Submission - Exercise [4]

Visual Data Analysis

[Maryam Assaedi, maryam.assaedi@rwth-aachen.de]
[Mst. Mahfuja Akter , s6msakte@uni-bonn.de]
[Mahpara Hyder Chowdhury, s6machow@uni-bonn.de]

April 28, 2019

Exercise 1 (Principal Component Analysis)

a) Read the breast-cancer-wisconsin.xlsx. Note that there are some instances with missing data, which have to be imputed before we can run PCA. Pandas offers convenient functions for this. Apply an imputation method that makes sense for this dataset, and briefly explain your decision.
Answer:
We read the data and applied imputed method for missing data by zero. Here we see that all valid data are in range [1 10]. If we drop those row which contains missing element, we might miss some important data prediction. And again we can not apply any value which are within the range, it can mislead our prediction. Here we use the following command:

```
#a
import pandas as pd
df = pd.read_excel("breast-cancer-wisconsin.xlsx")

data = pd.isna(df)
df[data] = 0
print(df)
```

which printed the following:

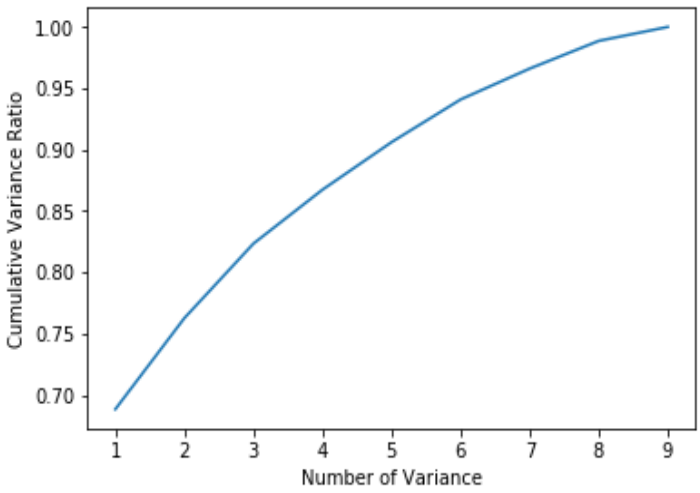
code	thickness	uniCelS	uniCelShape	marAdh	epiCelSize	bareNuc	\
1000025	5	1	1	1	2	1.0	
1002945	5	4	4	5	7	10.0	
1015425	3	1	1	1	2	2.0	
1016277	6	8	8	1	3	4.0	
1017023	4	1	1	3	2	1.0	
1017122	8	10	10	8	7	10.0	
1018099	1	1	1	1	2	10.0	
1018561	2	1	2	1	2	1.0	
1033078	2	1	1	1	2	1.0	
1033078	4	2	1	1	2	1.0	
1035283	1	1	1	1	1	1.0	
1036172	2	1	1	1	2	1.0	
1041801	5	3	3	3	2	3.0	
1043999	1	1	1	1	2	3.0	
1044572	8	7	5	10	7	9.0	
1047630	7	4	6	4	6	1.0	
1048672	4	1	1	1	2	1.0	

b) Create a plot that, for any number n, shows what fraction of the overall variance in the data is contained in the first n principal components. Make sure that you only include the nine relevant numerical attributes in the PCA, not the sample codes or class IDs. How many components do we need to cover more than 90 percent of the variance? (5P) Hint: You may use the implementation of PCA that is provided in the Python package scikit-learn.
Answer:
We have applied PCA and get variance for 9 numerical attributes. We see that we need to cover first 5 attributes to get more than 90 percent variance.
Here is the code below:

```
#b
import numpy as np
import seaborn as sns
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt
df2 = df.drop(['code','class'], axis=1)

pca  = PCA(n_components = 9)
principalComponents = pca.fit_transform(df2)
principalDf = pd.DataFrame(data = principalComponents, columns = ['principalComponent1','principalComponent2','principalComponent

x = [1,2,3,4,5,6,7,8,9]
y = np.cumsum(pca.explained_variance_ratio_)
print(np.cumsum(pca.explained_variance_ratio_))
plt.ylabel("Cumulative Variance Ratio")
plt.xlabel("Number of Variance")
plt.plot(x,y)
```

Here is the cumulative sum of 9 variance and plot the variance ratio against variance number.

