كلية الحوسبة والمعلوماتية |
College of Computing and Informatics

الجامعة السعودية الإلكترونية
SAUDI ELECTRONIC UNIVERSITY
2011-1432

Semester 1 – 2021/2022

Course Code	DS550
Course Name	Machine Learning
Assignment type	Critical Thinking
Module	04

Student ID	G200007615
Student Name	Abdulaziz Alqumayzi
CRN	15062

Solutions:

Critical Thinking Assignment 1

Introduction

In this activity, we will define and explain Principal Component Analysis (PCA) algorithm. Provide a python programming example on how to apply the algorithm on a dataset. The dataset used is **Credit Approval Data Set** from UCI Machine Learning Repository.

Data Set Information: This file is about applications for credit cards. In order to protect data confidentiality, all names and values of attributes have been replaced to meaningless symbols. This data collection is fascinating as a good blend of attributes - continuous, nominal with small value numbers and nominal with larger values.

Attribute information in Table 1.

Table 1

Attribute Information

Column	Data type
A1	a, b
A2	continuous
A3	continuous
A4	u, y, l, t
A5	g, p, gg
A6	c, d, cc, i, j, k, m, r, q, w,
A7	v, h, bb, j, n, z, dd, ff, o
A8	continuous
A9	t, f
A10	t, f
A11	continuous
A12	t, f
A13	g, p, s
A14	continuous
A15	continuous
A16	+,- (class attribute)

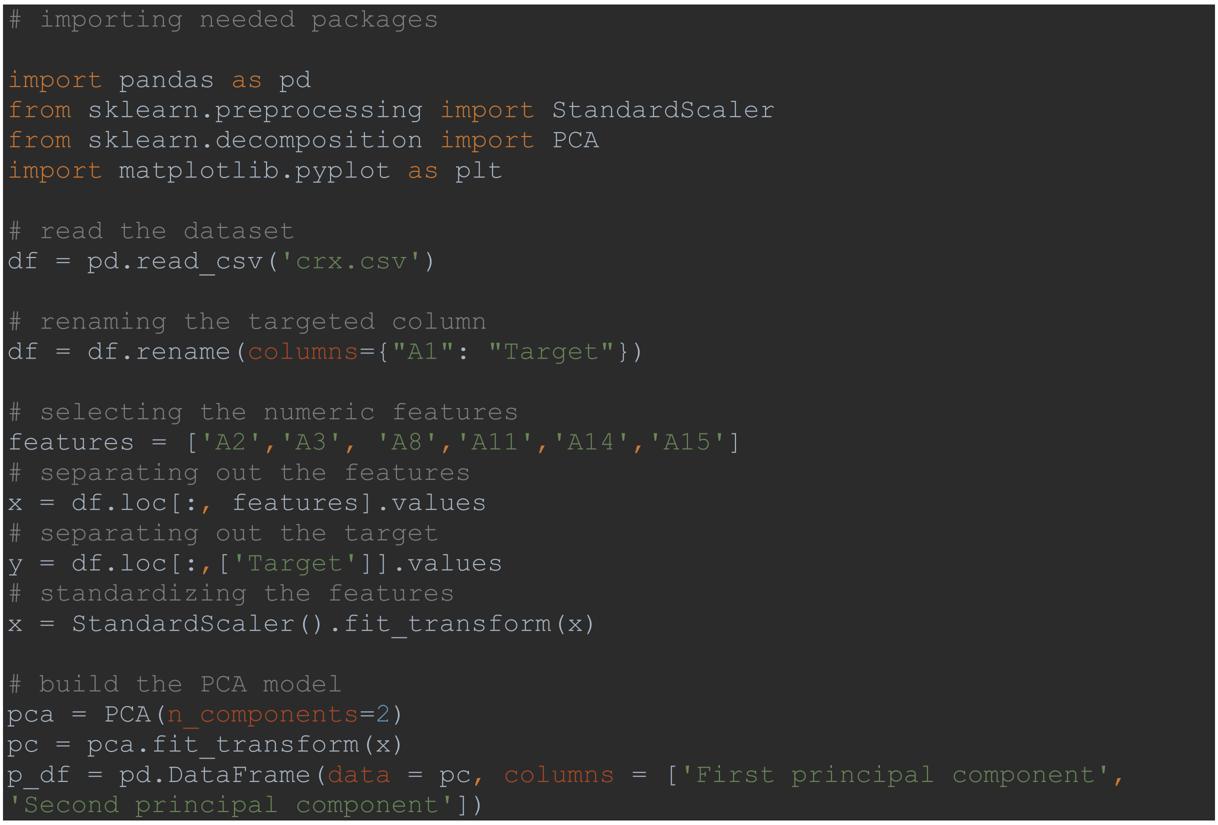
Principal Component Analysis (PCA)

