chapter:-4

Unsupervised learning

Clustering

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. It is an unsupervised learning technique.

→ Cluster - a collection of data object that are similar to one another within the same cluster and are dissimilar to the object in other clusters.

→ clustering can also be used for outlier detection where outliers may be more interesting than common cases.

→ Applications of outlier detection include the detection of credit card fraud and the monitoring of criminal activities in electronic commerce.

→ For example, exceptional cases in credit card transaction, such as very expensive and frequent purchases, may be of interest as possible fraudulent activity.

Similarity and Dissimilarity

Distance measures are used in order to find similarity or dissimilarity between data objects.

→ The most popular distance measure is Euclidean distance, which is defined as:

$$d(x,y) = \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$$ where $x=(x_1,y_1)$
$y=(x_2,y_2)$

Another well known metric is Manhattan distance, defined as:-

$$d(x,y) = |x_2-x_1|+|y_2-y_1|$$

Minkowski distance is a generalization of both Euclidean distance and Manhatten distance. It is defined as:

$$d(x,y) = \left(|x_2-x_1|^p + |y_2-y_1|^p\right)^{1/p}$$

where, p is a positive integer, such a distance is also called Lp norm, in some literature. It represents the Manhattan distance when p=1 (i.e L1 norm) and Euclidean distance

when $p = 2$ (i.e L_2 norm)

Categories of Clustering Algorithms:-

Many clustering algorithms exist in the literature. The major clustering methods can be classified into the following categories.

Partitioning Methods:-

- Given a database of n objects or data tuples, a partitioning method constructs k partitions of the data, where each partition represents a cluster and $k \leq n$.

Given k, the number of partitions to construct, a partitioning method creates an initial partitioning. ~~a partitioning method~~ It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another.

Hierarchical Method:-

A hierarchical method creates a hierarchical decomposition of the given set of data objects.
A hierarchical method can be classified as being agglomerative or divisive

Agglomerative Approach:-

⇒ The agglomerative approach follows the bottom-up approach. It starts with each object forming a seperate group. It successively merges the objects or groups that are close to one another, until a termination condition holds.

Divisive approach

⇒ The divisive approach follows the top-down approach. It starts with all of the objects in the same cluster. In each successive iteration, a cluster is split up into smaller cluster, until a termination condition hold

Density-based Method

- Most partitioning method cluster objects based on the distance between objects. Such method can find only spherical-shaped clusters and encounter difficulty at discovering clusters of arbitrary shaped.
- Other clustering method have been developed based on the notion of density

Model-based Methods:

→ Model-based methods hypothesize a model for each of the clusters and find the best fit of the data to the given model

- EM is an algorithm that performs expectation maximization analysis based on statistical modeling.

K-Mean Algorithm:-

K-Mean is the simplest partitioning based clustering algorithm.

The main idea is to define k centers, one for each cluster. These centers should be selected cleverly because of different location causes different results.

K-Algorithms:

1. Initialize cluster centroids $\mu_1, \mu_2, \mu_3, \ldots \mu_k$ belonging to data points randomly.
2. Repeat until Convergence.

{

for every i, set

$$c^{(i)} = \arg\min_j \|x^{(i)} - \mu_j\|^2$$

for every j, set

$$\mu_j = \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}$$

K-Mean Algorithm:

1. Specify number of clusters K.
2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids i.e assignment of data points to clusters isn't changing.

- Compute the sum of Squared distance between data points and all centroids.
- Assign each data point to the closest cluster
- Compute the centroid for the clusters by taking the ~~averg~~ average of the all data points that belong to each cluster.

The approach k-mean follows to solve the problem is called Expectation-Maximization

Example:-

Divide the data points {(2,10), (2,5), (8,4), (5,8), (7,5), (6,4)} into two clusters

Solution:

Let; $P_1 = (2,10)$ $P_2 = (2,5)$ $P_3 = (8,4)$
$P_4 = (5,8)$ $P_5 = (7,5)$ $P_6 = (6,4)$

Initial step

choose cluster centers randomly

Let;

$M_1 = (2,5)$ & $M_2 = (6,4)$ are two initial cluster centers.

Iteration 1

calculate distance between cluster centers and each data points.

$d(M_1, P_1) = 5$
$d(M_1, P_2) = 0$
$d(M_1, P_3) = 6.08$
$d(M_1, P_4) = 4.24$
$d(M_1, P_5) = 5$
$d(M_1, P_6) = 4.12$

$d(M_2, P_1) = 7.21$
$d(M_2, P_2) = 4.12$
$d(M_2, P_3) = 2$
$d(M_2, P_4) = 4.12$
$d(M_2, P_5) = 1.41$
$d(M_2, P_6) = 0$

Thus,

cluster 1 = {P1, P2}

cluster 2 = {P3, P4, P5, P6}

Iteration 2

New cluster centers:

$M_1 = (2, 7.5)$ $M_2 = (6.5, 5.25)$

$d(M_1, P_1) = 2.5$	$d(M_2, P_1) = 6.54$
$d(M_1, P_2) = 2.5$	$d(M_2, P_2) = 4.51$
$d(M_1, P_3) = 6.95$	$d(M_2, P_3) = 1.95$
$d(M_1, P_4) = 3.04$	$d(M_2, P_4) = 3.13$
$d(M_1, P_5) = 4.59$	$d(M_2, P_5) = 0.56$
$d(M_1, P_6) = 5.32$	$d(M_2, P_6) = 1.35$

cluster 1 = {P1, P2, P4} cluster 2 = {P3, P5, P6}

Iteration 3

New cluster Centers

$M_1 = (3, 7.67)$ $M_2 = (7, 4.33)$

$d(M_1, P_1) = 2.54$	$d(M_2, P_1) = 7.56$
$d(M_1, P_2) = 2.85$	$d(M_2, P_2) = 5.04$
$d(M_1, P_3) = 6.2$	$d(M_2, P_3) = 1.05$
$d(M_1, P_4) = 2.03$	$d(M_2, P_4) = 4.18$
$d(M_1, P_5) = 4.81$	$d(M_2, P_5) = 0.67$

CamScanner

$d(M_1, P_6) = 4.74$ $d(M_2, P_6) = 1.05$

Thus,

Cluster 1 = {P_1, P_2, P_4}

cluster 2 = {P_3, P_5, P_6}

Since, No data points are re-assigned

final clusters are :

cluster 1 = {P_1, P_2, P_4} cluster 2 = {P_3, P_5, P_6}

K-Means ++ Algorithm :

Randomization of picking k cluster centers in k-means algorithm results in the problem of initialization sensitivity.

