



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING (DATA SCIENCE)

COURSE CODE: DJ19DSC501

DATE:

COURSE NAME: Machine Learning - II

CLASS: AY 2023-24

Jhanvi Parekh

60009210033

D11

LAB EXPERIMENT NO. 8

AIM :

Compare the performance of PCA and Autoencoders on a given dataset

THEORY:

What is Dimensionality Reduction?

In machine learning classification problems, there are often too many factors on the basis of which the final classification is done. These factors are basically variables called features. The higher the number of features, the harder it gets to visualize the training set and then work on it. Sometimes, most of these features are correlated, and hence redundant. This is where dimensionality reduction algorithms come into play. Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables. It can be divided into feature selection and feature extraction.

Principle Component Analysis

Principle Component Analysis is an unsupervised technique where the original data is projected to the direction of high variance. These directions of high variance are orthogonal to each other resulting in very low or almost close to 0 correlation in the projected data. These features transformation is linear and the methodology to do it is:

Step 1: Calculate the Correlation matrix data consisting of n dimensions. The Correlation matrix will be of shape $n \times n$.

Step 2: Calculate the Eigenvectors and Eigenvalues of this matrix.

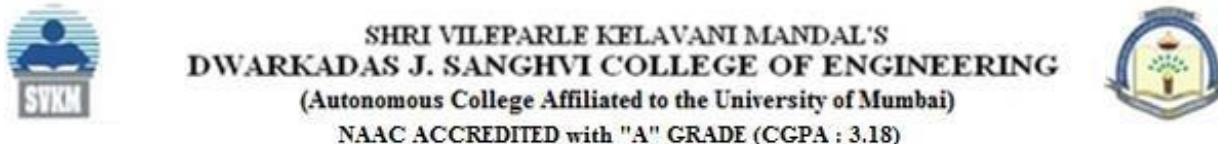
Step 3: Take the first k-eigenvectors with the highest eigenvalues.

Step 4: Project the original dataset into these k eigenvectors resulting in k dimensions where $k \leq n$.

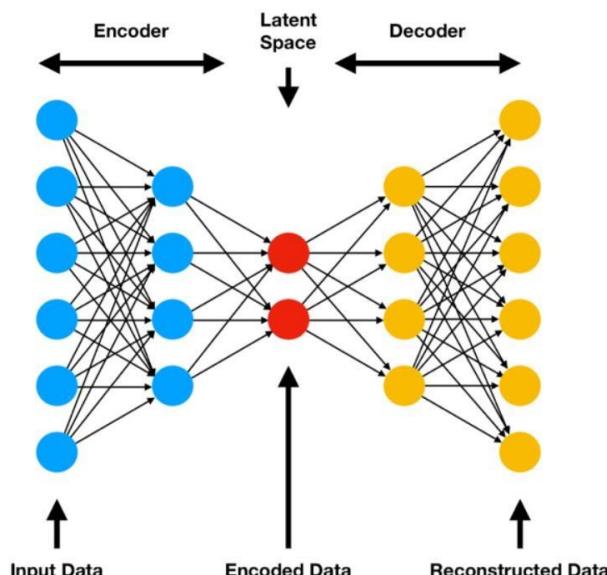
Autoencoders

Academic Year 2021-22

SAP ID:



Autoencoder is an unsupervised artificial neural network that compresses the data to lower dimension and then reconstructs the input back. Autoencoder finds the representation of the data in a lower dimension by focusing more on the important features getting rid of noise and redundancy. It's based on Encoder-Decoder architecture, where encoder encodes the high-dimensional data to lower-dimension and decoder takes the lower-dimensional data and tries to reconstruct the original high-dimensional data.



