Assignment 1

1) Explain how supervised learning is different from unsupervised learning.

Ans) Supervised Learning:
Supervised Learning is the type of machine learning in which machines are trained using well labelled training data, & on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct data output.

Unsupervised Learning:
Unsupervised learning is a type of machine learning in which models are trained using unlabelled dataset & are allowed to act on that without any supervision. ~~Insted~~ Instead, models itself find the hidden patterns & insights from the given data.

Difference between them:
- The main distinction between the two approaches is the used of labelled output & input data, while on ~~us~~ unsupervised learning algorithm does not.
- In supervised learning, the algorithm "learns" from the training dataset by iteratively making predictions on the data & adjusting for the correct answer, While supervised learning models tend to be more accurate than unsupervised learning models, they require upfront human intervention to label the data appropriately.
- Unsupervised learning models, in contrast, work on their own to discover the inherent structure of unlabelled data. Note that they still require some human intervention for validating output variables for example an unsupervised learning model can identify that online shoppers often purchase groups of products at the same time. However that it makes sense for a recommendation engine to group baby clothes ~~whithin~~ within an order of diapers, apple sauce & sippy cups.

(P.T.O)

- other key difference are:
- Goals: In supervised learning, the goal is to predict outcomes for the new data. With an, unsupervised learning algorithm, the goal is to get insights from large volumes of new data.
- Applications: Supervised learning models are ideal for spam detection, sentiment analysis, weather forecasting & pricing predictions, among ~~is great~~ other things. In contrast, unsupervised learning is great fit for anomaly detection, recommendation engines, customer personas & medical imaging.
- Complexity: Supervised learning is a simple method for machine learning typically using R or Python while unsupervised learning models are computationally complex because they need a large ~~do~~ training sets to produce intended outcomes.
- Drawbacks: Supervised learning models can be time-consuming to train, & the labels for input & output variables require expertise. Meanwhile, unsupervised learning methods can have wildly inaccurate results unless you have human intervention to validate the output variables.

2) Find SVD for the matrix $\begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix}$.

$A = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix}$

$A^T = \begin{bmatrix} 3 & 2 \\ 2 & 3 \\ 2 & -2 \end{bmatrix}$

$\therefore A \cdot A^T = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix} \begin{bmatrix} 3 & 2 \\ 2 & 3 \\ 2 & -2 \end{bmatrix}$

(P.T.O)

MAHAVIR EDUCATION TRUST'S

Shah & Anchor Kutchhi Engineering College

Date :

$\therefore A.A^T = \begin{bmatrix} 17 & 8 \\ 8 & 17 \end{bmatrix}$

Now,

using charecteristic equation, we get -

$|A.A^T - \lambda I| = 0$

$\therefore \begin{vmatrix} 17-\lambda & 8 \\ 8 & 17-\lambda \end{vmatrix} = 0$

$(17-\lambda)(17-\lambda) - (8 \times 8) = 0$

$\therefore 289 - 17\lambda - 17\lambda + \lambda^2 - 64 = 0$

$\therefore \lambda^2 - 34\lambda + 225 = 0$

$\therefore \lambda = 25, 9$

$\therefore$ Eigen values are 25, 9 -

Now,

getting eigen vector,

$[A.A^T - \lambda I][X] = 0$

for $\lambda = 25$,

$\begin{bmatrix} -8 & 8 \\ 8 & -8 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$

$\therefore \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$

$\therefore -x_1 + x_2 = 0$

$\therefore x_1 = x_2$

$\therefore$ Eigen vector for $\lambda = 25$ is $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

for $\lambda = 9$,

$\begin{bmatrix} 8 & 8 \\ 8 & 8 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$

$\therefore \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$

$\therefore x_1 + x_2 = 0 \quad \therefore x_1 = -x_2$

(P.T.O)

$\therefore$ Eigen vector for $\lambda = 9$ is $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$

Now,
finding L for $[1, 1]$ & $[1, -1]$, we get -

$L_1 = \sqrt{(1)^2 + (1)^2} = \sqrt{2}$

$L_2 = \sqrt{(1)^2 + (-1)^2} = \sqrt{2}$

Now, normalizing vectors to get U_1 & U_2

$\therefore U_1 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right), \quad U_2 = \left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right)$

$\therefore U = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$

Also, singular values are -

$\sigma_1 = \sqrt{25} = 5, \quad \sigma_2 = \sqrt{9} = 3$

$\therefore W = \begin{bmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix}$

