Assignment - 4

TITLE:
Statistical Processing.

PROBLEM STATEMENT:
Load Apply Basic PCA on the iris dataset.
- Describe the data set.
- Describe the structured of correlation among variables.
- Compute a PCA with maximum number of components.
- Compute commutative explained variance ratio. Determine number of component K by your computed values.
- Print K principle components and correlations of K principle component with original variables.
- Plot the samples projected into the K first PCs.
- Color sample by the species.

OBJECTIVES:
To be able to appropriately apply computational methodologies to real world statistical problems.

OUTCOMES:
Model and solve computing problem using correlation, and resampling using appropriate statistics algorithm.

PREREQUISITE:
→ Concept of data pre-processing.

THEORY:

PCA finds the principle components of data. It is often useful to measure data in terms of its principle components rather than on normal x-y axis. They are the directions where there is most variance, the directions where data is most spread out.

Principle component Analysis (PCA) is unsupervised reduction algorithm. It identifies hyperplane that lies closest to the data, and then it projects the data onto it preserving the variance.

PCA finds a new set of dimensions such that all dimensions are orthogonal and ranked according to variance of data along them. It means more important principle axis occurs first.

How does PCA works?

- Calculate covariance matrix X of data points.
- Calculate eigen vectors and corresponding eigen values.
- Select eigen vector according to decreasing eigen values.
- Choose first k eigen vectors and that will be new k dimensions.
- Transform the original n dimensional data points into k dimensions.

Implementing PCA on a 2-D dataset

→ Step-1: Normalise the data

First step is to normalise the data that we have so that PCA works properly.

→ Step-2: Calculate the covariance matrix

Since the dataset we took is 2-dimensional this will result in a 2×2 covariance matrix

$$\text{Matrix (Covariance)} = \begin{bmatrix} Var[X_1] & Cov[X_1, X_2] \\ Cov[X_2, X_1] & Var[X_2] \end{bmatrix}$$

→ Step-3: Calculate the eigen values and eigen vectors.

$$\det(\lambda I - A) = 0$$

$$(\lambda I - A)v = 0$$

→ Step-4: Choosing components and forming a feature vector.

We do not lose out some information in the process, but if eigen values are small, we do not lose much.

CONCLUSION:

Hence we conclude that, Principal Component Analysis (PCA) is useful to visualise the high-dimensional datasets, ~~as~~ it can compress it down to 2-dimensions.

AIML Assignment4.ipynb - Colab | Combine your Documents and In

colab.research.google.com/drive/1RFUigzH-fFLox70DOVffKN6xGW7zVRwQ

Apps HackerRank W3Schools Elearn Codeforces Django ERDPlus LeetCode Tinkercad Coursera My Captain Google Tech Dev G... Reading list

AIML Assignment4.ipynb

File Edit View Insert Runtime Tools Help Last edited on January 10

Comment Share

Files

.. sample_data

+ Code + Text Initializing Editing

Apply Basic PCA on the iris dataset. The data set is available at: https://raw.github.com/neurospin/pystatsml/master/datasets/iris.csv • Describe the data set. Should the dataset been standardized? • Describe the structure of correlations among variables. • Compute a PCA with the maximum number of components . • Compute the cumulative explained variance ratio. Determine the number of components�� by your computed values. • Print the�� principal components directions and correlations of the�� principal compo-nents with the original variables. Interpret the contribution of the original variables into the PC. • Plot the samples projected into the�� first PCs. • Color samples by their species.

```
[ ] import numpy as np
    import pandas as pd
    import seaborn as sns
    import matplotlib.pyplot as plt
```

```
[ ] iris_data = pd.read_csv("/content/drive/MyDrive/data/iris.csv")
    iris_data.head()
```

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa

Mounting Google Drive...

TCOB41 AIML Assi....pdf TCOB41 AIML Assi....pdf Show all

Type here to search 24°C Light rain ENG 13:41 13-01-2022

AIML Assignment4.ipynb

3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

Standardization 1)Each feature in the dataset has different mean and SD hence it is necessary to scale data for comparision with other features.

2)Standardization sets mean to zero and SD of 1 for all features.

3)As seen in the data below the SD and is different for each datasetalso the variation is different.

```
print("Sepal length range: [%s, %s]" % (min(iris_data["sepal_length"]),max(iris_data["sepal_length"])))
print("Sepal width range: [%s, %s]" % (min(iris_data["sepal_width"]),max(iris_data["sepal_width"])))
print("Petal length range: [%s, %s]" % (min(iris_data["petal_length"]),max(iris_data["petal_length"])))
print("Petal width range: [%s, %s]" % (min(iris_data["petal_width"]),max(iris_data["petal_width"])))

print("Sepal Length standard deviation :\t %f" %np.std(iris_data["sepal_length"]))
print("Sepal Width standard deviation :\t %f" %np.std(iris_data["sepal_width"]))
print("Petal Length standard deviation :\t %f" %np.std(iris_data["petal_length"]))
print("Petal Width standard deviation :\t %f" %np.std(iris_data["petal_width"]))
```

```
Sepal length range: [4.3, 7.9]
Sepal width range: [2.0, 4.4]
Petal length range: [1.0, 6.9]
Petal width range: [0.1, 2.5]
Sepal Length standard deviation :        0.825301
```

