Data Mining
IT 270
Midterm

Prof. Alexander Pelaez
apelaez@hofstra.edu

$17^{th} March, 2023$

1 Question 1 - Linear Algebra (15 Pts)

Please show all work. You may use R for part g, do not use `cov2cor()`. You may use R for part h & i

Given:

$$\mathbf{A} = \begin{bmatrix} 8 & 3 \\ 5 & 4 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 4 & 1 \\ 6 & 2 \\ 3 & 3 \end{bmatrix}, \mathbf{C} = \begin{bmatrix} 2 \\ 5 \\ 3 \end{bmatrix}, \mathbf{D} = \begin{bmatrix} 2 & 1 \\ -2 & 1 \end{bmatrix}, \mathbf{F} = \begin{bmatrix} 9 & 5 & 7 \\ 2 & 6 & 8 \\ 3 & 6 & 3 \end{bmatrix}, \mathbf{\Sigma} = \begin{bmatrix} 9 & 3 & 0 \\ 3 & 4 & 2 \\ 0 & 2 & 8 \end{bmatrix}$$

(a) $\mathbf{BA}$

(b) $\mathbf{B'F}$

(c) Find the determinant of $\mathbf{F}$. Show all work.

(d) $(\mathbf{AA})\mathbf{A'}$

(e) Find the trace of matrix $\mathbf{F}$.

(f) Assume Sigma $\mathbf{\Sigma}$ is the covariance matrix of the matrix $\mathbf{X}$; calculate $\mathbf{V}^{\frac{1}{2}}$.

(g) Calculate the correlation matrix (ρ) of Sigma $\mathbf{\Sigma}$ (do not use the `cov2cor()` function).

(h) Using `R` calculate the determinant of matrix Sigma $\mathbf{\Sigma}$.

(i) What is a singular matrix? What is a simple way to tell if a matrix is singular? Identify if (A, B, F, Sigma) are singular - provide your reasoning for each. $\mathbf{\Sigma}$.

2 Question 2 - PCA (25 Pts)

Using the Air Quality data set (See Data Files Link in Blackboard - column descriptions are in blackboard) of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multisensor Devices, examine the following:

(a) Provide a summary of this data, the definitions of the data are on blackboard to help you a bit. Your summary should include the types of data, interesting summary measures and appropriate box plots. Consider outliers and missing values.

(b) The analysis wishes to reduce the dimensions of the data set for further analysis. Conduct a principal component analysis and provide the following information :

- How many dimensions would you choose and why ?
- Are there any overlapping loadings based on a .6 cutoff? If so, which ones? Why would this be an issue
- Conduct a Varimax rotation on the result, how does this differ from a non rotated solution? Are there any overlapping loadings based on a .6 cutoff? Is this solution better?
- Using your new dimensions create new variables in your data set and compute values for each PCA dimension from the dimensions you choose. Please show the R code and the first observation. Correlation analysis between your new components, what does this correlation tell you?

3 Question 3 - Factor Analysis (25 Pts)

It is a well known fact that sports analytics are very popular. The data set `fifa.csv` contains information about soccer (football) players obtained from FIFA 19 information. It would be interesting if the ratings that are used could be analyzed as a simple set of factors, based on the skills of the player.

(a) Identify the columns that are the best candidates for analysis as factors.

(b) Conduct a factor analysis using these columns (if you get an error reduce the number of factors in the function, until the error disappears).

(c) Compare the results of a Varimax and Promax rotation, which gives you a better solution. (Note: Consider you may or may not need to change the number of factors).

(d) Name the factors, as best as you can - even if you don't know soccer, try and name them based on what you see in the factors.

(e) Are there any correlations between the factors? If so, explain what

(f) Create the diagram (by hand or Powerpoint) of the factors, be sure to label everything. Please do not use the function in R to create the diagram.

(g) Split the data set into two parts (Left footed players and right footed players). Conduct a factor analysis on each data set and explain if you see any differences between right and left footed players . Note you do not need to draw this one.

4 Question 4 - Decision Tree (25 Pts)

People Analytics is becoming a more popular and demanding area for Data Mining. The Head of Human Resources is looking to identify reasons for attrition (people leaving either voluntary or otherwise). The task is to develop a **decision tree** to determine this. You will need to merge three datasets `employee_survey_data.csv`, `general_data.csv`, `manager_survey_data.csv` to accomplish this task.

(a) Provide a summary of this data. Your summary should include the types of data (that are essential for this analysis), interesting summary measures and appropriate box plots. Consider outliers and missing values.

(b) Determine and list any columns that are not needed in a first “full” model. Explain why you removed those columns

(c) Using two algorithms C5.0 and Conditional Inference. Develop a full model decision tree to analyze the attrition.

(d) From your full model, develop a more appropriate model (lower variables) that will provide an ”attrition” prediction. What was the formula of your final model , in the C5.0 and conditional inference algorithm.

(e) Compare the methods based on appropriate measures of accuracy.

(f) Provide the charts from each algorithm of your final model.

5 Question 5 - Theoretical Questions (10 Pts)

Provide no more than a paragraph for each, be concise.

(a) Explain why the covariance matrix is important for analysis. Additionally, explain what the eigenvectors and eigenvalues are and why they are important.

(b) In machine learning, explain the importance of splitting up the dataset. What are the different ways to split and how should an analyst split the data.

(c) What are the differences and similarities between factor analysis and PCA. Focus on the equations that are produced.

(d) What are some of the challenges with k-NN. How would you advise someone who decides to use k-NN?

(e) What is the difference between orthogonal rotation and oblique rotations. When would we choose either, or neither.