The main component analysis is an unsupervised learning approach used for the reduction of dimensionality in machine learning. It is a statistical technique that, through orthogonal processing, turns observed correlated data into a number of linearly uncorrelated characteristics. The main components are dubbed these new altered features. It is one of the prominent tools for analyzing and predictive modeling exploratory data.

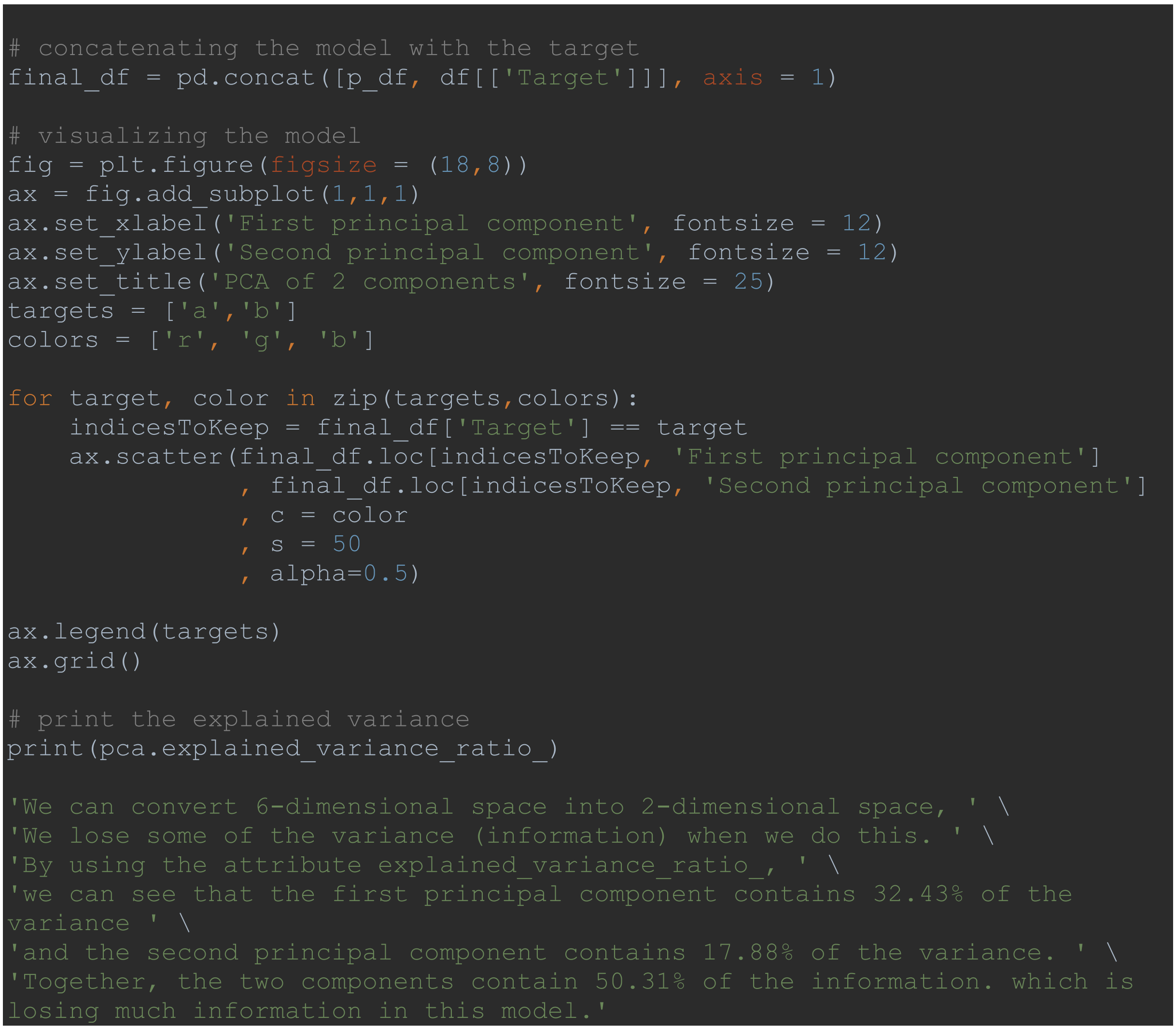
Applications of Principal Component Analysis: Mainly utilized in many AI applications such as computer vision, compression of images, etc. as dimensional reduction techniques. It can also be used to look for hidden patterns if the data are large. Some areas of PCA are finance, data mining, psychology, etc.

