CSOC20010

Introduction to Computational Social Science II
UCD School of Sociology
Spring, 2021-2022

Module Coordinator: Assoc Prof Taha Yasseri

Week 9 Assignment: Clustering and PCA

PART I:

In this assignment, we want to predict what would happen to you, if you were on Titanic! Would you have survived or ....?!

In the attached Python code we build a model that clusters the passengers into two groups based on their socio-economic features and we see that there is about X% overlap between group assignment and the actual faith of the passengers in each group, which shows the high predictive power of the socio-economic features.

For this assignment, you should think of an imaginary version of yourself with similar socioeconomic features, who has embarked on Titanic. Then you check to which cluster the model assigns you (imaginary you) and based on that what would be your best guess for the faith of the imaginary you. You can find the original Titanic data and the Python code in the attachment.

You need to submit a paragraph reporting the value of X above, the result of your prediction, and the modified ipynb code as an attachment. I should be able to run the code without error and see the prediction when I run the last line of the code.

HINT:

All you need to do is to add the following to the end of the code, replace the values in the first line with what would best describe the imaginary you, and run the whole code!

```
dfTaha = pd.DataFrame({'places': [3], 'name':['taha yasseri'], 'sex': ['male'], 'age':
[63], 'sibsp': [1], 'parch': [0], 'ticket':[10], 'fare':[0.95], 'cabin':['D7'],
'embarked':['S'], 'boat':[10], 'body': ['?'], 'home.dest':['Hudson']})

dfTaha.drop(['body','name'], 1, inplace=True)

dfTaha = handle_non_numerical_data(dfTaha)

predict_me = np.array(dfTaha.astype(float))

prediction = clf.predict(predict_me)

print(prediction)
```

PART II (Optional):

In the first part, we did the clustering in an 11-dimensional space (we considered 11 different parameters (sex, age, number of siblings, etc). However, some of these parameters are correlated. Therefore we can reduce the dimensionality of our data using Principal Component Analysis (PCA).

The Part II of the Python code runs the PCA for the variables and prints the Principal Components and their relationship with the original variables.

Focus on PCA1 and PCA2, what collective feature do they capture?

At the end of this section, all the data points are plotted in the transformed space (based on PCA1 and PCA2. The survival variable is used to set the colour of each data point (each passenger).

Write a paragraph with your interpretation of PCA1 and PCA2 and the final diagram. Which colour corresponds to which “outcome”?