**PA TASK 2: Dimensionality Reduction Methods**

**D212 – Data Mining II**

Michael Lawson

Western Governors University

D212 – Data Mining II

PA Task 2: Dimensionality Reduction Methods

Kesselly Kamara

09/12/2022

Michael C. Lawson | Student ID: #001479557 | Program: MS Data Analytics (5/1/21) |

Program Mentor: Emil Stoica | 859-265-1888 (EST) | [mlaws77@wgu.edu](mailto:mlaws77@wgu.edu)

**Part I:**

A.

1.

Will running Principal Component Analysis (PCA) reveal features that contribute the most to the data and help eliminate features that are not important?

2.

The goal of running PCA on the medical data set will be to reveal features that contribute the most to patterns in the data and to eliminate features that are redundant or have less impact on the patterns in the data.

**Part II:**

B.

1. How PCA analyzes:

“Principal Component Analysis, PCA for short, is an unsupervised learning technique used to surface the core patterns in the data” (Bento, 2020). PCA is a dimensionality reduction technique for continuous features. PCA can be used on other types of data, but there are better methods for variables that are not continuous. PCA works to minimize variance using squared deviations. The concept of squared deviations does not apply well to binary and categorical variables.

PCA searches for patterns in the data and works to minimize variance. The features that contribute the most to the pattern are called principal components. The algorithm sorts the principal components in descending order starting with the feature that contributes to the pattern of the data the most. The principal components that contribute the most to the patterns in the data can then be separated for a new model that does not have the noise from features that were not contributing to the patterns used for prediction. The result is a much stronger predictive model.

2.  **One** assumption of PCA:

PCA assumes that the data set has some linearity, meaning the variables combine in a linear manner and have relationships among each other. Some data sets do not actually have a relationship or linearity, and PCA would not be the best route for feature reduction. Other data sets may have linearity that formed by chance, but the features do not actually have a relationship.

**Part III:**

C.  Data preparation:

1.



2.  Standardize:

*# Import libraries/packages*

**import** numpy **as** np

**import** pandas **as** pd

**from** pandas **import** Series, DataFrame

plt**.**style**.**use('ggplot')

**import** seaborn **as** sns

**import** matplotlib.pyplot **as** plt

**import** sklearn **from** sklearn.decomposition **import** PCA

**from** sklearn.preprocessing **import** StandardScaler, LabelEncoder

**from** mpl\_toolkits **import** mplot3d

*#Import data*

med\_file **=** r"C:\Users\mlaws\OneDrive - Western Governors University\Documents\WGU\D212\medical\_clean.csv"

df\_load **=** pd**.**read\_csv(med\_file)

df\_load**.**info()

Table

Description automatically generated with medium confidence

*# Separate out the continuous variables for PCA*

df **=** df\_load[['Income','VitD\_levels', 'Lat', 'Lng', 'Initial\_days','TotalCharge','Additional\_charges']]

*# View size of table*

df**.**shape



*# Scale data*

d **=** df**.**values

scaler **=** StandardScaler()

scaler**.**fit(d)

d\_scaled **=** scaler**.**transform(d)

*#Convert scaled data back to .csv and export cleaned data set*

med\_scaled\_for\_PCA **=** pd**.**DataFrame(d\_scaled)

med\_scaled\_for\_PCA**.**to\_csv(r"C:\Users\mlaws\OneDrive - Western Governors University\Documents\WGU\D212\medical\_scaled\_for\_PCA.csv")

**Part IV:**

D.

1.

*# Perform PCA*

pca\_7 **=** PCA(n\_components**=**7, random\_state**=**7)

pca\_7**.**fit(d\_scaled)

pc **=** pca\_7**.**transform(d\_scaled)

print(pc)

Text, letter

Description automatically generated

*#Create a data frame of only principal components*

pc\_df **=** pd**.**DataFrame(pc, columns **=** {'PC1','PC2','PC3','PC4','PC5', 'PC6','PC7'})

print(pc\_df)

Text

Description automatically generated

*# Cmopute the covariance matrix*

Load **=** pd**.**DataFrame(pca\_7**.**components\_**.**T,

columns**=**['PC1','PC2','PC3','PC4','PC5', 'PC6','PC7'],

index**=**df**.**columns)

Load

Text, table

Description automatically generated

*# Calculate variance explained by all 5 components*

print('Variance explained by all 7 principal components = ', sum(pca\_7**.**explained\_variance\_ratio\_**\***100))



