Date ....

Dimensionality Reduction:-

Dimension - features

High dimensionality data

→ Training Time ↑

→ Computational Resources requirement ↑

→ chances of overfitting.

→ Visualualization (EDA) is difficult.

→ Most of the variables will be correlated.

Dimensionality Reduction:-

→ The process of reducing dimensions (features).

Principal Component Analysis (PCA)

→ Unsupervised Algorithm.

→ feature extraction technique.

Feature selection :- (subset of original features)

The process of selecting the most important feature using any of FS techniques (wrapper method, filter method, Embedded method) you can remove the irrelevant features.

Feature Extraction :- (Create new component)

Combine the existing features to create new components.

PCA :-

Extract / obtain the important data in the form of components.

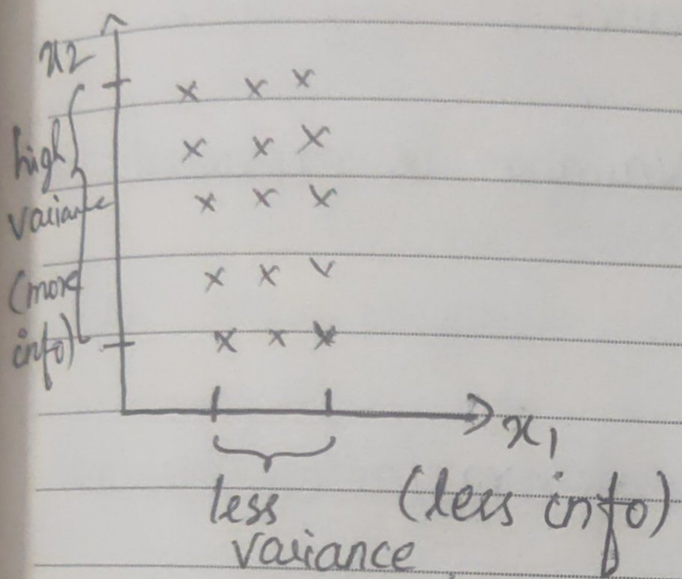
↓

Principal Components.

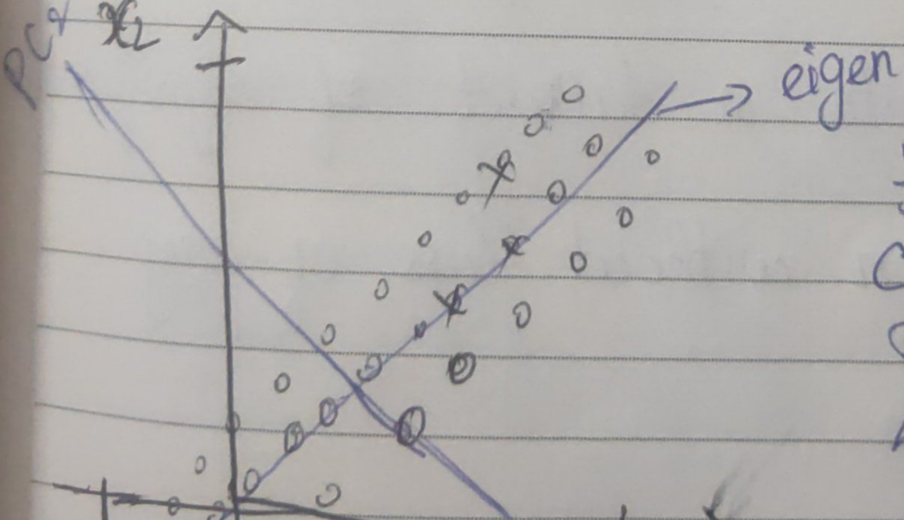
Date ....

Principal Component :-

Combination of original dimensions which has explained variance Ratio.



Select any 1 feature x_1 or x_2?
$x_1 \rightarrow$ dropped.
$x_2 \rightarrow$ bcoz, it has more information



From this we cannot select the feature, so here we can use PCA technique

Each PC is orthogonal/ to the first PC (each other) $\rightarrow$ perpendicular.

Date ....

If we have, 10 features → 10 principal Components.

PC_1, PC_2, PC_3 ..... PC_{10}
80% ↑ 15% ↑ 3% ↑ 2% ↑

PC_1 → Explained Variance (more Variance) → assume 80%

PC_1, PC_2, → Variance? → 95%.

PC_1, PC_2, PC_3 → Variance is explained → 98%.

Steps:-

→ PCA identifies the Correlation / pattern in the dataset so that it can be transformed into a dataset of significantly lower dimension without loss of any imp information.

PC_1 → most significant Component
PC_2 → second most " "
PC_3 → Third most " "

steps:

① → scale the data (PCA tries to get the features with maxi variance, the variance will be high for higher vari magnitude feature. so scale the data).

② Calculate the Covariance.

→ to understand the variables that are highly correlated.

③ Calculates eigen values & eigen vector:-

→ Computed by Co-variance.

Eigen vector - Determine in direction of new feature space.

Eigen Values - determines the magnitude (scalar of the eigen vectors)

→ This tells how the dataset is ~~sprad~~ spread out on the eigen vector.

④ Sort - the most significant component.

⑤ Remove the PCs that contains least information.

For eg:-

Let 3 features (x_1, x_2, x_3)

3PCs Variance of $PC_1 = 40$
Variance of $PC_2 = 20$
Variance of $PC_3 = 5$

Total Variance = 65 //.

How much Variance is explained by PCs?

$$\left.\begin{array}{r}\text{Explained Variance} \\ \text{ratio (EVR)}\end{array}\right\} = \frac{\text{explained Variance}}{\text{Total variance}}$$

$$\text{EVR of } PC_1 = \frac{40}{65} = 0.61$$

$$\text{EVR of } PC_2 = \frac{20}{65} = 0.31$$

Date ....

EVR of $PC_3 = \frac{5}{65} = 0.08$.

$PC_1 \rightarrow 61\%$ (0.61)
$PC_1 + PC_2 = 92\%$ (0.92)

Scree plot → Used to find the Optimal no. of PCs to be Considered.

Pros:-
→ Correlate features are removed
→ Model training time is reduced.
→ Overfitting is reduced.
→ Ability to handle noise

Cons:-
→ The resultant PC are less interpretable than the original data.
→ Can lead to information loss, if explained variance threshold is not considered appropriately.