This problem tends to affect the final formed clusters. The final formed clusters depend on how initial cluster centers were picked.

To overcome the above-mentioned drawback we use k-means ++. This algorithm ensures a smarter initialization of the cluster centers and improve the quality of the clustering.

Initialization Algorithm

1. Randomly select the first cluster center from the data points.
2. For each data point compute its distance from the nearest previously chosen cluster center.
3. Select the next cluster center from the data points such that the probability of choosing a point as cluster center is directly proportional to its distance from the nearest previously chosen cluster center.
4. Repeat step 2 and 3 until k cluster centers have been sampled.

Example :-

Consider the data points {(2,10), (2,5), (8,4), (5,8), (7,5), (6,4), (3,2), (4,6)}? Select three cluster center using K-mean ++ algorithm. Select a data point with highest probability as next cluster center.

Solution :-

Let $P_1 = (2,10)$ $P_2 = (2,5)$ $P_3 = (8,4)$
$P_4 = (5,8)$ $P_5 = (7,5)$ $P_6 = (6,4)$
$P_7 = (3,2)$ $P_8 = (4,6)$

Select the first cluster center randomly
Let $\mu_1 = (2,5)$

$d(\mu_1, P_1) = 5$
$d(\mu_1, P_2) = 0$
$d(\mu_1, P_3) = 6.08$
$d(\mu_1, P_4) = 4.24$
$d(\mu_1, P_5) = 5$
$d(\mu_1, P_6) = 4.12$
$d(\mu_1, P_7) = 3.6$
$d(\mu_1, P_8) = 2.12$

Thus probability of data point being selected as next cluster center is directly proportional to

{5, 0, 6.08, 4.24, 5, 4.12, 3.6, 2.12}

The data point $P_3 = (8, 4)$ has the highest probability of being selected as cluster center.

Thus, $\mu_2 = P_3 = (8, 4)$

Again, for each data point compute its distance from the nearest cluster center.

$\min(d(\mu_1, P_1), d(\mu_2, P_1)) = 5$

$\min(d(\mu_1, P_3), d(\mu_2, P_3)) = 0$

$\min(d(\mu_1, P_2), d(\mu_2, P_2)) = 0$

$\min(d(\mu_1, P_4), d(\mu_2, P_4)) = 4.24$

$\min(d(\mu_1, P_5), d(\mu_2, P_5)) = 5$

$\min(d(\mu_1, P_6), d(\mu_2, P_6)) = 4.12$

$\min(d(\mu_1, P_7), d(\mu_2, P_7)) = 5.39$

$\min(d(\mu_1, P_8), d(\mu_2, P_8)) = 4.47$

The data point $P_7 = (3, 2)$ has the highest probability of being selected as next cluster center.

Thus, three cluster centers are $\mu_1 = (2, 5)$ $\mu_2 = (8, 4)$ $\mu_3 = (3, 2)$.

Gaussian Mixture models and EM

If we model data points using mixture of two or more components where each component is modeled as a Gaussian distribution, it is called a Gaussian mixture model.

Partition, hierarchical, and density based clustering methods are hard clustering algorithms. This means these algorithms assign every data object to exactly one cluster.

Gaussian Mixture Model (GMM) is soft clustering algorithm. It may assign data object to multiple clusters with some probability.

For example, if $k=3$, a data point can be assigned to cluster $C1$ and $C2$ with probability 0.7 and 0.3 respectively.

Gaussian Mixture model (GMMs) assume that there are a certain number of Gaussian distribution and each of these distribution represent a cluster.

Hence, a Gaussian Mixture model tends to group the data points belonging to a single distribution together.

For a given set of data points, our GMM would identify the probability of each data point belonging to each of these distributions

Outline of EM Algorithm

1. Initialize the μ_k's, σ_k's and p_k's and evaluate the log-likelihood of these parameters.

2. E-step:- Evaluate the posterior probabilities $r_{zi}(k)$ using the current values of the μ_k's and σ_k's

3. M-step:- Estimate new parameters. μ_k, σ_k and p_k with the current value of ~~the μ_k's and σ_k's.~~ of $r_{zi}(k)$

4. Evaluate the log-likelihood of new μ_k's, σ_k's and p_k's.

5. If the log-likelihood is changed by less than some small ϵ stop. otherwise go back to step 2.

EM for GMM

1. Suppose we are given a set of m data observations $\{x_1, x_2, \ldots, x_m\}$ of a numerical variable x. Let x be a mix of k normal distributions.

2. Now GMM is mixture of k probability distributions:

$$f(x) = p_1 f_1(x) + p_2 f_2(x) + \ldots + p_k f_k(x)$$

$$p_i \geq 0 \text{ and } p_1 + p_2 + \ldots + p_k = 1$$

$$f_i(x) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\left(\frac{(x-\mu_i)^2}{2\sigma_i^2}\right)}$$

3. Initialize means μ_i, variance σ_i^2 and mixing coefficient p_k

4. For each point x_n, calculate the probability that it belongs to cluster / distribution c_i
$i = 1, 2, \ldots, k$ and $n = 1, 2, \ldots, m$

$$\gamma_{in} = \frac{p_i f_i(x_n)}{\sum_{i=1}^{k} p_i f_i(x_n)}$$

5. Calculate N_i as below:

$$N_i = \gamma_{i1} + \gamma_{i2} + \cdots + \gamma_{im}$$

6. Recalculate parameter as below:

$$\mu_i = \frac{1}{N_i}(\gamma_{i1} x_1 + \cdots + \gamma_{im} x_m)$$

$$\sigma_i^2 = \frac{1}{N_i}(\gamma_{i1}(x_1 - \mu_i)^2 + \cdots + \gamma_{im}(x_m - \mu_i)^2)$$

$$p_i = \frac{N_i}{N}$$

7. Check for convergence. If converge stop otherwise repeat steps 4-6.

Example:

We have divide the data points {1, 2, 3, 6, 10, 11, 12} into two clusters using GMM. Estimate parameter upto 2 iterations.