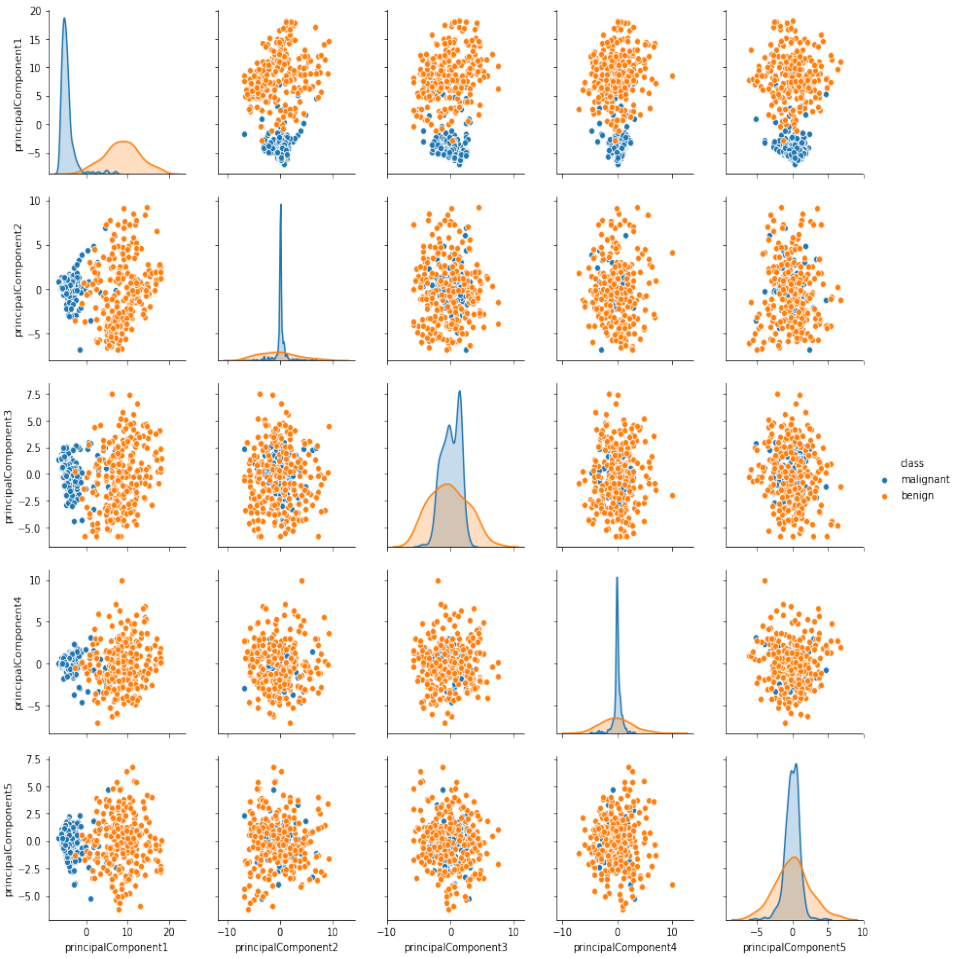
```
[0.68846327 0.76272114 0.82343024 0.86737153 0.90610211 0.94086948
 0.96603724 0.98861572 1.        ]

[<matplotlib.lines.Line2D at 0x124c98750f0>]
```



c) Each sample is now characterized by a point in PCA space. Create a scatter plot matrix that shows the first five principal components. Each diagonal cell should contain two overlaid density plots, one for the benign and one for the malignant class. Use different colors to distinguish between the classes, and add a legend that clearly states which samples are benign or malignant.
Answer:
We have plot first five PCA components and categorize them by labeling class as benign and malignant by different colors.
Here is the code below:

```
#c
newDf = principalDf.drop(['principalComponent6','principalComponent7','principalComponent8','principalComponent9'], axis=1)
df['class'].replace([2,4],['malignant','benign'], inplace = True)
finalDf = pd.concat([newDf, df[['class']]], axis = 1)
sns.pairplot(finalDf, hue = "class")
```

Here is the scatter plot of five principal components.



d) Which PCA mode shows the strongest difference between the benign and the malignant samples? Name the original variables that have the highest and lowest weights in its definition, respectively.

Answer:

From plot from c, we see that principalComponent1 gives the strongest difference between benign and the malignant samples.

We can see the highest weights from principalComponent1 and lowest value comes from principalComponent4 from above plot. Apart from visualization, we have applied some command to get the highest and lowest value and their corresponding original data variables.

