3/3/25

Lab-1

Date :
Page No.:

Write python code, consider filename as "housing.csv"

(i). To load .csv file into the data frame.

ii. To display information of all columns

iii). To display statistical information of all numerical.

iv). To display the count of unique labels for "Ocean Proximity" column.

v). To display which attributes (columns) in a dataset have missing values count greater than zero.

(i). df = pd.read_csv("housing.csv")

(ii) print(df)

(iii) print(df.describe()).

(iv). print(df["Ocean Proximity"].value_count())

(v). missing_value = df.isnull().sum()
missing_column = missing_values[missing_value > 0]
print(missing_column).

Date :
Page No.:

1. Which column in the dataset had missing values? How did you handle them?

Dropped the missing values with the help of ~~drop na~~ dropna since the number of rows ~~are~~ ~~less~~ with missing value is less compatibility.

2. What were the categorial ~~data~~ columns did you identify in the dataset? How did you encode them?

diabetes.csv : ["Gender", "Class"]
adult.csv : ["workclass", "education", "marx status", "occupation", "relationship", "gender", "native-country", "income"].

and we used LabelEncounter to encode it.

3. What is the difference between Min-Max Scaling and standardization? When would you use one over the other?

Both MinMax Scaling and standardization are feature scaling techniques.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad x' = \frac{x - \mu}{\sigma}$$

we use min max when we do not have significant outlier.
and we use standardization when we have outliers.

10/03/25

LAB - 2

Date :
Page No.:

Demonstrate the steps to build a machine-Learning model that predicts the median housing price using the California housing price dataset.

1. Perform the describe and info steps:

```
import pandas as pd
housing = pd.read_csv("content/drive/...")
print(housing.info())
print(housing.describe())
```

Output : column = 9
rows = 20640.

2. Plot histogram for each feature

The histogram for median_income & house_median_age can give us insights into the distribution of these features. For example, median_income might show a right-skewed distribution, indicating that most households have lower median income, while house_median_age might show how the ages of houses are distributed.

Date :
Page No.:

3. Demonstrate process of creating set & differences b/w random & stratified test.

Random	Stratified.
Splits data randomly into training & testing	Both are representative of the overall distribution of a feature.
May not preserve the distribution of key feature in training & testing set.	Preserves the distribution of key features across training & testing.

4. List geographical feature.:

This graph visualizes housing prices in relation to their geographic location, with the color representing median house value and size represents population

5. Which feature correlates to maximum:

median income usually shows the highest correlation with the median house value.

Date :
Page No.:

6. List the feature that could improve correlation.

By creating rooms_per_household & bedrooms_per_room.

The median house value is improved.

7. ~~total bedrooms~~ List the features that ~~leads~~ needs ~~data~~ to be cleaned & demonstrate cleaning process.

total bedrooms.

8. Categorial data to numerical data.

Yes, ocean_proximity is categorial feature and method used to convert is OneHotEncoder().

9. Important Scaling feature :-

- ensure all feature contribute equally to result.

Technique used StandardScaler & Minmax scaler.

17/03/25

LAB-3

Date :
Page No.:

Q Solving the following Linear Regression Problem using Matrix approach. Find linear Regression of the data of week and product sales.

x_i (week)	y_i (Sales in thousands)
1	2
2	4
3	5
4	9

$$\beta = ((X^T X)^{-1} X^T) y$$

$$X^T X = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{pmatrix} \times \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} = \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix}$$

$$(X^T X)^{-1} = \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix}^{-1} = \begin{bmatrix} 1.5 & -0.5 \\ -0.5 & 0.2 \end{bmatrix}$$

$$((X^T X)^{-1} X^T) = \begin{bmatrix} 1.5 & -0.5 \\ -0.5 & 0.2 \end{bmatrix} \times \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0.5 & 0 & -0.5 \\ -0.3 & -0.1 & 0.1 & 0.3 \end{bmatrix}$$