Solution;

Let $\mu_1 = 1$ $\quad \mu_2 = 10$

$\sigma_1 = 0.8$ $\quad \sigma_2 = 0.8$

$P_1 = P_2 = 0.5$

Now;

	N=1	2	3	4	5	6	7	
i=1								$N_1 =$
i=2								$N_2 =$

For the calculation of above table

$$\gamma_{ij} = \frac{P_i f_i(x_j)}{\sum P_i f_i(x_j) + P_2 f_2(x_j)}$$

$$= \frac{0.5 \times 1}{}$$

$$f_i(x_j) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\left(\frac{(x_j - \mu_i)^2}{2\sigma_i^2}\right)}$$

$$f_1(1) = \frac{1}{0.8\sqrt{2\pi}} e^{-\left(\frac{(1-1)^2}{2\times(0.8)^2}\right)} = 0.4996$$

Text clustering

- Text clustering is the application of cluster analysis to text-based documents.

It uses machine learning and natural language processing (NLP) to understand and categorize unstructured textual data.

Normally following steps are followed in text clustering:
Prepare Bag of words, Remove stop words, Perform stemming, Calculate TF, calculate IDF, calculate TF-IDF, Use clustering algorithm.

→ Prepare Bag of Words:-

A bag of words is a representation of text that describes the occurance of words within a document.

We just keep track of word counts and disregard the grammatical details and the word order.

It is called a "bag" of words because any

information about the order or structure of words in the documents is discarded.

The model is only concerned with whether a word occur in the document, not where in the document.

→ Remove Stop word :-

Stop word removal is one of the most commo-nly used preprocessing steps across different NLP application. The idea is simply removing the words that occur commonly across all the documents in the corpus.

Typically, artical and pronouns are generally classified as stop words. Words such as "the", "a", "an", "of" etc are stop words.

These words have no significance in some of the NLP tasks like information retrieval, classifi-cation and clustering, which means these words are not very discriminatives

⇒ Perform Word Stemming

- The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.
  For instance:

  am, are, is ⇒ be

  car, cars, car's, cars' ⇒ car

  The result of this mapping of text will be something like:

  The boy's cars are different colors ⇒ the boy car be differ color.

⇒ Calculated TF-IDF

- Term frequency (TF): The term frequency is simply the number of occurances of a word in a specific document. The term frequencies are calculated per word and document.

$$TF(t,d) = \frac{\text{Number of occurances of Term } t \text{ in document } d}{\text{Maximum TF in } d}$$

⇒ Inverse Document Frequency (IDF): The inverse document frequency is a measure of whether a term is common or rare in a given document corpus. IDF is only calculated per word.

$$IDF(t) = \frac{\text{Number of documents}}{\text{Number of documents containing term } t.}$$

⇒ TF-IDF :- TF-IDF (term frequency - inverse document frequency) is a statistical measure that evaluates a word is to a document in a collection of documents

This is done by multiplying two metrics: Term frequency and Inverse Document Frequency

$$TF\text{-}IDF = TF * IDF$$

Once we have document vectors we can use any clustering algorithm to perform documents clustering

Example :-

Compute TF-IDF vectors for following documents:-

D1 : Computer Science and Information Technology
D2 : Computer Science and Computer Engineering
D3 : Humanities and social Science
D4 : Department of social Science.
D5 : Application of Computer Science

Solution :-

Bag of Words (BW) = { computer, Science, and, Information, Technology, Engineering, Humanities, Social, ~~Science~~, Department, of, Applications }

Removing stop Words we get:

W = { computer, Science, Information, Technology, Engineering, Humanities, Social, Department, Applications, Computers }

Perform Stemming we get;

W = { Computer, Science, Information, Technology, Engineer, Humanity, Social, Department, Application }

Calculate TF-IDF.

Term	TF(D_1)	TF(D_2)	TF(D_3)	TF(D_4)	TF(D_5)	IDF
Computer	1	1	0	0	0	$\log(5/3)=0.51$
Science	1	0.5	1	1	1	0.223
Information	1	0	0	0	0	0.916
Technology	1	0	0	0	0	0.916
Engineer	0	0.5	0	0	0	1.609
Humanity	0	0	1	0	0	1.609
Social	0	0	1	1	0	0.916
Department	0	0	0	1	0	1.609
Application	0	0	0	0	1	1.609

Term	TF-IDF(D_1)	TF-IDF(D_2)	TF-IDF(D_3)	TF-IDF(D_4)	TF-IDF(D_5)
Computer	0.51	0.51	0	0	0.51
Science	0.223	0.1115	0.223	0.223	0
Information	0.916	0	0	0	0.916
Technology	1.609	0.0045	0	0	0.916
Engineer	0	0.0045	0	0	0
Humanity	0	0	1.609	0	0
Social	0	0	0.916	0.916	0
Department	0	0	0	1.609	0
Application	0	0	0	0	1.609

Dimensionality Reduction:-

Increasing the number of features will not always improve classification accuracy.

In practice, the inclusion of more features might actually lead to worse performance.

The number of training example required increases exponentially with dimensionality d (i.e k^d)

Therefore, we need to choose an optimum set of features to improve classification accuracy.

Reducing the number of variables of a dataset naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity.

Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithm without extraneous variables to process.

Dimensionality reduction is the process of reducing the number of variables under consideration by obtaining smaller set of principle variables.

