CSCI 550: Adv. Data Mining
Fall 2023

Project 1: Exploratory Data Analysis, Dimensionality Reduction and Clustering

Due date: Sep 29, 2023

This project is designed for team collaboration, accommodating groups of 2 to 4 members; unfortunately, individual submissions is not allowed.

While the use of Python is encouraged, you are permitted to use either R or Python to do this project. Please note that your final submission must include a well-structured and cleanly formatted PDF document. Additionally, it is mandatory to submit a notebook or a markdown file that integrates inline codes. Both these documents — the PDF and the code file — will be assessed collectively, hence they should mirror consistency; any alterations in the Jupyter notebook should be reflected in the PDF document. All results presented *must* have corresponding code. **Any answers/results given without the corresponding code that generated the result will be considered absent.** All code reported in your final project document should work properly. Please do not include any extraneous code or code which produces error messages. (Code which produces warnings is acceptable, as long as you understand what the warnings mean.)

For this project, you will be choosing your own dataset, given the following constraints: Pick one of the datasets published in 2022 or 2023. You may use UCI repository, or Tidy Tuesday project or any other repository to find a proper dataset. Your dataset should contain **at least 5 numerical attributes and at least 200 instances**. Try to choose a dataset that you believe might contain clusters.

I encourage you to be concise. A paragraph should typically not be longer than 5 sentences.

Part 1: Think about the data [20]

In one or two well-written introduction, answer the following questions about the data:

1.1 Why are you interested in this data set?

1.2 How many numerical attributes and categorical attributes are there in the data set?

1.3 Are there any missing values? (you may use a plot or table to summarize this information about missing values)? If there are missing values, how are you planning to handle these (will all data instances with missing values be removed? Will all attributes with missing values be removed? Will missing values be imputed? If so, how?)

1.4 Of the attributes used to describe this data, which do you think are the most descriptive of the data and why (before doing any data analysis)?

1.5 Why do you expect clusters to be present in the data?

1.6 Why might finding clusters in this data set be helpful (how might this help us understand or analyze the data)?

1.7 How many clusters do you expect to see in the data? Provide a range of values to answer this question. For example, 2 to 4. Why do you expect a number of clusters in this range?

1.8 Do you expect that the clusters will be of similar size (i.e., cluster 1 is about the same size as cluster 2, is about the same size as cluster 3, etc..)? Why or why not?

Part 2: Perform some exploratory data analysis [20].

Use any library that you are interested to answer the following questions in a well-written paragraph, and create the following plots from the numerical portion of the data.

2.1 What is the multivariate mean of the numerical data matrix (where categorical data have been converted to numerical values)?

2.2 What is the covariance matrix of the numerical data matrix (where categorical data have been converted to numerical values)?

2.3 Choose 2 pairs of attributes that you think could be related. Create scatter plots of all 2 pairs and include these in your report, along with a description and analysis that summarizes why these pairs of attributes might be related, and how the scatter plots do or do not support this intuition.

2.4 Which range-normalized numerical attributes have the greatest sample covariance? What is their sample covariance? Create a scatter plot of these range-normalized attributes.

2.5 Which Z-score-normalized numerical attributes have the greatest correlation? What is their correlation? Create a scatter plot of these Z-score-normalized attributes.

2.6 How many pairs of features have correlation greater than or equal to 0.5?

2.7 How many pairs of features have negative sample covariance?

2.8 What is the total variance of the data?

2.9 What is the total variance of the data, restricted to the five features that have the

3 greatest sample variance?

Part 3: Write functions for clustering in Python [30].

Write the following functions in Python. You may use scikit-learn or other packages to check the correctness of your implementation, but you may not use any existing clustering algorithm implementation in your code.

3.1 A function that implements the k-means clustering algorithm. The function should take a data matrix, a number of clusters k, and a convergence parameter epsilon, as input, and return the representatives (means) as well as the clusters found using k-means. If the distance is the same between a point and more than one representative (mean), then assign the point to the mean corresponding to the cluster with the lowest index.

3.2 A function that implements the DBSCAN clustering algorithm. The function should take a data matrix and the parameters minpts and epsilon as input, and return the clusters found using DBSCAN, where each data point is labeled as either a noise point, a border point, or a core point.

Part 4: Analyze your data [30]

Report the following, using tables or figures as appropriate. You may use scikit-learn’s implementation of k-means and DBSCAN, but you are encouraged to first try using your own implementations on real-world data.

4.1 Use sklearn’s PCA implementation to linearly transform the data to two dimensions. Create a scatter plot of the data, with the x-axis corresponding to coordinates of the data along the first principal component, and the y-axis corresponding to coordinates of the data along the second principal component. Does it look like there are clusters in these two dimensions? If so, how many would you say there are?

4.2 Use sklearn’s PCA implementation to linearly transform the data, without specifying the number of components to use. Create a plot with r, the number of components (i.e., dimensionality), on the x-axis, and f(r), the fraction of total variance captured in the first r principal components, on the y-axis. Based on this plot, choose a number of principal components to reduce the dimensionality of the data. Report how many principal components will be used as well as the faction of total variance captured using this many components.

4.3 For both the original and the reduced-dimensionality data obtained using PCA in question 4.2, do the following: Experiment with a range of values for the number of clusters, k, that you pass as input to the k-means function, to find clusters in the chosen data set. Use at least 5 different values of k. For each value of k, report the value of the objective function for that choice of k.

4.4 For both the original and the reduced-dimensionality data obtained using PCA in question 4.2, do the following: Experiment with a range of values for the mints and epsilon input parameters to the DBSCAN function to find clusters in the chosen data set. First keep epsilon fixed and try out a range of different values for minpts. Then keep minpts fixed, and try a range of values for epsilon. Use at least 5 values of epsilon and at least 5 values of minpts. Report the number of clusters found for each (minpts, epsilon) pair tested.