Finally,

$$((X^TX)^{-1}X^T)Y = \begin{bmatrix} 1 & 0.5 & 0 & -0.5 \\ -0.3 & -0.1 & 0.1 & 0.3 \end{bmatrix} \times \begin{bmatrix} 2 \\ 4 \\ 5 \\ 9 \end{bmatrix}$$

$$= \begin{bmatrix} -0.5 \\ -0.6 \quad -0.4 + 0.5 + \quad 2.7 \end{bmatrix}$$

$$= \begin{bmatrix} -0.5 \\ 2.2 \end{bmatrix}.$$

$$y = -0.5 + 2.2x.$$

→ Normal:

$\Sigma x = 10$ $\quad \Sigma x^2 = 30.$

$\Sigma y = 20.$ $\quad \Sigma xy = 61.$

$\bar{x} = 2.5 \quad \bar{y} = 5, \quad \bar{x}^2 = 7.5, \quad \bar{xy} = 15.25.$

$$\frac{\bar{xy} - \bar{x}\bar{y}}{\bar{x}^2 - (\bar{x})^2}$$

$$= \frac{15.25 - 5 \times 2.5}{7.5 - 6.25} = 2.2$$

$$x_0 = \bar{y} - \bar{x}\,\bar{x} = 5 - 2.2 \times 2.5 = -0.5$$

$$y = 2.2x - 0.5.$$

Date :
Page No.:

1. Did you perform any data processing steps?

Yes, handled missing value by filling them with the column mean. Also applied label encoding to categorial (like 'State') ca and scaled numerical features for 1000_companies.csv to normalize the data.

2. Did you visualize the regression line for cannada_per_capita_income.csv?

Yes.

Yes, the regression line was plotted. The plot shows a strong linear relationship between year & per-capita income, meaning that as the year increases, per capita income also rises.

3. Predicted salary for (12 years experience, 10 testscore, 10 interview score)? ₹

Sol

The predicted salary is printed in the script and depends on the trained model's coefficient.

4. Did you encode categorial variables for 1000_companies.csv?

Yes, the "State" column was encoded using LabelEncoder().

Date :
Page No.:

Did you scale the feature ? if yes, why?

Yes, because of R & D Spend, Administration, & Marketing Spend have different units, feature Scaling (using StandardScaler()) was applied to improve model performance.

24/03/25

Lab-4

Date :
Page No.:

Logistic Regression :-

(1). Student - passed or fail based on the study
Intercept $a_0 = -5$ coefficient : $a_1 = 0.8$.

(a). $P(z) = \frac{1}{1+e^{-z}}$ $\quad z = a_0 + a_1 x$

$z = -5 + 0.8x$.

$= \frac{1}{1+e^{5-0.8x}}$

(b). Probability that student who studies for 7 hours will pass :

$P(7) = \frac{1}{1+e^{5-5.6}} = \frac{1}{1+e^{-0.6}} = 0.6456$

(c). Determine predicted class (P/F) for student based on threshold 0.5

As $0.6456 > 0.5$ this statement will Pass.

Q.2. Consider $z = [2, -1, 0]$. for 3 classes. Apply ~~softmax~~ function to find probability value of 3 classes.

$\text{Softmax}(z_k) = \frac{e^{z_k}}{\sum_{j=1}^{k} e^{z_j}}$

$P(2) = \frac{e^2}{e^0 + e^{-1} + e^2} = 0.6652$

$$P(1) = \frac{e^1}{e^0+e^1+e^2} = 0.0900$$

$$P(0) = \frac{e^0}{e^0+e^1+e^2} = 0.2448$$

For HR_comma_sep.csv.

Q.1 Which variable had clear and direct impact on employee retention? Why?

Sol: satisfaction_level, time_spend_company, number_project, average_monthly_hours.

Reason: They showed strong correlation with retention based on EDA.

Q.2 Accuracy of logistic regression model? Good?