There are many ways to achieve dimensionality reduction but most of these techniques fall into one of two classes:

- Feature Elimination
- Feature Extraction:

Feature selection methods selects k dimension from set of d dimension in the original dataset

In feature Selection, we try to find a subset of the original set of variables, or features, to get a smaller subset which can be used to model the problem.

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \rightarrow Y = \begin{bmatrix} x_{i_1} \\ x_{i_2} \\ \vdots \\ x_{i_k} \end{bmatrix}$$

$k < N$

Feature Extraction

finds a set of new features through some mapping function $f()$ from the existing features. The mapping $f()$ could be linear or non linear.

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \xrightarrow{f(x)} y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_K \end{bmatrix} \quad K < N$$

The various feature extraction methods used for dimensionality reduction:

- Principle Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)
- Generalized Discriminant Analysis (GDA).

Principle Component Analysis (PCA)

PCA is unsupervised learning algorithm that is used for the dimensionality reduction in machine learning.

This method is often used to reduce the dimensionality of large data sets, by transforming a large set of variable into a smaller one that still contains most of the information in the large set.

- It is a statistical process that converts the observation of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation.

These new transformed features are called the principle components (PCs). Some properties of these principle components are given below:

- The PC must be the linear combination of the original features.
- These components are orthogonal, i.e, the correlation between the pair of variables is zero.
- The importance of each components decreases when going to 1 to n, it means the first PC has the most importance, and n^{th} PC will have the least importance.

Working of PCA

1. Get the Dataset with k features.

x_1	x_2	- - -	x_K
x_{11}	x_{21}	- -	x_{K1}
x_{12}	x_{22}	- - -	x_{K2}
:	:	- - -	:
:	:	- - -	:
x_{1n}	x_{2n}	- - -	x_{Kn}

2. Compute mean of Dataset :- Compute mean of each feature

$$\bar{x}_i = \frac{1}{n}(x_{i1} + x_{i2} + x_{i3} + \cdots + x_{in})$$

3. Compute correlation matrix of each feature pair

$$Cov(x_i, x_j) = \frac{1}{n-1}\sum_{K=1}^{n}(x_{iK} - \bar{x}_i)(x_{jK} - \bar{x}_j)$$

$$S = \begin{bmatrix} Cov(x_1,x_1) & Cov(x_1,x_2) & \cdots & Cov(x_1,x_K) \\ Cov(x_2,x_1) & Cov(x_2,x_2) & \cdots & Cov(x_2,x_K) \\ \vdots & \vdots & & \vdots \\ Cov(x_K,x_1) & Cov(x_K,x_2) & \cdots & Cov(x_K,x_K) \end{bmatrix}$$

4. Compute eigenvalues and normalized eigen vector of the covariance matrix.

- To find eigenvalues find solve the equation $|S - \lambda I| = 0$, we get n roots $\lambda_1, \lambda_2, \ldots \lambda_n$. which are eigen values.

- Then compute eigenvector (u) of each eigen-value λ_i, $i = 1, 2, \ldots n$ by solving $(S - \lambda_i I) u_i = 0$, where u_i is column vector of n unknowns.

- Finally normalize eigenvectors e_i by dividing by its length.

  Length of eigenvector is given by $|u|$

  $$= \sqrt{u_1^2 + u_2^2 + \ldots + u_n^2}$$

- The normalized eigenvector corresponding to the largest eigenvalue is the first principle component.

5. Derive new dataset by using most important n principle component

$$
\begin{bmatrix}
P_{C1} & P_{C2} & \cdots & P_{Cm} \\
P_{11} & P_{21} & & P_{m1} \\
P_{12} & P_{22} & & P_{m2} \\
\vdots & \vdots & & \vdots \\
P_{1n} & P_{2n} & & P_{mn}
\end{bmatrix}
\qquad P_{ij} = e_i^T
\begin{bmatrix}
x_{1j} - \bar{x}_1 \\
x_{2j} - \bar{x}_2 \\
\vdots \\
x_{kj} - \bar{x}_k
\end{bmatrix}
$$

Example :-

Given the following 2D dataset use PCA to determine principle components of the data and show 2D transformed dataset

x_1	4	8	13	7
x_2	11	4	5	14

Soln:-

No. of features $(K) = 2$

No. of samples $(n) = 4$

Now,

Compute mean as

$\bar{x}_1 = 8$

$\bar{x}_2 = 8.5$

Now;

Compute covariance

$$S = \begin{bmatrix} Cov(x_1,x_1) & Cov(x_1,x_2) \\ Cov(x_2,x_1) & Cov(x_2,x_2) \end{bmatrix}$$

$$Cov(x_i,x_j) = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)$$

$$Cov(x_1,x_1) = \frac{1}{4-1} \sum_{k=1}^{4} (x_{1k} - \bar{x}_1)(x_{1k} - \bar{x}_1)$$

$$= \frac{1}{3}\{(x_{11} - \bar{x}_1)(x_{11} - \bar{x}_1) + (x_{12} - \bar{x}_1)(x_{12} - \bar{x}_1) + (x_{13} - \bar{x}_1)(x_{13} - \bar{x}_1) + (x_{14} - \bar{x}_1)(x_{14} - \bar{x}_1)\}$$

$$\frac{1}{3}\{(4-8)^2 + (8-8)^2 + (13-8)^2 + (7-8)^2\}$$

$$= \frac{1}{3}(16 + 0 + 25 + 1)$$

$$= \frac{1}{3} \times 42 = 14$$

$$\text{Cov}(x_1, x_2) = \frac{1}{4-1} \sum_{k=1}^{4} (x_{1k} - \bar{x}_1)(x_{2k} - \bar{x}_2)$$

$$= \frac{1}{3} \{(x_{11} - \bar{x}_1)(x_{21} - \bar{x}_2) + (x_{12} - \bar{x}_1)(x_{22} - \bar{x}_2) + (x_{13} - \bar{x}_1)(x_{23} - \bar{x}_2) + (x_{14} - \bar{x}_1)(x_{24} - \bar{x}_2)\}$$

