$

$= \sqrt{1.8499} = 1.3601$

Now, divide

$$\frac{0.92194}{1.3601} = 0.6778$$

Similarly, $\frac{1}{1.3601} = 0.7351$

$$\lambda_2 = \begin{bmatrix} 0.6778 \\ 0.7351 \end{bmatrix}$$