Accuracy (eg ~ 79%)

Yes it has good accuracy, it is above baseline, through improvement may be possible.

Date :
Page No.:

For Zoo dataset.

i) Did you perform data processing steps?

Dropped animal_name as it's not useful for classification.

ii) Where there any missing values?

No missing values detected.

iii) What does the confusion matrix tell you about the performance.

~~Shows class-wise performance.~~
It shows how many instance of each class were correctly or incorrectly classified. It helps assess accuracy and identify area of miscalculation.

iv) Which class type were most frequently misclassified? Why?

Class types with similare features (eg mammal & humans) were most frequently misclassified. This happened due to overlapping features and insufficient distinction between classes.

Date :
Page No.:

Decision Tree :

$$\text{Entropy}(T) = \sum_{p=i}^{n} P_i \log_2 P_i$$

$$\text{Entropy}(T, A) = \sum_{i=1} \frac{|A_i|}{|T|} * \text{Entropy}(A_i)$$

$$\text{Gain}(A) = E(T) - E(T, A)$$

$$S = [1+, 4-]$$

$$E(S) = -\frac{1}{5} \log_2 \frac{1}{5} - \frac{4}{5} \log_2 \frac{4}{5}$$

Lab-6

K- Nearest Neighbour

Q. Consider following dataset, k=3 and test data (x, 35, 100) as (Person, Age, Salary) solve using KNN classifier and predict target.

Person	Age	Salary k	Target	Dist
A	18	50	N	52.81
B	23	55	N	46.57
C	24	70	N	31.95
D	41	60	Y	40.44
E	43	70	Y	31.04
F	38	40	Y	60.07
X	35	100		

Nearest neighbours → E, C, D

for class Y:

Scored $\in \rho(Y, f(x))$

$d(Y, f(E)) + d(Y, f(C)) + d(Y, f(D))$

$d(Y, Y) + d(Y, N) + d(Y, Y)$

$0 + 1 + 0 = 1$

Scoring No: $d(N, Y) + d(N, N)$

$+ d(N, Y)$

$0 + 1 + 0 = 1$

Date :
Page No.:

21/04/25

Target - 1

Q For iris dataset, how to choose the k value? Demonstrate using accuracy rate and error rate.

Ans Cross validation is used to find the best k value depending on accuracy or number of misclassification. 5 was chosen.

Q For diabetes dataset.
What is the purpose of feature scaling? How to perform it?

Ans (i) Feature scaling is used in KNN to have the same scale because if we use a different scale, the features with larger scale get more weightage so the following lower scale features will get supressed and won't have an impact on the model. To perform scaling, first the frequency distributions of all features were checked. The one having normal distributions were scaled using standard scaler and the rest using minmax scaler.

(ii)

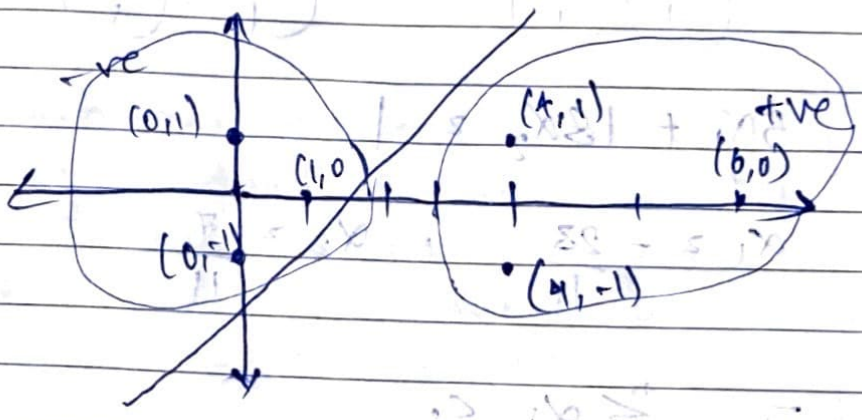
(iii)

8/4/25

Lab-6 - Support Vector Machine

Date :
Page No.:

Points (4,1), (4,-1) and (6,0) belong to positive class and points (1,0), (0,1) and (0,-1) belong to negative class. Draw an optimal hyperplane.