Tasks to be performed:

1. Use the Iris Dataset present in the scikit-learn library
2. Create an Auto Encoder and fit it with our data using 3 neurons in the dense layer
3. Use encoded layer to encode the training input
4. Plot loss for different encoders [PCA, Linear Autoencoder, Sigmoid based NonLinear Autoencoder, ReLU based Non-Linear Auto encoder]

Link: https://colab.research.google.com/drive/1n7aUhoBCC3ZOZW_r-SnrSYgxTpigDG5?usp=sharing

```

Assignments | Microsoft Teams x jhanviparekh-2702/Information x CD 60009210033_JhanviParekh_ML x + 
← → colab.research.google.com/drive/1n7alUh0BCC3ZQZW_r-SnStgxTpigD-G5#scrollTo=syhybVipOpis
File Edit View Insert Runtime Tools Help
+ Code + Text
import keras
import pickle
import pandas as pd
import numpy as np
from keras.models import Model,load_model
from keras.layers import Input,Dense
from keras.callbacks import ModelCheckpoint
from keras import regularizers
from keras.optimizers import Adam
from sklearn import datasets
from sklearn import decomposition
from sklearn.cluster import KMeans
from sklearn.preprocessing import MinMaxScaler
from sklearn import metrics
import matplotlib.pyplot as plt

RANDOM_SEED=371717
np.random.seed(RANDOM_SEED)

[2] iris=datasets.load_iris()

X=iris.data
y=iris.target
print("X:",X[0])
target_names=iris.target_names

scaler=MinMaxScaler()
scaler.fit(X)
X_scaled=scaler.transform(X)

def plot3clusters(x,title,vtitle):
    plt.figure()
    colors=['navy','turquoise','darkorange']
    l=2
    for color,i,target_name in zip(colors,[0,1,2],target_names):
        plt.scatter(X[y==i,0],X[y==i,1],color=color,alpha=1.,lw=l,label=target_name)
    plt.legend(loc='best',shadow=False,scatterpoints=1)
    plt.title(title)
    plt.xlabel(vtitle+"1")
    plt.ylabel(vtitle+"2")
    plt.show()

Executing (2m 30s) <cell line: 43> > error_handler() > fit() > error_handler() > evaluate() > enumerate_epochs() > __iter__() > __init__() > _create_iterator() > make_iterator()

33°C Smoke 12:31 18-11-2023

Assignments | Microsoft Teams x jhanviparekh-2702/Information x CD 60009210033_JhanviParekh_ML x + 
← → colab.research.google.com/drive/1n7alUh0BCC3ZQZW_r-SnStgxTpigD-G5#scrollTo=syhybVipOpis
File Edit View Insert Runtime Tools Help Saving...
+ Code + Text
[1] import matplotlib.pyplot as plt