PCA's real-world applications are image processing, a film recommendation system, and the optimization of power distribution in various channels.

Python Code



```
# importing needed packages

import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt

# read the dataset
df = pd.read_csv('crx.csv')

# renaming the targeted column
df = df.rename(columns={"A1": "Target"})

# selecting the numeric features
features = ['A2','A3', 'A8','A11','A14','A15']
# separating out the features
x = df.loc[:, features].values
# separating out the target
y = df.loc[:,['Target']].values
# standardizing the features
x = StandardScaler().fit_transform(x)

# build the PCA model
pca = PCA(n_components=2)
pc = pca.fit_transform(x)
p_df = pd.DataFrame(data = pc, columns = ['First principal component',
'Second principal component'])
```



```
# concatenating the model with the target
final_df = pd.concat([p_df, df[['Target']]], axis = 1)

# visualizing the model
fig = plt.figure(figsize = (18,8))
ax = fig.add_subplot(1,1,1)
ax.set_xlabel('First principal component', fontsize = 12)
ax.set_ylabel('Second principal component', fontsize = 12)
ax.set_title('PCA of 2 components', fontsize = 25)
targets = ['a','b']
colors = ['r', 'g', 'b']

for target, color in zip(targets,colors):
    indicesToKeep = final_df['Target'] == target
    ax.scatter(final_df.loc[indicesToKeep, 'First principal component']
               , final_df.loc[indicesToKeep, 'Second principal component']
               , c = color
               , s = 50
               , alpha=0.5)

ax.legend(targets)
ax.grid()

# print the explained variance
print(pca.explained_variance_ratio_)

'We can convert 6-dimensional space into 2-dimensional space, ' \
'We lose some of the variance (information) when we do this. ' \
'By using the attribute explained_variance_ratio_, ' \
'we can see that the first principal component contains 32.43% of the
variance ' \
'and the second principal component contains 17.88% of the variance. ' \
'Together, the two components contain 50.31% of the information. which is
losing much information in this model.'
```

References:

Bonaccorso, G. (2018). *Machine learning algorithms: Popular algorithms for Data Science and Machine Learning*. Packt.

Quinlan. (n.d.). *Credit Approval Data Set*. UCI Machine Learning Repository: Credit Approval Data Set. Retrieved October 8, 2021, from https://archive.ics.uci.edu/ml/datasets/Credit+Approval.

Principal component analysis - javatpoint. www.javatpoint.com. (2021). Retrieved October 8, 2021, from https://www.javatpoint.com/principal-component-analysis.

Galarnyk, M. (2021, February 3). *PCA using Python (scikit-learn)*. Medium. Retrieved October 9, 2021, from https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60.