Principal Component Analysis (PCA)

PCA is a dimensionality Reduction technique used in machine learning to reduce the number of features / variables in a dataset while retaining as much variance (information) as possible. PCA achieves this by identifying the directions (principal components) along which the variation in data is maximal.

* Why use PCA?

→ Many datasets have large number of features, which can lead to problems like overfitting in ML models. PCA helps by reducing the number of features while preserving the structure of the data.

→ It can filter out noise by discarding the components that contribute the least to data variance.

→ High dimensional data can be difficult to visualize. By reducing data to 2 or 3 dimensions, PCA enables easier visualization.

→ PCA makes modeling simpler, thus reduces computational cost and can improve model interpretability.

Key Concepts in PCA?

* **Variance:** PCA is based on the idea that the direction in which the data varies the most (where variance is high) are the most important. The goal is to find these directions and reduce dimensions along them.

* **Principal Components:** These are the new axes (directions) that PCA computes, which represents the directions of maximum variance in the data. The first principal component captures the largest variance, the second captures the second largest variance, and so on. These components are orthogonal (perpendicular) to each other, ensuring they capture independent aspects of the variance in the data.

* **Covariance Matrix:** The covariance matrix captures the relationship (covariance) between pairs of variables in the data. The eigenvectors of the covariance matrix represent the direction of maximum variance in the data, and the eigen values quantify how much variance is explained by each eigen vector.
* **Eigenvectors and Eigenvalues:** Eigen vectors represent the direction of maximum variance and these are also the direction of principal components. Eigenvalues indicate the magnitude (or importance) of the principal components. Larger eigenvalues correspond to principal components that explain more variance in the data.

→ Steps in PCA

1. Standardize the data → PCA is sensitive to scale of data, so it's common to standardize the data by subtracting the mean and dividing by standard deviation, ensuring that all features have equal importance.
2. Calculation of covariance matrix → Compute the covariance matrix to understand how features in the dataset are related. The covariance matrix helps in identifying the directions where data varies the most.
3. Compute Eigenvectors and Eigenvalues → Find the eigenvectors and eigenvalues of covariance matrix.
4. Select Principal Components → Sort the eigenvalues in descending order and choose the top k components that explain the majority of the variance. The value of k is typically chosen based on cumulative variance explained by the components (e.g., we might choose k such that 95% variance is retained).
5. Project the data → Once the principal components are chosen, the data is projected onto this lower dimensional

subspace. The transformed data is a compressed representation in terms of the selected principal components.

Applications of PCA:

1. Image Compression: PCA can be used to reduce the number of features in image data, making storage more efficient while retaining the essential features of the image.
2. Noise Filtering: By discarding principal components with low variance, PCA can filter out noise, keeping only the significant data.
3. Exploratory Data Analysis: By reducing the dimensionality, PCA allows for easier exploration of data pattern in fewer dimensions.
4. Preprocessing in ML: Before training a ML model, PCA is often used to reduce dimensionality, which can improve performance and speed.

Limitations of PCA:

1. PCA assumes that directions of maximum variations are linear. It may not perform well if the true underlying structure of the data is non-linear.
2. The principal components of linear combinations of the original variables, which can make them hard to interpret in terms of original features.
3. While PCA tries to retain as much variance as possible some information may still be lost, especially if too few principal components are selected.

→ **How to choose number of Principal Components:** A common approach is to plot the **explained variance ratio**, which shows how much variance each principal component explains. We can choose the number of components that cumulatively explains a desired amount of variance (usually 90% to 95%).