$$= \frac{1}{3} \{(4-8)(11-8.5) + (8-8)(4-8.5) + (13-8)(5-8.5) + (7-8)(14-8.5)$$

$$= \frac{1}{3} \{-4 \times 2.5 + 0 + 5 \times -3.5 + (-1)(5.5)\}$$

$$= \frac{1}{3} (-10 - 17.5 - 5.5)$$

$$= \frac{1}{3} \times -33 = -11$$

$$\text{Cov}(x_2, x_1) = \frac{1}{4-1} \sum_{k=1}^{4} (x_{2k} - \bar{x}_2)(x_{1k} - \bar{x}_1)$$

$$= \frac{1}{4-1} \{(x_{21} - \bar{x}_2)(x_{11} - \bar{x}_1) + (x_{22} - \bar{x}_2)(x_{12} - \bar{x}_1) + (x_{23} - \bar{x}_2)(x_{13} - \bar{x}_1) + (x_{24} - \bar{x}_2)(x_{14} - \bar{x}_1)$$

$$= -11$$

$$\text{Cov}(x_2, x_2) = \frac{1}{4-1} \sum_{k=1}^{4} (x_{2k} - \bar{x}_2)(x_{2k} - \bar{x}_2)$$

$$= \frac{1}{4-1} \{ (x_{21} - \bar{x}_2)^2 + (x_{22} - \bar{x}_2)^2 + (x_{23} - \bar{x}_2)^2 + (x_{24} - \bar{x}_2)^2 \}$$

$$= \frac{1}{3} \{ (11 - 8.5)^2 + (4 - 8.5)^2 + (5 - 8.5)^2 + (14 - 8.5)^2 \}$$

$$= \frac{1}{3} \{ (2.5)^2 + (-4.5)^2 + (-3.5)^2 + (5.5)^2 \}$$

$$= \frac{1}{3} \times 69 = 23$$

$$\therefore S = \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$$

Now,

For eigen values,

$$|S - \lambda I| = 0$$

$$\left| \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0$$

$$\begin{vmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{vmatrix} = 0$$

$$(14-\lambda)(23-\lambda) - 121 = 0$$

$$322 - 37\lambda + \lambda^2 - 121 = 0$$

$$\lambda^2 - 37\lambda + 201 = 0$$

Solving using quadratic eqn

$$\lambda = \frac{37 \pm \sqrt{565}}{2}, \quad \lambda = \frac{37 - \sqrt{565}}{2}$$

$$\lambda_1 = 30.38 \qquad \lambda_2 = 6.61$$

Now,
Computing eigen vector

$$(S - \lambda_1 I) U = 0$$

$$\begin{bmatrix} 14-\lambda_1 & -11 \\ -11 & 23-\lambda_1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = 0$$

$$\begin{bmatrix} 14-30.38 & -11 \\ -11 & 23-30.38 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = 0$$

$$\begin{bmatrix} -16.38 & -11 \\ -11 & -7.38 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = 0$$

$$-16.38 u_1 - 11 u_2 = 0$$

$$-11 u_1 - 7.38 u_2 = 0$$

$$16.38 u_1 + 11 u_2 = 0$$

$$11 u_1 + 7.38 u_2 = 0$$

Solving both eqn

$$\frac{u_1}{11} = \frac{-u_2}{16.38}$$

$\therefore u_1 = 11$

$u_2 = -16.38$

$\therefore$ Thus, Eigen vector of $\lambda_1 (u_1) = \begin{bmatrix} 11 \\ -16.38 \end{bmatrix}$

Now for $\lambda_2 = 6.61$

Computing eigen vectors.

$$[S - \lambda_2 I] u_2 = 0$$

$$\begin{bmatrix} 14 - 6.61 & -11 \\ -11 & 23 - 6.61 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = 0$$

$$\begin{bmatrix} 7.39 & -11 \\ -11 & 16.39 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = 0$$

$$7.39 u_1 - 11 u_2 = 0$$
$$-11 u_1 + 16.39 u_2 = 0$$

$$\frac{u_1}{11} = \frac{u_2}{7.39}$$

$$\therefore \lambda_2 (u_2) = \begin{bmatrix} 11 \\ 7.39 \end{bmatrix}$$

Now,

Normalize the eigen vectors

$$U_1(e_1) = \begin{bmatrix} \frac{11}{\sqrt{11^2+(16.38)^2}} \\ \frac{-16.38}{\sqrt{11^2+(-16.38)^2}} \end{bmatrix}$$

$$U_1(e_1) = \begin{bmatrix} 0.557 \\ -0.83 \end{bmatrix}$$

$$U_2(e_2) = \begin{bmatrix} \frac{11}{\sqrt{11^2+7.39^2}} \\ \frac{7.39}{\sqrt{11^2+7.39^2}} \end{bmatrix}$$

$$= \begin{bmatrix} 0.83 \\ 0.557 \end{bmatrix}$$

Now,

Compute new dataset using principle component

$$P_{11} = e_1^T \begin{bmatrix} x_{11} - \bar{x}_1 \\ x_{21} - \bar{x}_2 \end{bmatrix}$$

$$= \begin{bmatrix} 0.557 & -0.83 \end{bmatrix} \begin{bmatrix} 4 - 8 \\ 11 - 8.5 \end{bmatrix}$$

$$= \begin{bmatrix} 0.557 & -0.83 \end{bmatrix} \begin{bmatrix} -4 \\ 8.5 \end{bmatrix}$$

$$= 0.557 \times (-4) - 0.83 \times 3.5$$

$$= -5.133$$

$$P_{12} = e_1^T \begin{bmatrix} x_{12} - \bar{x}_1 \\ x_{22} - \bar{x}_2 \end{bmatrix}$$

$$= \begin{bmatrix} 0.557 & -0.83 \end{bmatrix} \begin{bmatrix} 8 - 8 \\ 4 - 8.5 \end{bmatrix}$$