RANDOM_SEED=371717
np.random.seed(RANDOM_SEED)

[2] iris=datasets.load_iris()

X=iris.data
y=iris.target
print("X:",X[0])
target_names=iris.target_names

scaler=MinMaxScaler()
scaler.fit(X)
X_scaled=scaler.transform(X)

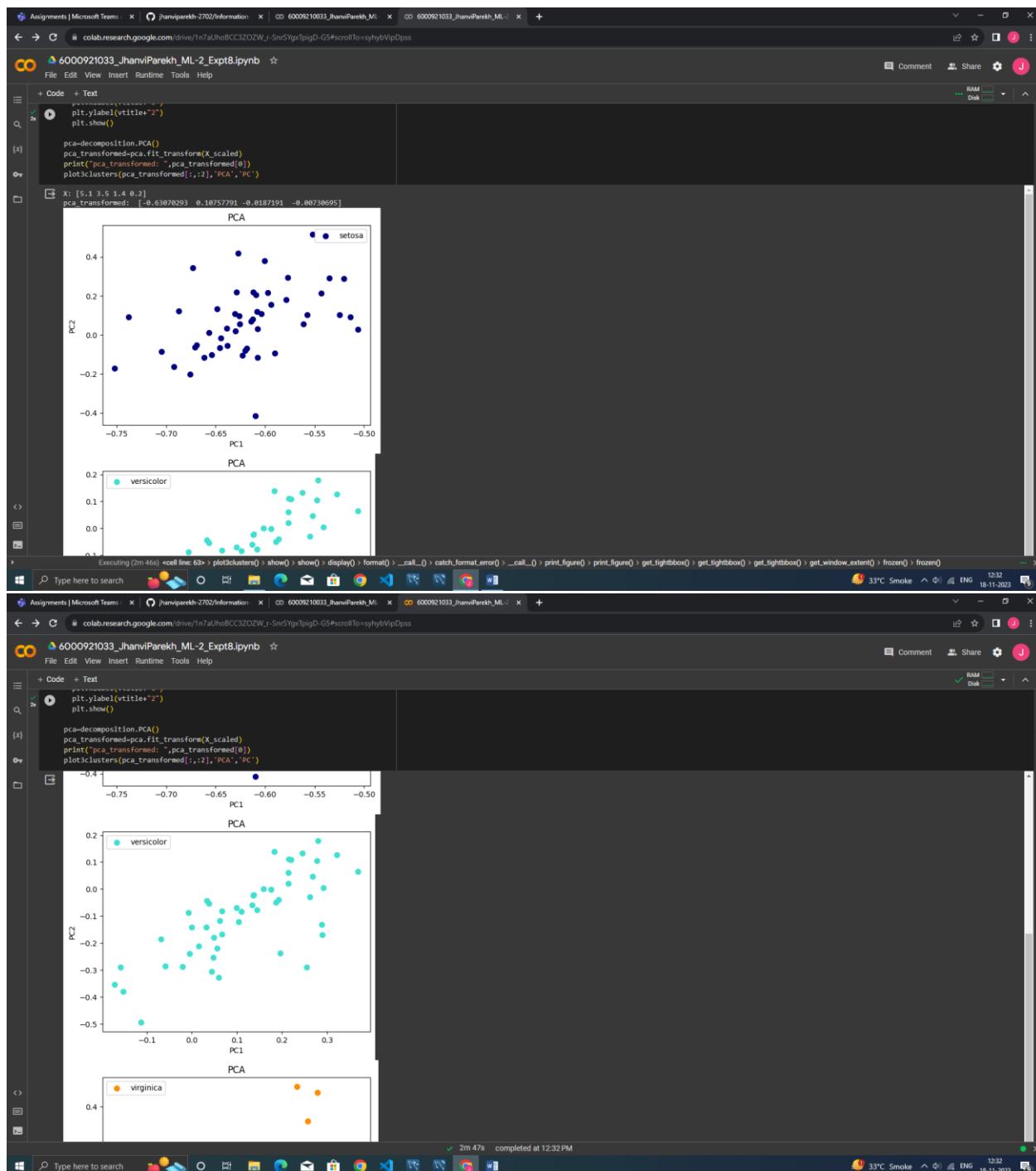
def plot3clusters(x,title,vtitle):
    plt.figure()
    colors=['navy','turquoise','darkorange']
    l=2
    for color,i,target_name in zip(colors,[0,1,2],target_names):
        plt.scatter(X[y==i,0],X[y==i,1],color=color,alpha=1.,lw=l,label=target_name)
    plt.legend(loc='best',shadow=False,scatterpoints=1)
    plt.title(title)
    plt.xlabel(vtitle+"1")
    plt.ylabel(vtitle+"2")
    plt.show()

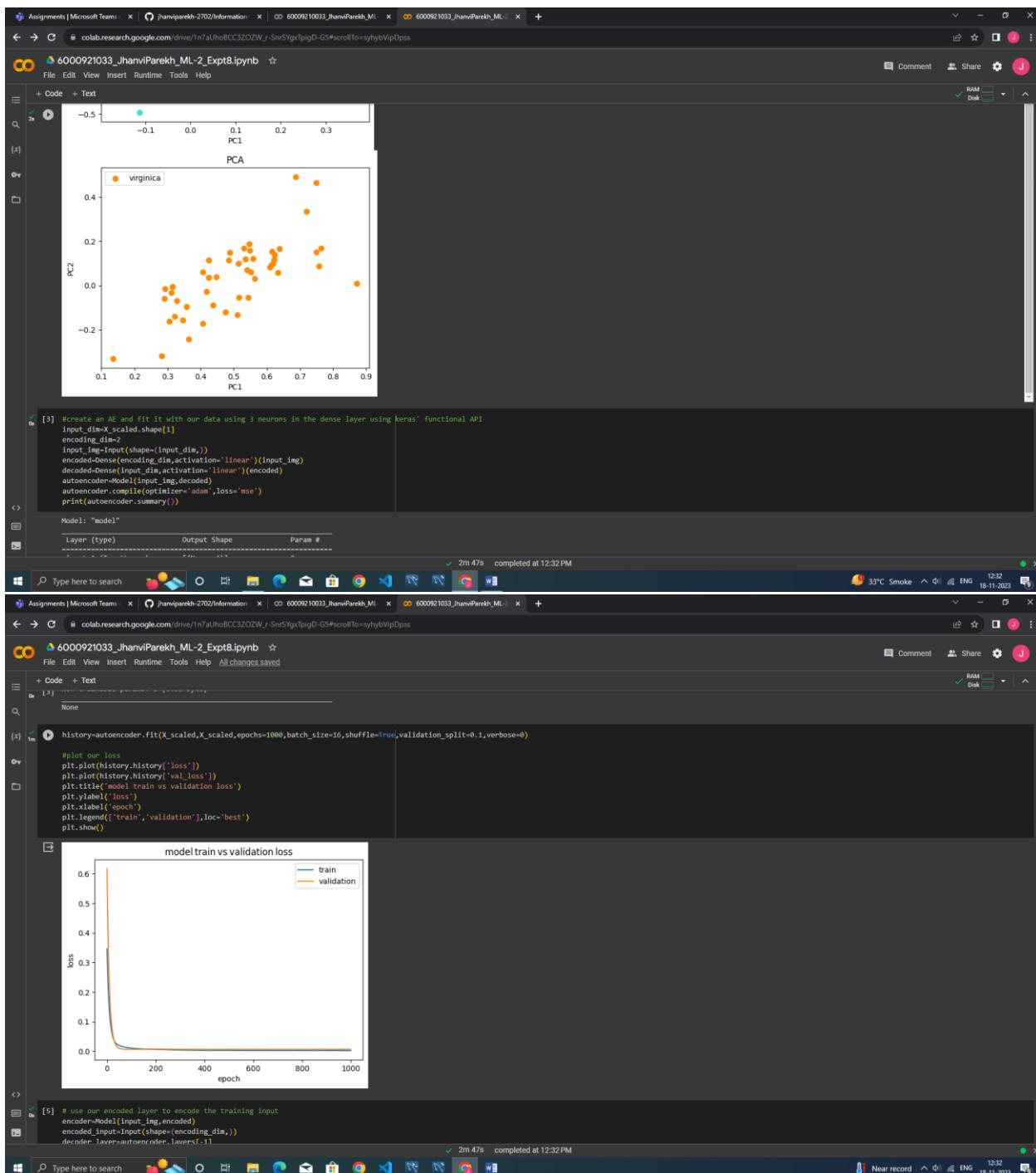
pca=decomposition.PCA()
pca_transformed=pca.fit_transform(X_scaled)
print(pca_transformed[:,:2])
plot3clusters(pca_transformed[:,2],'PCA','PC')

PCA
versicolor
-0.4
-0.75 -0.70 -0.65 -0.60 -0.55 -0.50
PC1
0.2
0.1
Executing (2m 36s) <cell line: 43> > error_handler() > fit() > error_handler() > __call__() > call_function() > call_flat() > flat_call() > __call__() > call_function() > quick_execute()

33°C Smoke 12:32 18-11-2023