S_1 S_2

$$\begin{pmatrix}1\\0\end{pmatrix} \quad \begin{pmatrix}4\\1\end{pmatrix}$$

$$\alpha_1 \Rightarrow \begin{matrix}S_1\\ \begin{pmatrix}1\\0\\1\end{pmatrix}\end{matrix} \quad \begin{matrix}S_2\\ \begin{pmatrix}4\\1\\1\end{pmatrix}\end{matrix} \Leftarrow \alpha_2$$

$$\alpha_1 S_1 \cdot S_1 + \alpha_2 S_2 \cdot S_2 = -1$$

$$\alpha_1 \begin{pmatrix}1\\0\end{pmatrix} \cdot \begin{pmatrix}1\\0\end{pmatrix} + \alpha_2 \begin{pmatrix}4\\1\end{pmatrix} \cdot \begin{pmatrix}1\\0\end{pmatrix} = -1$$

$$\alpha_1 \begin{pmatrix}1\\0\\1\end{pmatrix}\begin{pmatrix}1\\0\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}4\\1\\1\end{pmatrix}\begin{pmatrix}1\\0\\1\end{pmatrix} =$$

$$2\alpha_1 + 5\alpha_2 = -1$$

$$\alpha_1 S_1 \cdot S_2 + \alpha_2 S_2 \cdot S_2 = 1 \quad (1)$$

$$\alpha_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}\begin{pmatrix} 4 \\ 1 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 4 \\ 1 \\ 1 \end{pmatrix}\begin{pmatrix} 4 \\ 1 \\ 1 \end{pmatrix} = 1$$

$$5\alpha_1 + 18\alpha_2 = 1$$

$$\alpha_1 = -\frac{23}{11}, \quad \alpha_2 = \frac{7}{11}$$

$$\bar{w} = \sum_{i=1}^{n} \alpha_i \cdot S_i \quad (2)$$

$$= \alpha_1 S_1 + \alpha_2 S_2$$

$$= -\frac{23}{11}\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \frac{7}{11}\begin{pmatrix} 4 \\ 1 \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} 5/11 \\ 7/11 \\ -16/11 \end{pmatrix} \quad \frac{5}{11}, \frac{7}{11} \rightarrow$$

orientation line

$$b = -\frac{16}{11}$$

$$\boxed{b + \frac{16}{11} = 0}$$

$$m = \tan\theta$$

$$\theta = \tan^{-1}(m)$$

orientation line $= \tan^{-1}\left(\frac{5}{7}\right)$

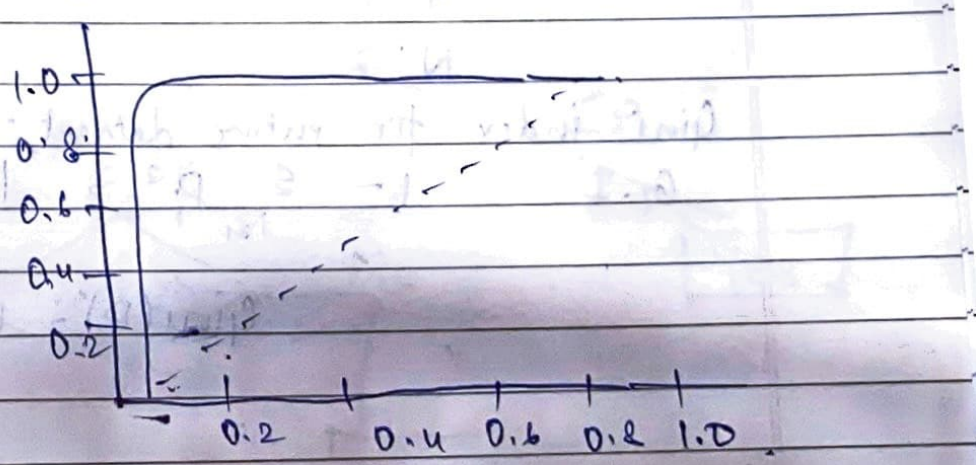
line cuts x-axis at 16/11.

compare accuracy of kernel. both linear RBF kernel gave 100 % accuracy on the test set.