AIML Assignment4.ipynb - Colab | Combine your Documents and In

colab.research.google.com/drive/1RFUigzH-fFLox70DOVffKN6xGW7zVRwQ

Apps HackerRank W3Schools Elearn Codeforces Django ERDPlus LeetCode Tinkercad Coursera My Captain Google Tech Dev G... Reading list

AIML Assignment4.ipynb

File Edit View Insert Runtime Tools Help Last [redacted] on [redacted]uary 10

Comment Share

Files

drive

sample_data

+ Code + Text RAM Disk

```
Petal Width standard deviation :        0.759693
```

Correlation Matrix :

```
[ ] corr = iris_data.corr(method ="pearson")
    display(corr)
    sns.heatmap(corr,cmap =sns.diverging_palette(10,0,as_cmap=True),annot= True,fmt = "f")
```

	sepal_length	sepal_width	petal_length	petal_width
sepal_length	1.000000	-0.117570	0.871754	0.817941
sepal_width	-0.117570	1.000000	-0.428440	-0.366126
petal_length	0.871754	-0.428440	1.000000	0.962865
petal_width	0.817941	-0.366126	0.962865	1.000000

```
&lt;matplotlib.axes._subplots.AxesSubplot at 0x7f9ee30b5750&gt;
```

Disk 65.98 GB available

TCOB41 AIML Assi....pdf TCOB41 AIML Assi....pdf Show all

Type here to search 24°C Light rain ENG 13:42 13-01-2022

1)Sepal_length <-> Petal_Length : 0.871754 (Positive Correlation)

2)Sepal_width <-> Petal_Length : -0.428440 (Negative Correlation)

3)Petal_length <-> Petal_width : 0.962865 (Highly Positive Correlation)

4)Petal_width <-> Sepal_Length : 0.817941 (Positive Correlation)

PCA (Principal Component Analysis):

```
[ ] from sklearn.decomposition import PCA
    pca = PCA()
    x_new1 =pca.fit_transform(iris_data.drop(["species"],axis =1))
    x_new1[:5]
```

```
array([[-2.68412563e+00,  3.19397247e-01, -2.79148276e-02,
        -2.26243707e-03],
       [-2.71414169e+00, -1.77001225e-01, -2.10464272e-01,
        -9.90265503e-02],
       [-2.88899057e+00, -1.44949426e-01,  1.79002563e-02,
        -1.99683897e-02],
       [-2.74534286e+00, -3.18298979e-01,  3.15593736e-02,
         7.55758166e-02],
       [-2.72871654e+00,  3.26754513e-01,  9.00792406e-02,
         6.12585926e-02]])
```

Explained Variance of R^2:

AIML Assignment4.ipynb - Colab | Combine your Documents and In

colab.research.google.com/drive/1RFUigzH-fFLox70DOVffKN6xGW7zVRwQ

Apps HackerRank W3Schools Elearn Codeforces Django ERDPlus LeetCode Tinkercad Coursera My Captain Google Tech Dev G... Reading list

AIML Assignment4.ipynb

File Edit View Insert Runtime Tools Help Last [redacted] on [redacted]uary 10

Comment Share

Files

+ Code + Text

RAM Disk

..
drive
sample_data

Explained Variance of R^2:

```
[ ] explained_variance= pca.explained_variance_ratio_

    explained_variance
```

```
array([0.92461872, 0.05306648, 0.01710261, 0.00521218])
```

```
[ ] plt.figure(figsize =(6,4))
    plt.bar(range(4),explained_variance, alpha=0.5, align='center', label ="Individual Explained Variance", color ="darkblue")
    plt.ylabel("Explained Variance Ratio")
    plt.xlabel("Principal [redacted]t")
    plt.legend(loc ="best")
    plt.tight_layout()
```

Disk 65.98 GB available

TCOB41 AIML Assi....pdf TCOB41 AIML Assi....pdf Show all

Type here to search 24°C Polluted air ENG 13:42 13-01-2022

There are 3 Principal Components. Applying PCA

```
[ ] pca  = PCA(n_components=3)
    x_new =pca.fit_transform(iris_data.drop(['species'],axis =1))
    x_new[:5]
```

```
array([[-2.68412563,  0.31939725, -0.02791483],
       [-2.71414169, -0.17700123, -0.21046427],
       [-2.88899057, -0.14494943,  0.01790026],
       [-2.74534286, -0.31829898,  0.03155937],
       [-2.72871654,  0.32675451,  0.09007924]])
```