```





```

Assignments | Microsoft Teams | jhanviparekh-2702/Information | 60009210033_JhanviParekh_ML | 6000921033_JhanviParekh_ML | + - _ X
← → colab.research.google.com/drive/1n7alUh0BCC3ZQZW_r-SnrSYgxTpigD-G5#scrollTo=syhybVipOpis
File Edit View Insert Runtime Tools Help All changes saved
Comment Share J RAM Disk
+ Code + Text
1m 1s epoch
Search
[+] # use our encoded layer to encode the training input
encoder=Model(input_img,encoded)
encoded_input=Input(shape=(encoding_dim,))
decoder_layer=autoencoder.layers[-1]
decoder=Model(encoded_input,decoder_layer(encoded_input))
encoded_data=encoder.predict(X_scaled)
plot3clusters(encoded_data[:,2:],['Linear AE','AE'])

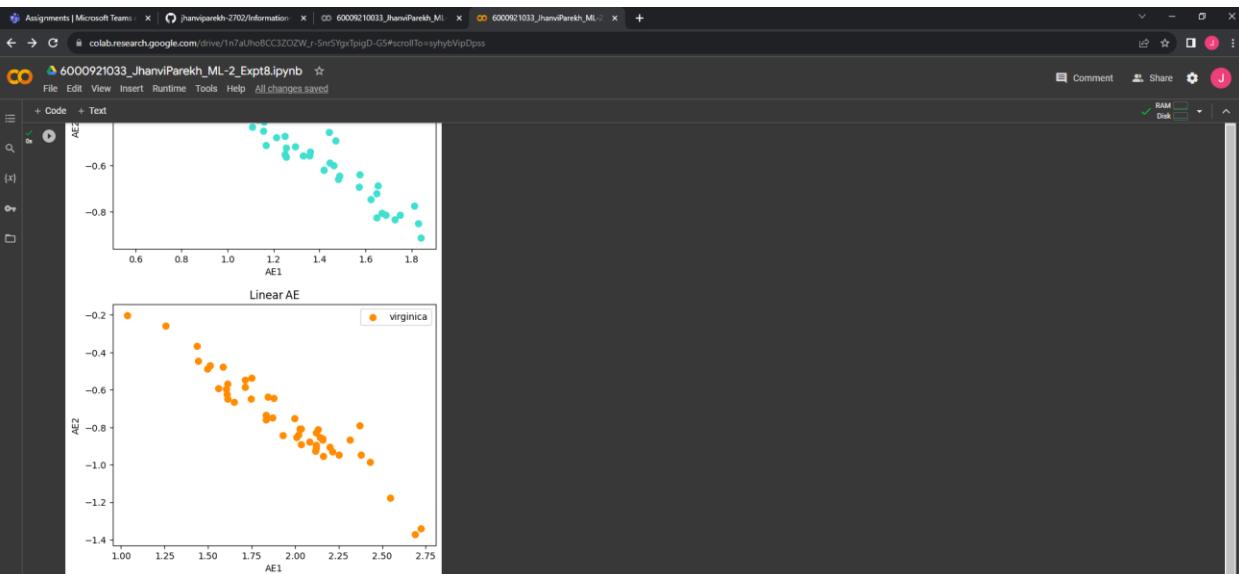
5/5 [=====] - 0s 2ms/step
Linear AE

Linear AE
setosa
versicolor
virginica
AE2
AE1
2m 47s completed at 12:32 PM
Near record ^ ⌂ DNG 12:32 18-11-2023
Assignments | Microsoft Teams | jhanviparekh-2702/Information | 60009210033_JhanviParekh_ML | 6000921033_JhanviParekh_ML | + - _ X