$$= \begin{bmatrix} 0.557 & -0.83 \end{bmatrix} \begin{bmatrix} 0 \\ -4.5 \end{bmatrix}$$

$$= 0 + 0.83 \times 4.5$$

$$= 3.735$$

$$P_{21} = e_2^T \begin{bmatrix} x_{11} - \bar{x}_1 \\ x_{21} - \bar{x}_2 \end{bmatrix}$$

$$= \begin{bmatrix} 0.83 & 0.577 \end{bmatrix} \begin{bmatrix} -4 \\ 3.5 \end{bmatrix}$$

$$= 0.83 \times (-4) + 0.577 \times 3.5$$

$$= -1.3$$

$$P_{22} = e_2^T \begin{bmatrix} x_{12} - \bar{x}_1 \\ x_{22} - \bar{x}_2 \end{bmatrix}$$

$$= \begin{bmatrix} 0.83 & 0.577 \end{bmatrix} \begin{bmatrix} 0 \\ -4.5 \end{bmatrix}$$

$$= 0 - 4.5 \times 0.577$$

$$= -2.59$$

∴ ∎

$$P_{13} = e_1^T \begin{bmatrix} x_{13} - \bar{x}_1 \\ x_{23} - \bar{x}_2 \end{bmatrix}$$

$$= [0.557 \quad -0.83] \begin{bmatrix} 13 - 8 \\ 5 - 8.5 \end{bmatrix}$$

$$= 0.557 \times 5 - 0.83 \times (-3.5)$$

$$= 5.69$$

$$P_{14} = e_1^T \begin{bmatrix} x_{14} - \bar{x}_1 \\ x_{24} - \bar{x}_2 \end{bmatrix} =$$

$$[0.557 \quad -0.83] \begin{bmatrix} 7 - 8 \\ 14 - 8.5 \end{bmatrix}$$

$$= 0.557 \times (-1) - 0.83 \times (5.5)$$

$$= -5.122$$

$$P_{23} = e_2^T \begin{bmatrix} x_{13} - \bar{x}_1 \\ x_{23} - \bar{x}_2 \end{bmatrix}$$

$$= 0.83 \times 5 - 0.557 \times (-3.5)$$

$$= 2.2$$

$$P_{24} = e_2^T \begin{bmatrix} x_{14} - \bar{x}_1 \\ x_{24} - \bar{x}_2 \end{bmatrix}$$

$$= \begin{bmatrix} 0.83 & 0.557 \end{bmatrix} \begin{bmatrix} -1 \\ 5.5 \end{bmatrix}$$

$$= 0.83 \times -1 + 0.557 \times (5.5)$$

$$= 2.2335$$

$$\therefore \begin{pmatrix} P_{11} & P_{21} \\ P_{12} & P_{22} \\ P_{13} & P_{23} \\ P_{14} & P_{24} \end{pmatrix} = \begin{pmatrix} -5.133 & -1.3 \\ 3.735 & -2.59 \\ 5.69 & 2.2 \\ -5.122 & 2.2335 \end{pmatrix}$$

Low Rank Approximation

The maximum number of its linearly independent columns (or rows) of a matrix is called the rank of a matrix. The rank of a matrix cannot exceed the number of its rows or columns.

Assume we have a matrix A, with rank d and we wish to produce matrix B such that (a) the rank of B is s (which is less than d) and (b) such that $\|A-B\|^2$ is minimized. The resulting matrix B is called a low rank approximation to A.

- Low rank approximation is used in compression, denoising, and matrix completion.

Singular Value Decomposition (SVD) is one of the method used for finding low rank approximation.

Eigen decomposition is only possible with square matrices. For rectangular matrices having dimension $m \times n$ SVD can be used.

Let A be matrix of dimension $m \times n$. Now the matrix A^TA is matrix of order $n \times n$.
There exist n eigenvalues of matrix A^TA:

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_n > 0$$

Let $\sigma_i = \sqrt{\lambda_i}$ here $\sigma_i, i = 1, 2, \cdots n$ are called Singular values of A.

Let A be $m \times n$ matrix, then SVD of A is defined as:

$$A = USV^T$$ where U is an $m \times m$ orthogonal matrix, V is $n \times n$ orthogonal matrix and S is diagonal matrix whose diagonal elements are singular values of A.

We know that, eigenvalue equation of matrix A is

$$Ax = \lambda x$$

$$\Rightarrow A^T A x = \lambda x$$

$$\Rightarrow A^T A V_i = \lambda_i V_i$$

Matrix V & S can be found by solving above eqn. The matrix V is combination of all V_i

Similarly we know that eigen value equation of matrix A is

$$Ax = \lambda x$$

$AA^T x = \lambda x$

$AA^T u_i = \lambda_i u_i$

Matrix U can be found by solving above equation
The matrix U is combination of all u_i.

Example :-

Find SVD of following matrix

$$A = \begin{bmatrix} 0 & 1 & 1 \\ \sqrt{2} & 2 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

Solution :-

$$B = AA^T$$

$$= \begin{bmatrix} 0 & 1 & 1 \\ \sqrt{2} & 2 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & \sqrt{2} & 0 \\ 1 & 2 & 1 \\ 1 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 2 & 2 & 2 \\ 2 & 6 & 2 \\ 2 & 2 & 2 \end{bmatrix}$$

Now,

$$|B - \lambda I| = 0$$

$$\left| \begin{bmatrix} 2 & 2 & 2 \\ 2 & 6 & 2 \\ 2 & 2 & 2 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right| = 0$$

$$\begin{vmatrix} 2-\lambda & 2 & 2 \\ 2 & 6-\lambda & 2 \\ 2 & 2 & 2-\lambda \end{vmatrix} = 0$$

$$(2-\lambda)\left((6-\lambda)(2-\lambda) - 4\right) - 2\left(2(2-\lambda) - 4\right) + 2\left(4 - 2(6-\lambda)\right) = 0$$

$$(2-\lambda)\left(12 - 8\lambda + \lambda^2 - 4\right) - 2(4 - 2\lambda - 4) + 2\left(4 - 12 + 2\lambda\right) = 0$$

$$24 - 16\lambda + 2\lambda^2 - 8 - 12\lambda + 8\lambda^2 - \lambda^3 + 4\lambda + 4\lambda + 8 - 24 + 4\lambda = 0$$

$$-\lambda^3 + 10\lambda^2 - 16\lambda = 0$$

$$\lambda_3 = 0 \quad \lambda_2 = 2 \quad \lambda_1 = 8$$

Thus;

