→ Dimensionality Reduction →

① PCA [Principal Component Analysis]

Consider

→

weight	Height	DBP	SBP	Health?

Diastolic BP

Systolic B.P.

Say Health of Person depends on 4 features

- weight ✓
- Height ✓
- DBP ✓
- SBP ✓

To represent this data ⇒ We need 4 Dimension

For more features ⟶ We need more Dimension.

* Visualizing data in More than 3 dimension is difficult.

* Computation on 4 features is also Complex

Possible Solution →

Height	weight	DBP	SBP	? Health.

[Body Mass Index] BMI

BP [Blood Preserve]

① look for Strong Correlated features: Here Height & weights } are Strongly correlated

Also DBP and SBP }

... effective Correlation of Height & Weight ⇒ BMI

→ We can combine the effective correlation of Height & Weight ⇒ BMI

" " " " " " " DBP and SBP ⇒ BP.

Now Instead of 4 columns we have 2 columns.

BMI	BP.

→ This is 2-D data. ✓ Easy to Compute
✓ Easy to Visualize.

We have Reduced the Dimension of data from 4-D to 2-D.

Note → Consider 2 columns:→

Let → X_1 (DBP), X_2 (SBP)

X_1 DBP	X_2 SBP
78	126
80	128
81	127
82	130
84	130
86	132

$y = \alpha_1 X_1 + \alpha_2 X_2$

(BP)

α_1 & α_2 are weights of features

X_1 and X_2 are features

Ex

→ ① $BP = 0.8\,DBP + 0.6\,SBP$

We are giving more weights to DBP as compared to SBP.

Ex ② Let BP = mean of DBP & SBP.

$BP = \frac{DBP + SBP}{2} = 0.5\,DBP + 0.5\,SBP$

$\alpha_1 = 0.5$

$\alpha_2 = 0.5$

for mean

DBP	SBP	BP mean
78	126	102
80	128	104
81	127	104
82	130	106
84	130	107
86	132	109

[illegible] BP = [illegible] and SBP.

82	150	107
84	130	109
86	132	

Ex (3) let BP = Sum of DBP and SBP.

$BP = DBP + SBP \Rightarrow \alpha_1 = 1$

$\alpha_2 = 1$

For Sum

DBP	SBP	Sum
78	126	204
80	128	208
81	127	208
82	130	212
84	130	214
86	132	218

Now PCA [⇒ Principal Component Analysis ✓

⇒ It is a method to find the Linear Combination that accounts for as much variability as possible in Combined Variable ↳(Maximum Variance)]

[To understand ⇒ $y = \alpha_1 x_1 + \alpha_2 x_2$

we want value of α_1 and α_2 such that the variance in the value of (y) should be Maximum

B'coz more the Variance → More is the information]

Note:⇒ for $y = \alpha_1 x_1 + \alpha_2 x_2$ (combined variable)

→ Here To maximize variance large value of α_1 and α_2 can be proposed

→ We will place restriction on value of weights i'e $(\alpha_1, \alpha_2 \ldots)$ such that

$$\alpha_1^2 + \alpha_2^2 + \ldots + \alpha_n^2 = 1$$

* If there are n features, we will need 'n' weights

In our Ex ⇒ we had 2 features ∴ we need α_1 and α_2

$$\therefore BP = \alpha_1 DBP + \alpha_2 SBP$$

constraints $\alpha_1^2 + \alpha_2^2 = 1$

but $\alpha_1 = 0.8$ $(0.8)^2 + (0.6)^2 = 0.64 + 0.36 = 1.00$ ✓

Let $\alpha_1 = 0.8$

$\alpha_2 = 0.6$

$(0.8)^2 + (0.6)^2 = 0.64 + 0.36 = 1.00$ ✓

Let Let $\alpha_1 = 0.8$ and $\alpha_2 = 0.6$

$BP = \alpha_1 DBP + \alpha_2 SBP$

DBP	SBP	BP
78	126	138
80	128	140.8
81	127	141.0
82	130	143.6
84	130	145.2
86	132	148.0
	Mean	142.8

$$\text{Variance} = \frac{1}{n-1} \sum_{i=1}^{n} (BP - \text{mean } BP)^2$$

$$= \frac{1}{5} \left((138 - 142.8)^2 + (140.8 - 142.8)^2 + (141.0 - 142.8)^2 + \dots \right)$$

$$= 12.74$$

Let for above 6 input Sample.

* we can take Many values of α_1 and α_2 Such that $\alpha_1^2 + \alpha_2^2 = 1$
=> and calculate variance for Each α_1 and α_2

α_1	α_2	Var(y)
0.8	0.6	12.74
0.6	0.8	11.8
0.98	0.2	10.4
0.2	0.98	7.4

Of the above 4 Cases: we see for $\alpha_1 = 0.8$ and $\alpha_2 = 0.6$ we get Largest Variance = 12.74

Taxika 1 :- $BP = 0.8\,DBP + 0.6\,SBP$

To find α_1 and α_2 we got over α_1 and α_2

We got our a_1 and a_2

Taarika 2 →

Note → Consider again →

DBP	SBP
78	126
80	128
81	127
82	130
84	130
86	132
x	y

Construct Covariance Matrix $=$

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

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

For above Sample data set

Covariance matrix ⇒

	DBP	SBP
DBP	8.17	5.97
SBP	5.97	4.97

Eigen values = from above Covariance Matrix.

$$\begin{bmatrix} -0.8 \\ -0.6 \end{bmatrix} \checkmark \checkmark$$

$\therefore$ $BP = -0.8\,DBP - 0.6\,SBP$

$\Rightarrow$ $BP = 0.8\,DBP + 0.6\,SBP$ ✓

Applications →

(1) Consider

DBP	SBP	weight	Height

BP BMI

$\Rightarrow$

BP	BMI

Reduced Dimension by Combining features.

BP BMI Reduced Dimension "Combining features.

(2)

DBP	SBP	weight	Height	Pulse
78	126	82	5.1	90
80	128	74	5.4	80
81	127	64	5.11	70
82	130	82	6.2	90
84	130	92	6.3	95
86	132	85	5.10	94

Pat1
Pat2

Here if we have many no of features and we need to compare the Health chart/ parameters of samples, it is difficult

But if we combine the above 5 col in a col.

C1	C2
V1	V2
V3	V4
V5	V6

Now with less no of features comparing the Health chart/ parameters of samples is relatively Easy:

* PCA → It is Dimensionality Reduction Technique.

Consider: For House Price Prediction

Size	Location	YOC	OC Y/N	Builder	Color	Garden	Swimming	Metro Distance	Price

* Here Price of House depends on above features.
* Here Many No of features (Problem of Plenty).
* Issues:
  * Represent" Problem ✓
  * Overfitting Problem ✓
* We need to Reduce Dimension
* We need to Identify Important Components.

Dimensionality Reduction:-

* Reduces the dimension of feature space

  Ex: if there are 100 features/col in dataset and you want to get only 10 features then with dimensionality reduction technique we can achieve this.

* It transforms dataset which is in n dimension space to n' dimension space where $n' < n$.

to n' dimension space where

Why Dimensionality Reduction =>

- Normally it is argued that many features gives More accurate result
- But after some point the model performance decreases (overfitting) with increase in number of features.
- This is Curse of Dimensionality
- So Dimension Reduction is Crucial.

* PCA enables us to identify the correlation and pattern in the dataset so that it can be transformed into new dataset with lower dimension without loss of important information.