Homework 12 - Principal Component Analysis (PCA)

Mall Customers data set (65 points)

Due Date: Monday April 26 at 11:59 pm

Instruction:

- This is a group-work assignment!
- You are expected to submit the **.ipynb** file and the exported **.html**.
- Only one member in each group needs to submit the assignment. It will be automatically submitted for the rest of group members.

  Good luck and enjoy machine learning!

Question 1 Linear PCA (35 points)

Import the Mall_Customers.csv data frame, define your feature space as X and do the followings:

1. Import the pca Python package and define your model with 3 principal components. Because the units of feature space are relatively close, you don't need to scale the data for this exercise. **(5 points)**
2. Fit the model and report the PC loadings and PC scores. **(5 points)**
3. Report the cumulative proportion variance explained (PVE) for each principal component **(5 points)**
4. Scree plot: Plot the scree plot and interpret what you see **(5 points)**
5. Biplot: Plot the biplot with two features only and interpret what you see **(10 points)**
6. From the biplot you visualized in part 4, how many customer segments do you recommend to the management team **(5 points)**

Question 2 Kernel PCA (30 points)

From sklearn.decomposition import kernelPCA and answer the following questions:

1. Fit your kernelPCA with 3 components using rbf kernel. **(5 points)**
2. Find the proportion variance explained (PVE) for each principal component **(15 points)**
3. Report the cumulative PVE and compare it with your findings in Question 1 part 3. **(5 points)**
4. If, for visualization purposes only, you had to work with the first two principal components, which method do you preferer. The linear PCA or Kernel PCA? Why? **(5points)**