$\sigma_3 = \sqrt{\lambda_1}$ $\qquad \sigma_2 = \sqrt{2}$

$= \sqrt{0} = 0$

$\sigma_1 = \sqrt{\lambda_3} = \sqrt{8} = 2\sqrt{2}$

for eigen vector of each eigen value

$\lambda_3 = 0$

$$[B - \lambda_1 I]\, u = 0$$

$$\begin{bmatrix} 2 & 2 & 2 \\ 2 & 6 & 2 \\ 2 & 2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0$$

$$2x_1 + 2x_2 + 2x_3 = 0$$

$$2x_1 + 6x_2 + 2x_3 = 0$$

$$2x_1 + 2x_2 + 2x_3 = 0$$

Solving we get

$$\frac{x_1}{\begin{vmatrix} 2 & 2 \\ 6 & 2 \end{vmatrix}} = \frac{x_2}{\begin{vmatrix} 2 & 2 \\ 2 & 2 \end{vmatrix}} = \frac{x_3}{\begin{vmatrix} 2 & 2 \\ 2 & 6 \end{vmatrix}}$$

$$\frac{x_1}{-8} = \frac{-x_2}{0} = \frac{x_3}{8}$$

$$\frac{x_1}{8} = \frac{x_2}{0} = \frac{-x_3}{8} \Rightarrow x_1 = 1$$

$x_2 = 0$

$x_3 = -1$

$$\therefore u_3 = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

for $\lambda_2 = 2$

$$\begin{bmatrix} 0 & 2 & 2 \\ 2 & 4 & 2 \\ 2 & 2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0$$

$$0x_1 + 2x_2 + 2x_3 = 0$$

$$2x_1 + 4x_2 + 2x_3 = 0$$

$$2x_1 + 2x_2 = 0$$

Solving we get

$$\frac{x_1}{\begin{vmatrix} 2 & 2 \\ 4 & 2 \end{vmatrix}} = \frac{-x_2}{\begin{vmatrix} 0 & 2 \\ 2 & 2 \end{vmatrix}} = \frac{x_3}{\begin{vmatrix} 0 & 2 \\ 2 & 4 \end{vmatrix}}$$

$$\frac{x_1}{-4} = \frac{-x_2}{-4} = \frac{x_3}{-4}$$

$$\frac{x_1}{-1} = \frac{x_2}{1} = \frac{x_3}{-1}$$

$$u_2 = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \\ -1 \end{bmatrix}$$

For $\lambda_3 = 8$

$$\begin{bmatrix} -6 & 2 & 2 \\ 2 & -2 & 2 \\ 2 & 2 & -6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0$$

$$-6x_1 + 2x_2 + 2x_3 = 0$$
$$2x_1 - 2x_2 + 2x_3 = 0$$
$$2x_1 + 2x_2 - 6x_3 = 0$$

Solving we get

$$\frac{x_1}{\begin{vmatrix} 2 & 2 \\ -2 & 2 \end{vmatrix}} = \frac{-x_2}{\begin{vmatrix} -6 & 2 \\ 2 & 2 \end{vmatrix}} = \frac{x_3}{\begin{vmatrix} -6 & 2 \\ 2 & -2 \end{vmatrix}}$$

$$\frac{x_1}{8} = \frac{-x_2}{-16} = \frac{x_3}{+8}$$

$$\frac{x_1}{1} = \frac{x_2}{2} = \frac{x_3}{1}$$

$$\therefore u_3 = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$$

Now, ~~Normalized~~ for normalized Eigen vector.

$$u_3 = \begin{bmatrix} 1/\sqrt{2} \\ 0 \\ -1/\sqrt{2} \end{bmatrix} \quad u_2 = \begin{bmatrix} -1/\sqrt{3} \\ 1/\sqrt{3} \\ -1/\sqrt{3} \end{bmatrix} \quad u_1 = \begin{bmatrix} 1/\sqrt{6} \\ 2/\sqrt{6} \\ 1/\sqrt{6} \end{bmatrix}$$

Thus;

$$U = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{3} & 1/\sqrt{6} \\ 0 & 1/\sqrt{3} & 2/\sqrt{6} \\ -1/\sqrt{2} & -1/\sqrt{3} & 1/\sqrt{6} \end{bmatrix} \quad U = \begin{bmatrix} 1/\sqrt{6} & -1/\sqrt{3} & 1/\sqrt{2} \\ 2/\sqrt{6} & 1/\sqrt{3} & 0 \\ 1/\sqrt{6} & -1/\sqrt{3} & -1/\sqrt{2} \end{bmatrix}$$

$$S = \begin{bmatrix} 2 & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Now;

for calculation of V

$$B = A^T A$$

$$= \begin{bmatrix} 0 & \sqrt{2} & 0 \\ 1 & 2 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 1 \\ \sqrt{2} & 2 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0+2+0 & 0+2\sqrt{2}+0 & 0+0+0 \\ 0+2\sqrt{2}+0 & 1+4+1 & 1+0+1 \\ 0+0+0 & 1+0+1 & 1+0+1 \end{bmatrix}$$

$$= \begin{bmatrix} 2 & 2\sqrt{2} & 0 \\ 2\sqrt{2} & 6 & 2 \\ 0 & 2 & 2 \end{bmatrix}$$

