CS 422: Data Mining

Department of Computer Science
Illinois Institute of Technology
Vijay K. Gurbani, Ph.D.

Spring 2022: Homework 9 (10 points)

Due date: Sunday May 01, 2022, 11:59:59 PM Chicago Time

You cannot use any late exceptions for this homework.
Please read all of the parts of the homework carefully before attempting any question. If you detect any ambiguities in the instructions, please let me know right away instead of waiting until after the homework has been graded.

1. Exercises (Please submit a PDF file containing answers to these questions. Any other file format will lead to a loss of 0.5 point. Non-PDF files that cannot be opened by the TAs will lead to a loss of all points.)

1.1 Tan, Chapter 7 (Cluster Analysis: Basic Concepts and Algorithms) (2 points)

Exercise 7, 12, 16. (For 16, note that Table 7.13 for Exercise 16 has a similarity matrix, not a distance matrix. Similarity and distance are related to each other by the formula *distance = 1.0 – similarity*.)

2 Practicum problems (7 points divided as specified by each assignment) Please label your answers clearly, see Homework 0 R notebook for an example (Homework 0 R notebook is available in "Blackboard → Assignment and Projects → Homework 0".) Each answer must be preceded by the R markdown as shown in the Homework 0 R notebook (### Part 2.1-A-ii, for example). Failure to clearly label the answers in the submitted R notebook will lead to a loss of 2 points per problem below.

2.1 DBScan (4 points, distributed as indicated below)

Read in the dataset s1.csv uploaded in Blackboard. (s1.csv is extracted from "s1.txt", **Clustering Basic Benchmark**, P. Fränti and S. Sieranoja, http://cs.joensuu.fi/sipu/datasets/). S1 is a set of Gaussian clusters. There are 5,000 observations of two dimensions in the dataset.

(a) **[0.2 pts]** Do you think it is necessary to standardize the dataset? Justify your answer.

(b) **[1 pt.]** Plot the dataset and describe in 1-2 sentences what you observe (visually) in the plot: how many clusters do you see? Are they well-separated?

(c) **[0.5 pt.]** Let's see how many clusters K-Means finds.
(i) Using the "wss" method, draw the scree plot for the optimal number of clusters.
(ii) Using the "silhouette" method, draw the scree plot for the optimal number of clusters.
(iii) What do you think is the appropriate number of clusters if we were to use K-Means clustering on this dataset?

(d) **[0.5 pt.]** (i) Using the answer to (c)(iii), perform K-Means clustering on the dataset and plot the results.
(ii) Comment on how K-Means has clustered the dataset. **(1-2 sentences.)**

(e) **[1.8 pts.]** We will now perform dbscan on this dataset.
(i) What is the value of MinPts that you think is reasonable for this dataset? Why?
(ii) In order to find the value of ε (eps), we need to calculate the average distance of every point to its *k* nearest neighbors. Set the value of *k* to be the result you obtained in (e)(i). Then, using this value determine what the correct value for ε should be. (Hint: Look at the online manual page for the function kNNdistplot()).

Using the scree plot from kNNdistplot(), you should find the best value of ε that clusters the dataset into the expected number of clusters determined in (c)(iii). To do this, perform a grid search on ε, and for each value of ε, run dbscan algorithm and visualize the clustering results. (You can do this manually in the R REPL and find the best value for ε, you do not need to write a loop.)

Using the best value of ε, plot the results of the dbscan algorithm on the dataset and state how many clusters you see in the plot in the form below:

```
At minPts = ___, eps = ____, there are ____ clusters.
```

2.2 Principal component analysis (4 points, distributed as indicated below.)

In this assignment, you will perform PCA on a small dataset and interpret its results.

The dataset is derived from the 2007 CIA World Factbook and contains 8 attributes. These are:

Variable	Meaning
GDP	GDP per capita in US dollars
HIV	HIV prevalence as a percentage of the population
Lifeexp	Life expectancy (in years)
Mil	Military spending as percentage of GDP
Oilcons	Oil consumption in barrels per annum per capita
Pop	Population (in millions)
Tel	Number of fixed telephone lines per 1,000 people
Unempl	Percentage of population unemployed

The dataset is available in Blackboard in the file countries.csv. Read in the dataset in a data frame taking care to allocate the first column to the name of the row instead of an attribute; this will allow the row name to be displayed in a PCA biplot. (Hint: See row.names parameter for the read.csv(…) method.)

(a) **[0.1 pts]** (i) Print a summary of all of the attributes in the dataset to become familiar with their values and ranges.
(ii) **[0.4 pts]** Plot a boxplot of all of the attributes. There are two outliers associated with the **Pop** boxplot. What do you think they represent?

(b) **[1 pt.]** Perform a PCA transformation on this dataset.
(i) Print the summary of the PCA object. How many components explain at least 90% of the variance?
(ii) Draw a screeplot of the PCA object.
(iii) Based on the screeplot, how many principal components would you use for modeling?

(c) Print the PCA components (the "rotation" field of the PCA object). Let's focus on PC1 and PC2.
(i) **[0.5 pts.]** Which attributes is PC1 positively correlated with, and which attributes is it negatively correlated with? Based on this, what is your expectation of PC1?

(ii) **[0.5 pts.]** Which attributes is PC2 positively correlated with, and which attributes is it negatively correlated with? Based on this, what is your expectation of PC2?

(d) **[1.5 pts.]** Draw a biplot with the first and second components. Then, answer the following questions:
(i) Examine the rotated variables (the "x" field of the PCA object) for the first and second component for Brazil, UK, and Japan. Print these two columns out.
(ii) Using the information in (d)(i)(ii), provide reasons whether the values for PC1 and PC2 for Brazil, UK, and Japan make sense.