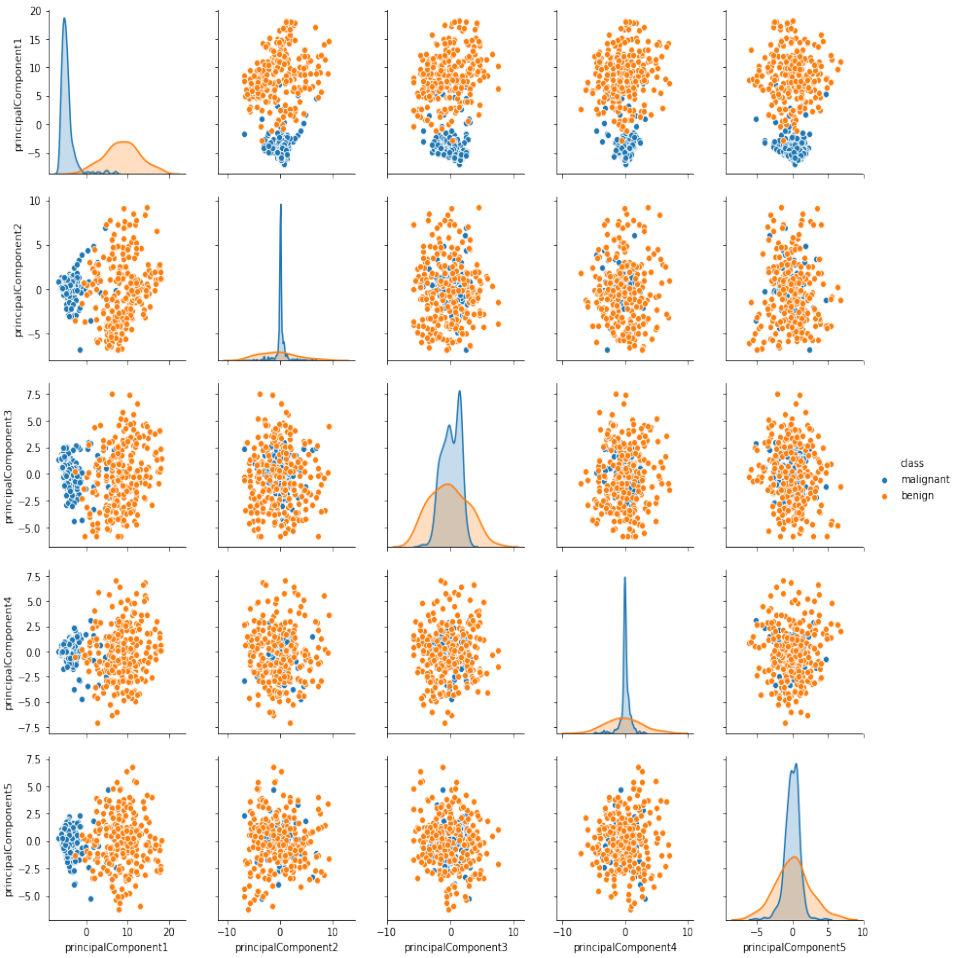
Here is the code below:

```
#d
print(max(finalDf['principalComponent1']))
print(min(finalDf['principalComponent4']))
maxweight = finalDf['principalComponent1']>18.13
minweight = finalDf['principalComponent4']<-7.07
print(finalDf[maxweight])
print(finalDf[minweight])
print(df.loc[425])
print(df.loc[288])
```

e) Scatterplot matrices often reveal outliers in the data. Visually identify at least one sample that is far away from the others, remove it from the dataset, and re-generate the scatterplot matrix without it.
Answer:
From plot we can see that one PCA value from principalComponent4 is far away from others data. This data is supposed to be maximum weights of that PCA. We drooped corresponding data from data set and replot again.
Here is the code below:

```
#e
maxBenignData = principalDf['principalComponent4']>9.8
print(finalDf[maxBenignData])
finalDf = finalDf.drop(finalDf.index[167])

sns.pairplot(finalDf, hue = "class")
```

Here is the scatter plot after drooping isolated data from dataset. And we see the plot, for principalComponent4 range of weight is decreased after drooping.



f) In the breast cancer dataset, all variables xi have a similar range, xi 2 [1; 10]. If the variables of a dataset have very different ranges, for example one variable x1 2 [1000; 2000] and another one x2 2 [1; 5], how would this affect the PCA? Could it make sense to pre-process the data in such cases? Why and how?

Answer:

If the ranges of variables varies much more then PCA effects. The variation ratio goes belong to large ranged variable. So we can not get accurate variance. It is make sense that, we have to pre-process those data into one form. We can apply normalized method for that type of imbalanced data.

g) Explain why, on this dataset, we cannot use Linear Discriminant Analysis (LDA) to create an alternative 5D embedding in which the classes are more clearly separated. Use the LDA imple- mentation in scikit-learn to perform a 1D LDA embedding and plot it (in a scatterplot) against the first principal component. Do they show a clear correlation? Is this true in general, or specific to the dataset?