Correlation and Direction of PCA converting categorial data to numerical

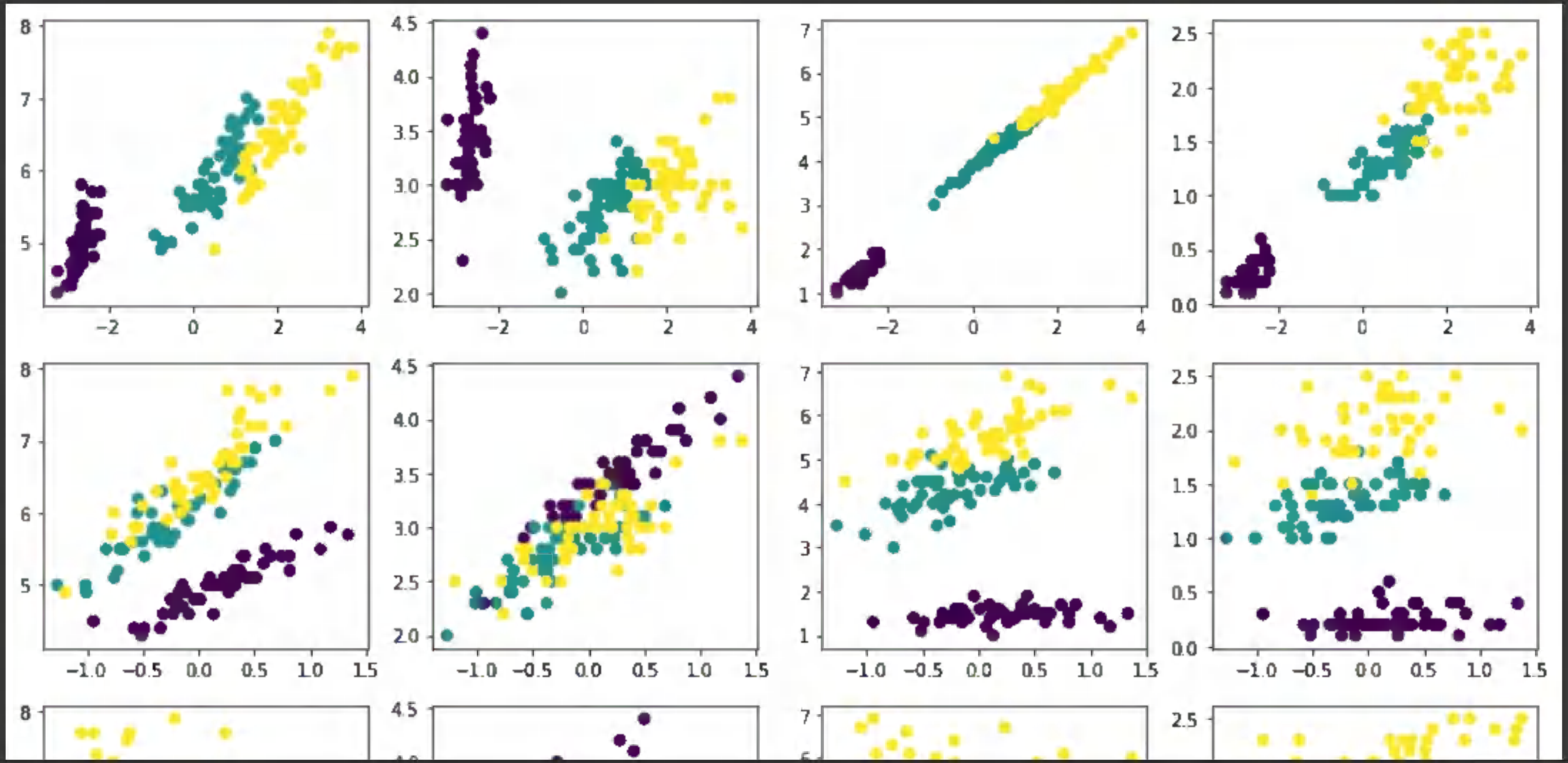
```
[ ] categ_num ={"species":{"setosa":0,"versicolor":1,"virginica":2}}
    iris_data1 = iris_data.replace(categ_num)
    columns = list(iris_data.columns[:4])
```

Comparing Principal Components with original variables

```
[ ] fig, axes =plt.subplots(3,4,figsize =(15,10))
    k= 0
    for i in range(axes.shape[0]):
```

AIML Assignment4.ipynb

```
k= 0
for i in range(axes.shape[0]):
    for j in range(axes.shape[1]):
        axes[i,j].scatter(x_new[:,i],iris_data[columns[j]],c =iris_data1["species"])
plt.show()
```




AIML Assignment4.ipynb - Colab
Combine your Documents and Im
colab.research.google.com/drive/1RFUigzH-fFLox70DOVffKN6xGW7zVRwQ
Apps HackerRank W3Schools Elearn Codeforces Django ERDPlus LeetCode Tinkercad Coursera My Captain Google Tech Dev G...
Reading list
AIML Assignment4.ipynb
File Edit View Insert Runtime Tools Help Last edited on January 10
Comment Share
Files
drive
sample_data
+ Code + Text
RAM
Disk
Editing
Disk 65.98 GB available
TCOB41 AIML Assi....pdf
TCOB41 AIML Assi....pdf
Show all
Type here to search
24°C Polluted air
ENG
13:42
13-01-2022