Which one preferred & why?
Both prefered equally. The IRIS dataset is relatively simple & linearly separated that is why the linear kernel performs just as well as RBF.

Letter Recognition Dataset

Accuracy : 0.9055 AUC : 1.00



05/05/25

Lab. 7 ..

Date :
Page No. :

Random Forest
Ensemble

① For sample S1 draw decision tree considering CGPA as root node.

S.No	CGPA	Interactiveness	Communn Knowled	Job
1	≥9	Y	Good	Y
2	<9	N	Moderate	Y
3	≥9	N	Avg	N
4	≥9	N	Avg	N
5	≥9	Y	Good	Y

Ans as root node is CGPA we can seperate into unique groups as ≥9 & <9.

For CGPA ≥9

Y : 2

N : 2

Gini index for entire dataset : Y-3, N-2

$$G.I \quad 1 - \sum_{i=1} P_i^2 = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2$$

$$Gini(D) = 1 - 0.36 - 0.16$$

$$= 0.48.$$

For CGPA ≥9 Y : 2, N : 2

$$G(\geq 9) = 1 - (0.5)^2 - (0.5)^2 = 0.5$$

for CGPA <9 → 1 sample

Y = 1, N = 0.

Gini (<9) = 0.

Weighted Gini$_{CGPA}$ = $\frac{4}{5} \times 0.5 + \frac{1}{5} \times 0 = 0.4$

Information Gain (Gini reduction)

= 0.48 - 0.4 = 0.08

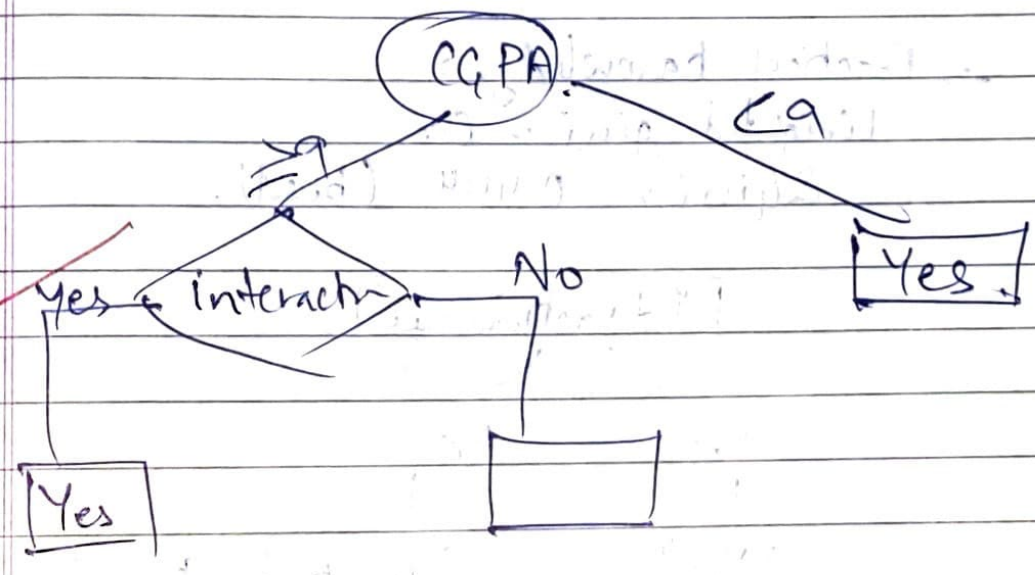
Split CGPA ≥9 further to get split on interaction

Y - Yes - 2

N - N - 2

Gini = $\frac{2}{4} \times 0 + \frac{2}{4} \times 0 = 0$.