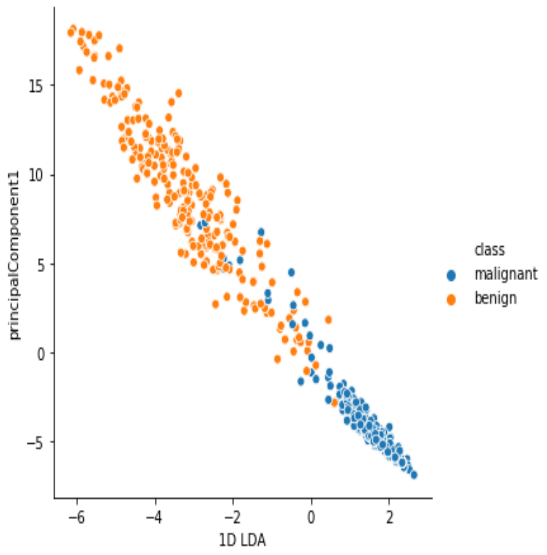
Answer:

It is not possible to create an alternative 5D embedding for this dataset because it requires more classs but we have only 2 classes (benign and malignant).

Here we have applied 1D LDA with all(9) numeric data from dataset. And plots that 1D LDA against PCA1 values.
Here is code below:

```
#g
import matplotlib.cm as cm
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis as LDA
X = df.iloc[:, 1:10].values
y = df.iloc[:, 10].values
lda = LDA(n_components=1)
lda = lda.fit(X,y).transform(X)

newLdaDf = pd.DataFrame(data = lda, columns = ['1D LDA'])
ldadf = pd.concat([newLdaDf, principalDf['principalComponent1'],df[['class']]], axis = 1)

ax = sns.relplot(x="1D LDA", y = "principalComponent1", hue="class", data = ldadf)
plt.show(ax)
```

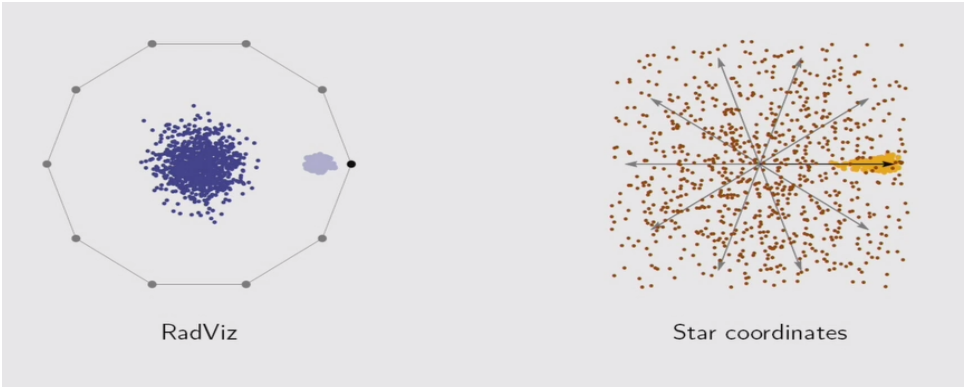
Here the scatter plot shows the clear correlation between PCA and LDA. It is supposed to be correlated to each other because the PCA and LDA is transformed from same dataset.



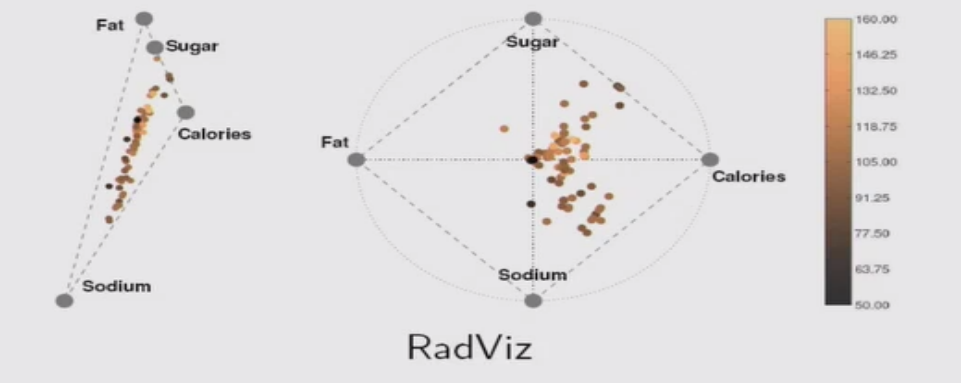
Exercise 2 (Comparing RadViz and Star Coordinates)

a) In general, the authors of this paper prefer star coordinates over RadViz. Briely explain two reasons for their preference.
Answer: The authors prefer SC over RadViz because of the ability to get the original high-dimensional attribute values from the visual elements in the plot more accurately.

