ASSIGNMENT: 9

Name: FALAK KHAN
Batch: MACHINE LEARNING AND AI BATCH A3

1)What is the difference between Normalization and Standardization?

Ans:: Standardization is a scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

2) Explain your understanding of n_components in PCA().

Ans: n_components: number of principal components to identify.

n_components should be equal to the features which contribute a large number to the overall variance. The number depends on the business logic.

3) Explain your understanding of explained_variance_ratio.

Ans: explained_variance_ratio_ = explained_variance_ / np.sum(explained_variance_)

explained_variance_ratio_ : array, shape (n_components,) Percentage of variance explained by each of the selected components.

If n_components is not set then all components are stored and the sum of the ratios is equal to 1.0.

explained_variance_: array, shape (n_components,) The amount of variance explained by each of the selected components.
Equal to n_components largest eigenvalues of the covariance matrix of X.

4)Explain your understanding of Pearson, Kendall, and Spearman's method of calculating variance in PCA.

Ans: Pearson r correlation: Pearson r correlation is the most widely used correlation statistic to measure the degree of the relationship between linearly related variables. For example, in the stock market, if we want to measure how two stocks are related to each other, Pearson r correlation is used to measure the degree of relationship between the two. The point-biserial correlation is conducted with the Pearson correlation formula except that one of the variables is dichotomous. The following formula is used to calculate the Pearson r correlation:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

r_{xy} = Pearson r correlation coefficient between x and y
n = number of observations
x_i = value of x (for ith observation)
y_i = value of y (for ith observation)

Kendall rank correlation: Kendall rank correlation is a non-parametric test that measures the strength of dependence between two variables. If we consider two samples, a and b, where each sample size is n, we know that the total number of pairings with a b is n(n-1)/2. The following formula is used to calculate the value of Kendall rank correlation:

Nc = number of concordant
Nd= number of discordant

Spearman rank correlation: Spearman rank correlation is a non-parametric test that is used to measure the degree of association between two variables. The Spearman rank correlation test does not carry any assumptions about the distribution of the data and is measured on a scale that is at least ordinal.
The following formula is used o calculate the Spearman rank correlation:

ρ = Spearman rank correlation
di = the difference between the ranks of corresponding variables
N= number of observations