← → colab.research.google.com/drive/1n7alUh0BCC3ZQZW_r-SnrSYgxTpigD-G5#scrollTo=syhybVipOpis
File Edit View Insert Runtime Tools Help All changes saved
Comment Share J RAM Disk
+ Code + Text
1m 1s epoch
Search
[+] # use our encoded layer to encode the training input
encoder=Model(input_img,encoded)
encoded_input=Input(shape=(encoding_dim,))
decoder_layer=autoencoder.layers[-1]
decoder=Model(encoded_input,decoder_layer(encoded_input))
encoded_data=encoder.predict(X_scaled)
plot3clusters(encoded_data[:,2:],['Linear AE','AE'])

5/5 [=====] - 0s 2ms/step
Linear AE

Linear AE
versicolor
virginica
AE2
AE1
2m 47s completed at 12:32 PM
Near record ^ ⌂ DNG 12:32 18-11-2023

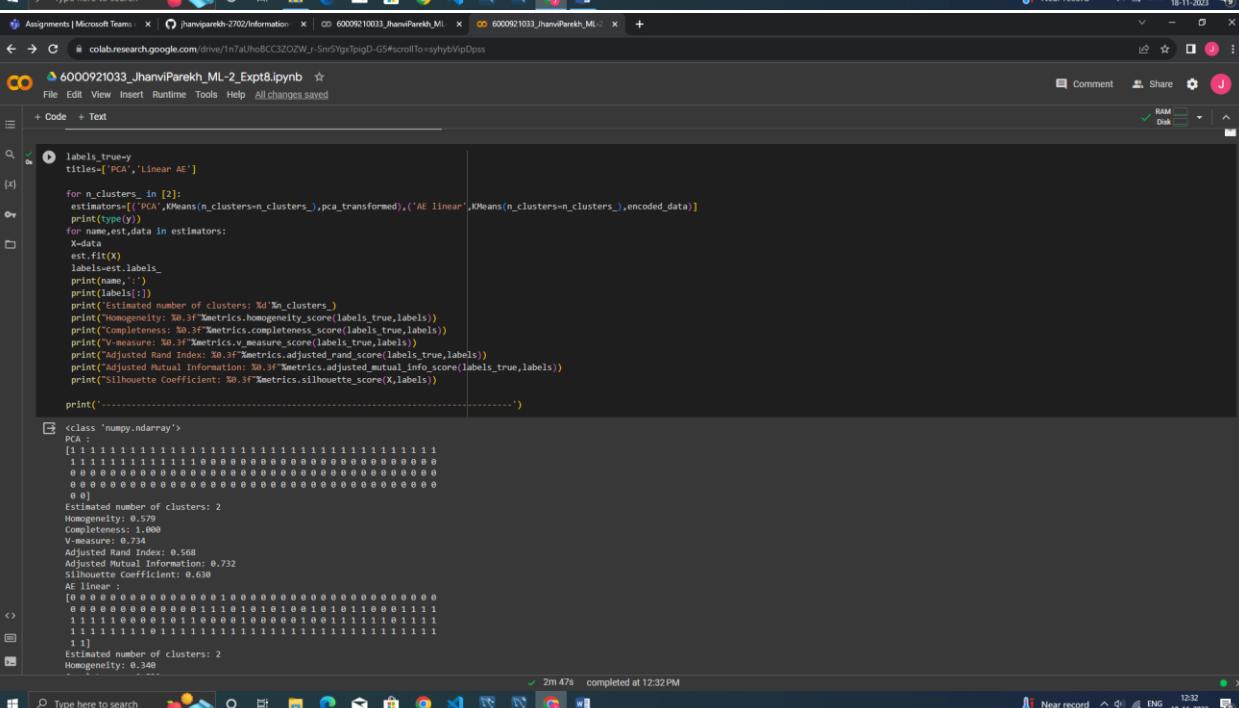
```



```
[6]: labels_true=y
titles=['PCA','Linear AE']

for n_clusters_ in [2]:
    estimators=[('PCA',KMeans(n_clusters=n_clusters_),pca_transformed),('AE linear',KMeans(n_clusters=n_clusters_),encoded_data)]
    plot_kmeans(estimators,titles)
```