In RadViz, the cluster in low-dimensional plot depends on shape, size and location. Whereas, SC resolves these problems regarding cluster size and overlaps by producing linear mapping.
b) Despite this, the authors mention a specifc use case where they consider RadViz to be superior to star coordinates. Briely explain what this use case is and why they prefer RadViz in this case.
Answer: RadViz is specially beneficial when the data is sparse.As RadViz pulls non-parse data close to the center and parse data close to the anchor points.



c) Neither RadViz nor star coordinates provide a one-to-one mapping between a data point's high- dimensional location and its two-dimensional projection. Interactively modifying the anchor points or axes, respectively, can reduce the resulting ambiguities. However, the paper mentions two examples in which, unlike star coordinates, RadViz introduces ambiguities that cannot be resolved with any re-arrangement of the anchor points. Point out these two examples. Could the extended RadViz that we learned about in the lecture resolve them?
Answer: In the paper, the authors have used a dataset for US breakfast cereal data with standardize four variables where using RadViz was not effective. As with those four variables it was pulling the points towards the vertices of the convex hull and according to the authors, those points cannot have same directions.Moreover, the convex hull was small so the ordering of the caloric content was degraded.
Despite of updating the anchor points, the resulting mapped data points with respect to the caloric content remains unordered.
Yes, by using extended RadViz, we can resolve the problem as it aims to reduce overlapping by extending the anchor points along the line segments.



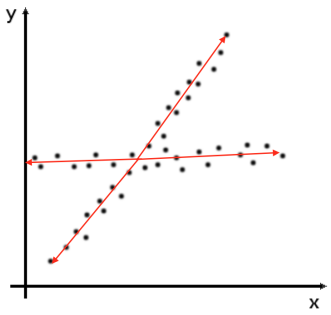
d) Several methods have been proposed to optimize the locations of anchor points in RadViz to obtain an improved separation of classes in the resulting projection. How did two such algorithms compare to supervised linear dimensionality reduction techniques in the experiments reported in this paper, in terms of (i) computational effort and (ii) achieved class separation?
Answer: Applying LDA to RidViz, the resulting plot separates the data very clearly. One the other hand, applying CDL to RidViz, the error rates for both the wine dataset and Olives dataset are consecutively 4.49 percent and 17.13 percent.

e) The authors propose that star coordinates could be combined with Linear Discriminant Analysis in order to perform a manual feature selection. Briefly explain how that approach works.
Answer: In SC short axis vectors refers possible candidates to be discarded, since the effect of a variable in a plot depends on the length of its axis vector. The classes can be separate while reducing the set of variables avoiding overlaps.

f) Briefly explain why anchor points in RadViz are commonly chosen such that they form a corner of the overall convex hull. Is this property also met for the star coordinate axis configuration in Fig. 12 (a)? Is it true in Fig. 12 (b)?
Answer: Anchor points in RadViz are commonly chosen such that they form a corner of the overall convex hull so that it can analyze sparse data efficiently. No, this property does not met for SC configuration. Here, for SC, the plot 12(a) reveals the importance of the variables and the error rate is much lower.

Exercise 3 (Dimensionality Reduction)

a) Given the following plot, sketch all principal modes (i.e., eigenvector directions) that would (approximately) result from a Principal Component Analysis (PCA).



b) If the given dissimilarities correspond to a valid metric, Multidimensional Scaling (MDS) can be solved as an eigenvector problem, in analogy to kernel PCA. What problem can arise if we attempt to use this approach even though the metric assumption is violated?
Answer: Before using this approach on nonmetric distances, we need to do some transformation to the data beforehand. First we have to find the optial monotonic transformation of the proximities. Then we have to arrange the points of configuration so that their distances match the scaled proximities as closely as possible.
c) The ISOMAP algorithm involves construction of a neighborhood graph. Briefly explain two different approaches for this step and describe a situation in which you would prefer one of them over the other.
Answer: The two different approaches to contruct neighbor Graphs are:
1) ϵ-graph: This approach connects each point to all the points that are at distance equal to or less than ϵ
2)KNN: This approach connects each points to to nearest K-neighboring points. In the situation of the presence of outliers, using knn would connect these points to their neighbors regardless of how far they are. In this situation, I would prefer to use an ϵ graph so they would not be connected to points that are far away. In my opinion, a combination of both approaches would be the optimal solution. In other words, connecting each point to its knn neighbors that lie within distance ϵ. In this case, outliers will not be connected to thier knn as they are further than distance ϵ.