*# Variance explained by each PC*

varex1 **=** pca\_7**.**explained\_variance\_ratio\_**\***100

var\_df1 **=** pd**.**DataFrame(varex1**.**round(2), columns**=**['Captured Variance Per PC'],

index **=** ['PC1','PC2','PC3','PC4','PC5', 'PC6','PC7'])

var\_df1

Table

Description automatically generated

*# Eigenvalues by principal component*

eigenvalues **=** pca\_7**.**explained\_variance\_

eigen\_df **=** pd**.**DataFrame(eigenvalues**.**round(4), columns**=**['Eigenvalues Per PC'], index**=**['PC1','PC2','PC3','PC4','PC5', 'PC6','PC7'])

eigen\_df

Table

Description automatically generated

*# Calculate cumulative sum of each principal component*

np**.**cumsum(pca\_7**.**explained\_variance\_ratio\_**\***100)



2.

The scree plot bends at an “elbow” and principal components before the sharp curve in the elbow are the features to keep. The medical data set produced a scree plot that is very straight with a sharp angled elbow. The first six features appear before the elbow. The shape of the scree plot curve indicates the data may not be the best for detecting variance or running PCA. The scree plot has a curve at the top rather than a sharp angle in most cases.

# Scree Plot

plt**.**plot(np**.**cumsum(pca\_7**.**explained\_variance\_ratio\_))

plt**.**xlabel('# Components')

plt**.**ylabel('Variance')

plt**.**savefig('elbow\_plot.png', dpi**=**100)

Chart, line chart

Description automatically generated

3. Variance of principal components

# Calculate cumulative sum of each principal component

np**.**cumsum(pca\_7**.**explained\_variance\_ratio\_**\***100)



# print the cumulative sum of each PC in an easier to read way

print('Variance explained by the first principal component =',

np.cumsum(pca\_7.explained\_variance\_ratio\_\*100)[0])

print('Variance explained by the first 2 principal components =',

np.cumsum(pca\_7.explained\_variance\_ratio\_\*100)[1])

print('Variance explained by the first 3 principal components =',

np.cumsum(pca\_7.explained\_variance\_ratio\_\*100)[2])

print('Variance explained by the first 4 principal components =',

np.cumsum(pca\_7.explained\_variance\_ratio\_\*100)[3])

print('Variance explained by the first 5 principal components =',

np.cumsum(pca\_7.explained\_variance\_ratio\_\*100)[4])

print('Variance explained by the first 6 principal components =',

np.cumsum(pca\_7.explained\_variance\_ratio\_\*100)[5])

Text

Description automatically generated

# run the PCA model on the most significant 6 principal components

pca2 = PCA(n\_components=6, random\_state=7)

pca2.fit(d\_scaled)

trans\_pca2 = pca2.transform(d\_scaled)

var=pca2.explained\_variance\_ratio\_\*100

var\_df = pd.DataFrame(var.round(2), columns = ['Explained Variance'], index=['PC1','PC2','PC3', 'PC4','PC5','PC6'])

var\_df

Table

Description automatically generated

# Create heat map of top 6 principal components

sns.heatmap(var\_df1, annot=True, linewidth=0.5, fmt='.1g')

plt.show

Chart, funnel chart

Description automatically generated

4.  T*otal* variance of principal components.

The first six principal components capture 99.83% of the variance in the data set. Only one feature was deemed insignificant. The first three principal components capture 58.9% of variance.

5.  Summary

From the 7 continuous variables in the data set, PCA discovered that six features account for 99.83% of the variance in the data set. With only one variable being eliminated and the scree plot having straight lines rather than curved, it appears the data in this set most likely does not have predictable or strong patterns even after using the principal components to create predictive models later. After the first principal component, the next five principal components contribute nearly the same to variance according to PCA. The recommendation for this data would be to find additional data or to throw out this data set and start fresh with new data.

**Part V: References**

Bento, Carolina. June 3, 2020. *Principal Component Analysis algorithm in Real-Life: Discovering patterns in a real-estate dataset.* <https://towardsdatascience.com/principal-component-analysis-algorithm-in-real-life-discovering-patterns-in-a-real-estate-dataset-18134c57ffe7>

BioTuring Team. June 18, 2018. *How to read PCA biplots and scree plots.* [https://bioturing.medium.com/how-to-read-pca-biplots-and-scree-plots-186246aae063#:~:text=Use%20a%20scree%20plot%20to,enough%20to%20describe%20the%20data.](https://bioturing.medium.com/how-to-read-pca-biplots-and-scree-plots-186246aae063%23:~:text=Use%20a%20scree%20plot%20to,enough%20to%20describe%20the%20data.)

Kamara, Kesselly. May 15, 2022. *pca\_webinar\_recording.* <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=60fa4159-94ba-4f41-9ba8-aea0011f18f9>