Data Reduction or Dimensionality Reduction :-

Dimensionality reduction involves deriving new features from old ones.

Generally this is done by applying transforms to the dataset, which change the axes (cordinate system) of the graph by moving and rotating them. This can be written as a matrix that we apply to the data. It enables us to combine features and identify useful data.

① Principal Component Analysis (PCA) →

PCA is a useful statistical technique and is often ~~useful~~ used in finding patterns in high dimensional datasets.

Some Background Maths :-

① Std. Deviation : Measure of the spread of data.

$$s = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n-1}}$$

Eg1: $x = [0, 8, 12, 20]$

Ans → $s = 8.3266$

Eg2: $x = [8, 9, 11, 12]$

Ans → 1.8257

Eg1 has larger SD as compared to Eg2 because eg1 data has larger spread than the second.

(2) Variance: Variance is another measure of spread of data in a dataset. It is almost identical to SD.

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

(3) Covariance: Mean & SD are one-dimensional calculations.

Covariance is always measured between two dimensions.

If we calculate covariance b/w one dimension and itself, we get variance.

$$\text{cov}(x, y) = \frac{\sum_{i=1}^{n} (x-\bar{x})(y-\bar{y})}{n-1}$$

If the covariance answer is:

positive → Means both dimensions increase together.
neg. → one dimension ↑ and other ↓.

zero → Two dimensions are independent of each other.

Covariance Matrix: Covariance is always measured between two dimensions.

If we have a dataset with more than two dimensions, then more than one covariance measurem[ent] can be calculated.

Parshvanath Charitable Trust's
A. P. SHAH INSTITUTE OF TECHNOLOGY
(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

Eg: for 3 dimensional (x, y, z) dataset

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

* Down the main diagonal, values are the variances for that dimension.

* Since $cov(x,y) = cov(y,x)$, the matrix is symmetrical about the main diagonal.

4 Eigen Values and Eigen Vectors:

Let A be an $n \times n$ matrix. A scalar λ is called Eigen Value of A if there is a nonzero vector x such that

$$Ax = \lambda x$$

Such a vector x is called eigen vector of A corresponding to λ.

Finding Eigen Values:

To find eigen values of A, we have to find the values of λ which satisfy the charastic eq^n

$$\det(A - \lambda I) = 0 \qquad I = \text{Identity matrix}$$

1. form the matrix $A - \lambda I$.
2. calculate $\det|A - \lambda I|$:
3. find sol^n to $\det|A - \lambda I| = 0$