(Gini reduction) = 0.5 - 0 = 0.5



1. Gini = $1 - \left(\frac{4}{5}\right)^2 - \left(\frac{1}{5}\right)^2 = 1 - 0.64 - 0.04$

$= 0.32$.

2. Gini = $\frac{2}{5} \times 0 + \frac{3}{5} \times 0.444 = 0.2664$.

Δ Gini = $0.32 - 0.2664 = 0.0536$

3. growing the "no" beach with other features.

→ CGPA:

weighted gini = $\frac{1}{3} \times 0 + \frac{2}{3} \times 0.50 = 0.333$

Δ Gini = 0.111.

→ Practical knowledge:

Weighted gini = 0.

ΔGini = 0.444 (best).

[Interactiveness]

Yes — "Yes"

No — [Practical knowledge]

Good — Yes

Average — No

Date :
Page No.:

RF	n-estimates	mean accuracy.
1	10	0.9667
2	50	0.9667
3	100	0.9667
4	150	0.9667
5	200	0.9667.

Accuracy = 1.0

Best number of trees = 20

confusion matrix = $\begin{bmatrix} 10 & 0 & 0 \\ 0 & 9 & 0 \\ 0 & 0 & 11 \end{bmatrix}$.

05/05/25.

Lab - 8

Date :
Page No.:

Boosting Ensemble method

CGPA	Intrac	Practi	Com	Tot
>=9	Y	G	G	Y
<9	N	G	M	Y
>=9	N	A	M	N
<9	N	A	G	N
>=9	Y	G	M	Y
>=9	Y	G	M	Y

Initial weight = 1/6.

$$\sum_{j=1}^{1} Hi(dj).wt(dj)$$

$$= 2 \times \frac{1}{6} \rightarrow 0.33$$

Pre	Act
Y	Y
N	Y
Y	N
N	N
Y	Y
Y	Y

$$\alpha \sim \frac{1}{2} \frac{\ln(1 - \varepsilon_{CGPA})}{\varepsilon_{CGPA}}$$

$$\alpha_{CGPA} = \frac{1}{2} \ln\left(\frac{1 - 0.33}{0.33}\right)$$

$$\alpha_{CGPA} = 0.347.$$

$$= \frac{1}{6} \times 4 \times e^{-0.347} + \frac{1}{6} \times 2 \times e^{0.347}$$

$$\rightarrow 0.9428.$$

new = 0.1249

0.2501.

Date :
Page No.:

Interact... Auth...

Y	Y	0.1244
N	Y	0.2501
Y	N	0.2501
N	N	0.1249
Y	Y	0.1249
Y	Y	0.1249

Adaboost

1	10	0.2037
2	50	0.1300
3	100	0.83050
4	150	0.8322
5	200	0.8326

Confusion

[7028 382
1243 1112]

12/oct/25.

Lab - 9

Date :
Page No. :

Q. Compute two clusters using kmeans algorithm for clustering whose cluster centers are (1,1) & (5,7) execute for 2 iteration.

→ Iteration 1:

Calc Euclidean dist to centroids:

Record	(1,1) Close to C1	(5,7) Close to C2	Assignment
(1,1)	0	7.21	C1
(1.5,2)	1.12	6.12	C1
(3,4)	3.61	3.61	C1
(5,7)	7.21	0.0	C2
(3.5,5)	4.12	2.5	C2
(4.5,5)	5.31	2.06	C2
(3.5,4.5)	4.30	2.92	C2