Now;

$$|B - \lambda I| = 0$$

$$\left| \begin{bmatrix} 2 & 2\sqrt{2} & 0 \\ 2\sqrt{2} & 6 & 2 \\ 0 & 2 & 2 \end{bmatrix} - \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix} \right| = 0$$

$$\begin{vmatrix} 2-\lambda & 2\sqrt{2} & 0 \\ 2\sqrt{2} & 6-\lambda & 2 \\ 0 & 2 & 2-\lambda \end{vmatrix} = 0$$

$$(2-\lambda)\left((6-\lambda)(2-\lambda)-4\right) - 2\sqrt{2}\left(2\sqrt{2}(2-\lambda)-0\right) = 0$$

$$(2-\lambda)\left(12-8\lambda+\lambda^2-4\right) - 2\sqrt{2}\left(4\sqrt{2}-2\sqrt{2}\lambda\right) = 0$$

$$(2-\lambda)(8-8\lambda+\lambda^2) - (8\times 2 - 4\times 2\lambda) = 0$$

$$16 - 16\lambda + 2\lambda^2 - 8\lambda + 8\lambda^2 - \lambda^3 - 16 + 8\lambda = 0$$

$$-\lambda^3 + 10\lambda^2 - 16\lambda = 0$$

$$\lambda^3 - 10\lambda^2 + 16\lambda = 0$$

$$\lambda_3 = 0 \quad \lambda_2 = 2 \quad \lambda_1 = 8$$

for $\lambda_1 = 8$

$$\begin{bmatrix} -6 & 2\sqrt{2} & 0 \\ 2\sqrt{2} & -2 & 2 \\ 0 & 2 & -6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0$$

$$-6x_1 + 2\sqrt{2}x_2 + 0x_3$$

$$2\sqrt{2}x_1 - 2x_2 + 2x_3 = 0$$

$$0x_1 + 2x_2 - 6x_3 = 0$$

Solving;

$$\frac{x_1}{\begin{vmatrix} 2\sqrt{2} & 0 \\ -2 & 2 \end{vmatrix}} = \frac{-x_2}{\begin{vmatrix} -6 & 0 \\ 2\sqrt{2} & 2 \end{vmatrix}} = \frac{x_3}{\begin{vmatrix} -6 & 2\sqrt{2} \\ 2\sqrt{2} & -2 \end{vmatrix}}$$

$$\frac{x_1}{4\sqrt{2}} = \frac{-x_2}{-12} = \frac{x_3}{4}$$

$$\frac{x_1}{\sqrt{2}} = \frac{x_2}{3} = \frac{x_3}{1}$$

$$\therefore u_1 = \begin{bmatrix} \sqrt{2} \\ 3 \\ 1 \end{bmatrix}$$

for; $\lambda_2 = 2$

$$\begin{bmatrix} 0 & 2\sqrt{2} & 0 \\ 2\sqrt{2} & 4 & 2 \\ 0 & 2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0$$

$$0x_1 + 2\sqrt{2}x_2 + 0x_3 = 0$$

$$2\sqrt{2}x_1 + 4x_2 + 2x_3 = 0$$

$$0x_1 + 2x_2 + 0x_3 = 0$$

Solving for x_1, x_2 & x_3

$$\frac{x_1}{\begin{vmatrix} 2\sqrt{2} & 0 \\ 4 & 2 \end{vmatrix}} = \frac{-x_2}{\begin{vmatrix} 0 & 0 \\ 2\sqrt{2} & 2 \end{vmatrix}} = \frac{x_3}{\begin{vmatrix} 0 & 2\sqrt{2} \\ 2\sqrt{2} & 4 \end{vmatrix}}$$

$$\frac{x_1}{4\sqrt{2}} = \frac{-x_2}{0} = \frac{x_3}{-8}$$

$$\frac{x_1}{\sqrt{2}} = \frac{x_2}{0} = \frac{x_3}{-2}$$

$$\therefore u_2 = \begin{bmatrix} \sqrt{2} \\ 0 \\ -2 \end{bmatrix}$$

For $\lambda_3 = 0$

$$\begin{bmatrix} 2 & 2\sqrt{2} & 0 \\ 2\sqrt{2} & 6 & 2 \\ 0 & 2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0$$

$$2x_1 + 2\sqrt{2}x_2 + 0x_3 = 0$$

$$2\sqrt{2}x_1 + 6x_2 + 2x_3 = 0$$

$$0x_1 + 2x_2 + 2x_3 = 0$$

$$\frac{x_1}{\begin{vmatrix} 2\sqrt{2} & 0 \\ 6 & 2 \end{vmatrix}} = \frac{-x_2}{\begin{vmatrix} 2 & 0 \\ 2\sqrt{2} & 2 \end{vmatrix}} = \frac{x_3}{\begin{vmatrix} 2 & 2\sqrt{2} \\ 2\sqrt{2} & 6 \end{vmatrix}}$$

$$\frac{x_1}{4\sqrt{2}} = \frac{-x_2}{4} = \frac{x_3}{4}$$

$$\frac{x_1}{\sqrt{2}} = \frac{-x_2}{1} = \frac{x_3}{1}$$

$$\therefore u_3 = \begin{bmatrix} \sqrt{2} \\ -1 \\ 1 \end{bmatrix}$$

Now normalizing we get

$$v_2 \begin{bmatrix} \frac{\sqrt{2}}{\sqrt{2+9+1}} & \frac{\sqrt{2}}{\sqrt{2+4}} & \frac{\sqrt{2}}{\sqrt{2+1+1}} \\ \frac{3}{\sqrt{2+9+1}} & \frac{0}{\sqrt{2+4}} & \frac{-1}{\sqrt{2+1+1}} \\ \frac{1}{\sqrt{2+9+1}} & \frac{-2}{\sqrt{2+4}} & \frac{1}{\sqrt{2+1+1}} \end{bmatrix}$$

$$V = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} \\ 3/\sqrt{12} & 0 & -1/2 \\ 1/\sqrt{12} & -2/\sqrt{6} & 1/2 \end{bmatrix}$$

$$S = \begin{bmatrix} 2\sqrt{2} & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & 0 \end{bmatrix}$$