Now,

$A^T . A = \begin{bmatrix} 3 & 2 \\ 2 & 3 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix}$

$\therefore A^T . A = \begin{bmatrix} 13 & 12 & 2 \\ 12 & 13 & -2 \\ 2 & -2 & 8 \end{bmatrix}$

Now,
using charecteristic equation, we get -

$(A^T A - \lambda I) = 0$

$\therefore \begin{vmatrix} 13-\lambda & 12 & 2 \\ 12 & 13-\lambda & -2 \\ 2 & -2 & 8-\lambda \end{vmatrix} = 0$

$\therefore \lambda^3 - 34\lambda^2 + (100 + 100 + 25)\lambda - 200 = 0$

$\cancel{\therefore \lambda^3 - 34\lambda^2 + 233\lambda - 200 = 0}$

(P.T.O)

~~$\therefore \lambda = 25, 8, 1$~~

$\therefore \lambda^3 - 34\lambda^2 + 225\lambda - 0 = 0$

$\therefore \lambda = 25, 9, 0$

Now,

getting eigen vector,

$$[A^T \cdot A - \lambda I][X] = 0$$

for $\lambda = 25$,

$$\begin{bmatrix} -12 & 12 & 2 \\ 12 & -12 & -2 \\ 2 & -2 & -17 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0$$

Using Cramers rule, we get -

$$\frac{x_1}{200} = \frac{x_2}{200} = \frac{x_3}{0}$$

$$\therefore \frac{x_1}{1} = \frac{x_2}{1} = \frac{x_3}{0}$$

$\therefore$ The eigen vector for $\lambda = 25$ is $\begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$.

for $\lambda = 9$,

$$\begin{bmatrix} 4 & 12 & 2 \\ 12 & 4 & -2 \\ 2 & -2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0$$

$\therefore$ By cramers rule, we get -

$$\frac{x_1}{-32} = \frac{x_2}{32} = \frac{x_3}{-128}$$

$\therefore$ Eigen vector for $\lambda = 9$ is $\begin{bmatrix} 1 \\ -1 \\ 4 \end{bmatrix}$.

(P.T.O)

for $\lambda = 0$

$$\begin{bmatrix} 13 & 12 & 2 \\ 12 & 13 & -2 \\ 2 & -2 & 8 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0$$

By cramers rule, we get -

$$\frac{x_1}{-50} = \frac{x_2}{50} = \frac{x_3}{25}$$

$\therefore$ Eigen vector for $\lambda = 0$ is $\begin{bmatrix} 2 \\ -2 \\ -1 \end{bmatrix}$

Now,

finding L for $[1, 1, 0]$, $[1, -1, 4]$, $[2, -2, -1]$, we get-

$L_1 = \sqrt{(1)^2 + (1)^2 + (0)^2} = \sqrt{2}$

$L_2 = \sqrt{(1)^2 + (-1)^2 + (4)^2} = 3\sqrt{2}$

$L_3 = \sqrt{(2)^2 + (-2)^2 + (-1)^2} = 3$

Now, normalized vector v, will be -

$$V = \begin{bmatrix} 1/\sqrt{2} & 1/3\sqrt{2} & 2/3 \\ 1/\sqrt{2} & -1/3\sqrt{2} & -2/3 \\ 0 & 4/3\sqrt{2} & -1/3 \end{bmatrix} = \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/6 & -2/3 \\ \sqrt{2}/2 & -\sqrt{2}/6 & 2/3 \\ 0 & 2\sqrt{2}/3 & 1/3 \end{bmatrix}$$

Here, we get- ~~SVD~~

~~A = U~~ $\therefore$ SVD is given as -

$A = UWV^T$

$$\therefore A = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} 5 & 0 & 0 \\ 0 & 3 & 0 \end{bmatrix} \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 & 0 \\ \sqrt{2}/6 & -\sqrt{2}/6 & 2\sqrt{2}/3 \\ -2/3 & 2/3 & 1/3 \end{bmatrix}$$