New centroids:

$$C_1 = \frac{1+1.5+3}{5}, \frac{4+2+1}{3} \rightarrow 1.83, 2.33$$

$$C_2 = \frac{5+3.5+4.5+3.5}{4}, \frac{7+5+5+4.5}{4}$$

$$= 4.12, 5.37$$

Iteration 2 Record	(1.83, 2.33) Close to C_1	(4.12, 5.37) Close to C_2	Assignment
(1,1)	1.57	5.37	C_1
(1.5,2)	0.47	4.27	C_1
(3,4)	2.04	1.77	C_2
(5,7)	5.61	1.85	C_2
(3.5,5)	3.15	0.72	C_2
(4.5,5)	3.78	0.53	C_2
(3.5,4.5)	2.74	1.07	C_2

∴ New cluster are

$C_1 = \{R_1, R_2\}$, $C_2 = \{R_3, R_4, R_5, R_6\}$

New centroids

$C_1 = \frac{2.5}{2}, \frac{3}{2}$ | $C_2 = \frac{19.5}{5}, \frac{25.5}{5}$

For Iris dataset

The elbow plot (inertia vs k) shows a sharp elbow at k=3, indicating that the cluster is the normal choice for petal length and width of the Iris.

12/08/25.

Lab 11

PCA

Q. Reduce elimination from 2 to 1.

X_1	4	8	13	7
X_2	11	4	5	14

Standadize the dataset

$$\bar{X_1} = \frac{4+8+13+7}{4} = 8$$

$$\bar{X_2} = \frac{(11+4+5+14)}{4} = 8.5$$

$$X_C^T = \begin{bmatrix} 4-8 & 8-8 & 13-8 & 7-8 \\ 11-8.5 & 4-8 & 5-8 & 14-8 \end{bmatrix}$$

$$X_C^T = \begin{bmatrix} -4 & 0 & 5 & -1 \\ 2.5 & -4.5 & -3.5 & 6.5 \end{bmatrix}$$

Covariance matrix

$$C = \left(\frac{1}{n-1}\right) \cdot X \cdot X^T$$

$$= \frac{1}{3}\begin{bmatrix} -4 & 0 & 5 & -1 \\ 2.5 & -4.5 & -3.5 & 5.5 \end{bmatrix}$$

$$= \frac{1}{3}\begin{bmatrix} -4 & 0 & 6 & -1 \\ 2.5 & -4.5 & -3.5 & 5.5 \end{bmatrix}\begin{bmatrix} -4 & 2.5 \\ 0 & -4.5 \\ 5 & -3.5 \\ -1 & 5.5 \end{bmatrix}$$

$$C = \frac{1}{3}\begin{bmatrix} 42 & -33 \\ -33 & 69 \end{bmatrix}$$

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

Let $(C - \lambda I) = 0$,

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

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

$$\lambda = 30.3849,$$

Consider larger λ

$$\begin{bmatrix} 14 - 30.3849 & -11 \\ -11 & -23 - 30.3849 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} -16.3849x & -11y \\ -11x & -7.3849y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$-16.3849x - 11y = 0$$
$$x = -0.6713y$$

taking $y = 1$

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -0.6713 \\ 1 \end{bmatrix}$$

Normalize

$$x = \frac{x}{\sqrt{x^2+y^2}} \quad y = \frac{y}{\sqrt{x^2+y^2}}$$

$$C_1 = \begin{bmatrix} -0.5574 \\ 0.8303 \end{bmatrix}$$

Calculate principle component.

$$Z = X_c C_1 = \begin{bmatrix} -4 & 2.5 \\ 0 & -4.5 \\ 5 & 3.5 \\ -1 & 5.5 \end{bmatrix} \begin{bmatrix} -0.5574 \\ 0.8303 \end{bmatrix}$$

$$Z = \begin{bmatrix} +4.305 \\ -3.736 \\ -5.693 \\ +5.124 \end{bmatrix}$$

Date :
Page No.:

Accuracy before PCA

logistic regression : 0.9016
SVM : 0.8525.
Random forest : 0.8361

Accuracy After PCA
logistic : 0.8689
SVM : 0.8689.
Random forest : 0.88.