✓ 2m 47s completed at 12:52 PM



Assignments | Microsoft Teams | jhanviparekh_2702/information | 60009210033_JhanviParekh_ML | 6000921033_JhanviParekh_ML | +

cola.research.google.com/drive/n7rUho8C3zQZW_rSmSYgxTpigD-G5#scrollTo=yhybVipDps

6000921033_JhanviParekh_ML-2_Expt8.ipynb

File Edit View Insert Runtime Tools Help All changes saved

+ Code + Text

#create an AE and fit it with our data using 3 neurons in the dense layer using keras' functional API

```
input_dim=X_scaled.shape[1]
encoding_dim=2
input_img2=Input(shape=(input_dim,))
encoded2=Dense(encoding_dim,activation='sigmoid')(input_img2)
decoded2=Dense(input_dim,activation='sigmoid')(encoded2)
autoencoder2=Model(input_img2,decoded2)
autoencoder2.compile(optimizer='adam',loss='mse')
print(autoencoder2.summary())
history2=autoencoder2.fit(X_scaled,X_scaled,epochs=2000,batch_size=16,shuffle=True,validation_split=0.1,verbose=0)
```

#plot our loss

```
plt.plot(history2.history['loss'])
plt.plot(history2.history['val_loss'])
plt.title('model train vs validation loss')
plt.ylabel('loss')
plt.xlabel('epoch')
plt.legend(['train','validation'],loc='best')
plt.show()
```

use our encoded layer to encode the training input

```
encoder2=Model(input_img2,encoded2)
encoded_input2=Input(shape=(encoding_dim,))
decoder_layer2=autoencoder2.layers[-1]
decoder2=Model(encoded_input2,decoder_layer2(encoded_input2))
encoded_data2=encoder2.predict(X_scaled)

plotIClusters(encoded_data2[:,2:],'Non-Linear sigmoid-based AE','AE')
```

#create an AE and fit it with our data using 3 neurons in the dense layer using keras' functional API

```
input_dim=X_scaled.shape[1]
encoding_dim=2
input_img3=Input(shape=(input_dim,))
encoded3=Dense(encoding_dim,activation='relu',activity_regularizer=regularizers.l1(10e-5))(input_img3)
decoded3=Dense(input_dim,activation='sigmoid')(encoded3)
autoencoder3=Model(input_img3,decoded3)

autoencoder3.compile(optimizer='adam',loss='mse')
print(autoencoder3.summary())
```

```

+ Code + Text
(encoded3=dense(encoding_dim3,activation='relu',activity_regularizer=regularizers.l1(10e-5))(input_img3)
decoded3=dense(input_dim3,activation='sigmoid')(encoded3)
autoencoder3=Model(input_img3,decoded3)

autoencoder3.compile(optimizer='adam',loss='mse')
print(autoencoder3.summary())

history3=autoencoder3.fit(X_scaled,X_scaled,epochs=400,batch_size=16,shuffle=True,validation_split=0.1,verbose=0)

#plot on loss
plt.plot(history3.history['loss'])
plt.plot(history3.history['val_loss'])
plt.title('model train vs validation loss')
plt.ylabel('loss')
plt.xlabel('epoch')
plt.legend(['train','validation'],loc='best')

plt.show()

#use the encoded layer to encode the training input
encoder3=Model(input_img3,encoded3)
encoded_input3=Input(shape=(encoding_dim3,))
decoder_layer3=autoencoder3.layers[-1]
decoder3=Model(encoded_input3,decoder_layer3(encoded_input3))
encoded_data3=encoder3.predict(X_scaled)

plot3clusters(encoded_data3[:,2], 'Non-Linear relu-based AE', 'AE')
plot3clusters(encoded_data3[:,2], 'PCA', 'PC')
plot3clusters(encoded_data3[:,2], 'Linear AE', 'AE')
plot3clusters(encoded_data3[:,2], 'Non-Linear sigmoid-based AE', 'AE')
plot3clusters(encoded_data3[:,2], 'Non-Linear relu-based AE', 'AE')

```

```

+ Code + Text
plot3clusters(encoded_data[:,2], 'Linear AE', 'AE')
plot3clusters(encoded_data[:,2], 'Non-Linear sigmoid-based AE', 'AE')
plot3clusters(encoded_data[:,2], 'Non-Linear relu-based AE', 'AE')

Model: "model_3"
-----  

Layer (type)      Output Shape       Param #
Input_3 (InputLayer) [(None, 4)]        0
dense_2 (Dense)   (None, 2)           10
dense_3 (Dense)   (None, 4)           12
-----  

Total params: 22 (88.00 Byte)
Trainable params: 22 (88.00 Byte)
Non-trainable params: 0 (0.00 Byte)
-----  

None

```