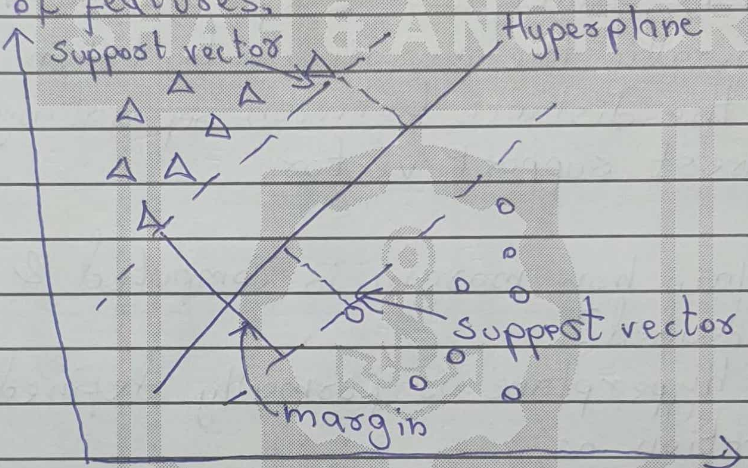
//

(P.T.O)

3) What is SVM? Explain the following terms: Separating hyper plane, margin & suppor vectors with suitgble example.

Ans) - Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification & regression.

- The objective of the SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.
- The dimension of the hyperplane depends upon the number of features.



- Supporting Hyperplane:

There can be multiple lines/decision boundries to segregr-ate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane depend on the features present in the dataset. which means if there are two features, then the hyperplane will be a straight line. And if there are 3 features, then hyperplane will be a 2dimension plane.

(P.T.O)

- Suppor vectors:
The data points or vectors that are the closest to the hyperplane and which affect the position of hyper plane are termed as support vector. Since these vectors support the hyperplane, hence they are also known as ~~sy~~ support vector.
- Margins:
The distance between the vectors & the hyperplane is called as margin. The best hyperplane will be whose ~~can be taken as 2*p, where~~ margin is the maximum. Generally the margin can be taken as 2*p, where p is the distance between separating hyperplane & nearest support vector.

4) Explain how margin is computed & optimal hyper-plane is decided?

Ans) - A Hyperplane is formally defined by the following notation as,

$$f(x) = w^t x + b,$$

In the above equation, w represents the weight vector & b represents the bias.
- By scaling the values of w & b we can represent ~~the bias~~ the optimal Hypeoplane in many ways. As a matter of fact, among all possible notations of the hyperplane the one selected is,

$$|w^t x + b| = 1$$

This notation is called as the canonical hyper plane.
- The distance between a point x & a hyperplane (w,b) is given as,

(P.T.O)

$$\text{Distance}_{sv} = \frac{|w^t x + b|}{||w||} = \frac{1}{||w||}$$

- Margin is twice the distance to nearest samples
$$M = \frac{2}{||w||}$$

- The optimal separating hyperplane H is the one that maximizes the margin
$$(w^*, b^*) = \underset{w,b}{\arg\max} \min_{\in\{1...N\}} \frac{|w_i x_i) + b|}{||w||^2}$$

$$\rho(M) = \frac{2}{||w||} \quad \forall (\bar{x_i}, \bar{y_i}) \in (D) ; y_i(\bar{w}^T \bar{x_i} + b) \geq 1$$

maximising Margin is same as minimizing the $\frac{1}{\rho} = \frac{||w||}{2}$ that is we need to find w & b such that:

$\frac{1}{2} \bar{w}^T \bar{w}$ is minimum. $\forall (\bar{x_i}, \bar{y_i}) \in D ; y_i(\bar{w}^T x_i + b) \geq 1$

Here, we are optimizing a quadratic equation with linear constraint. Now, this leads us to find the solution dual problems.

- Duality problem:

- In optimization, the duality principle states that optimization can either be viewed from a different perspective; the primal problem & the dual problem. The solution to dual problem provides a lower bound to the solution of the primal (minimization) problem.

- An optimization problem can be typically written as:
minimize$_x$ $f(x)$
subject to $g_i(x) = 0$, $i = 1, ..., p$.

(P.T.O)

$\therefore h_i(x) \leq 0 \quad i = 1 \ldots, m$

where, f is objective function g & h are constraint function. The above problem can be solved by a technique such as Lagrange multipliers.

- Lagrange Multipliers:

- Lagrange multiplier is a way of finding local minima & maxima for the functions with an equality constraint. Lagrange multipliers can be described as follows;

In Lagrange equation:

$\nabla f(x,y) = \nabla \lambda g(x,y)$ or

$\nabla f(x,y) - \nabla \lambda g(x,y) = 0$

Suppose, we define the function such that,

$\nabla L(x,y,\lambda) = \nabla f(x,y) - \nabla \lambda g(x,y)$

The above function is known as Lagrangian, now, we need to find $\nabla L(x,y,\lambda)$ is 0 i.e. point where gradient of functions f & g are parallel.