DSC 324/424: Homework 4

Due Friday, August 19, 2022 by midnight

1) **(20 points)** Submit a rough draft of your final project (Executive Summary) and Journal Article including the following, which will help you prepare your final document for the following week. **Look to How to Read a Journal Article within the Syllabus folder to assist in writing your journal article.**
    - **a. Executive Summary (2 Pages)**
        - i. Non-technical report (imagine writing a report to your boss, who is a business manager without any technical/statistical expertise)
        - ii. A paragraph or two about the application and dataset including the research question
        - iii. If needed, any necessary background information needed to understand the application or the results (be brief or short in this section)
        - iv. Brief description of the methods
        - v. Results in laymen terms and how it relates to application
        - vi. Limitations of the research and future work
        - vii. Final Conclusions about the research
    - **b. Abstract (250 words)**
        - i. Introductory sentence explaining the research question, which intrigues the audience to continue reading
        - ii. A sentence about the methods used
        - iii. A sentence or two about the results
        - iv. A sentence about the conclusion and how the results tie into the research question and the application of the dataset.
    - **c. Introduction**
        - i. Explain the dataset and provide background on the application
    - **d. Literature Review**
        - i. Can be combined with the introduction
        - ii. Provide any previous literature on the topic, research questions, or methods used to answer the topic related to your research question
    - **e. Methods**
        - **i.** The list of the techniques that you are using to analyze your data. Include descriptions of why each topic is appropriate.
    - **f. Discussion and Results**
        - **i.** For each technique, a preliminary analysis and preliminary results remember to answer your research question.
        - **ii.** Plots or visualizations that back up or reinforce your analysis (in journals, they are labeled as figures)
        - **iii.** Include a section that describes the common threads that link your analyses. Is there anything that is reinforced by two approaches? Are there any common threads that you are seeing? If your analyses are so distinct that they don't cross-over like this, what do each add to your understanding of the data?
        - **iv.** Limitations of the research / how would you have completed the research differently
        - **v.** Future work that could be conducted on the dataset or the application
    - **g. Conclusion**

i. Final conclusions about the research

DONE

Due Sunday, August 21, 2022 by 11:59PM

2) (20 points): The Excel spreadsheet heart.csv contains one sheet named cardiovascular. These are data from a sample of 70,000 with 12 variables for each patient. These are:

1) Age-age of patient
2) Gender-1-Male and 2-Female
3) Height in cm
4) Weight in kg
5) ap_hi-arterial pressure
6) ap_lo- arterial pressure
7) Cholesterol
8) Glucose
9) Smoke-1-Smoking and 0-No Smoking
10) Alcohol-1 Drinking Alcohol and 0-No Drinking Alcohol
11) Active-1 Staying Active and 0-No Staying Active
12) Cardio-1 Has Cardiovascular Disease and 0 No Cardiovascular Disease

Develop a Linear Discriminant Analysis model to classify the cardio event from the other variables.

a) What is the performance of the classifier using cross-validation?

b) What is the performance of the classifier using training and testing?

c) Would certain misclassification errors be worse than others? If so, how would you suggest measuring this?

Yes, misclassification in smoke, alcohol, and active would be worse than others. The first way is to balance the samples in each class by comparing the points: (1) Accuracy of training, (2) validation, (3) test data.

3) **(10 points-Cluster Analysis): Using Google Scholar**, locate **a journal article**, which uses cluster analysis in your field of interest. **Write a summary** of the journal article and how it utilizes the cluster analysis in **two to three paragraphs**. **Cite the paper in APA format.**

In this article, researchers used two analyses: (1) Principle component analysis

and (2) Cluster Analysis. The use of each analysis, together with nine technical stock indicators, helped to estimate a stocks potential returns. The time frame of the returns is measured in the day of training by choosing the top thirty stocks based on earnings. The paper used PCA to transform the variables in order to eliminate influence as there are

> "different dimensions of indicators so teh influence of dimensions should be eliminated before calculations and the original data should be standardized" (Guo, 4).

However, in this summary we will be focusing on cluster analysis.

Guo used a clustering algorithm called, classical k-means clustering analysis algorithm. This algorithm divides the data set based on the centroid. This allows for data to only be apart of one cluster. After, Guo divided the data into the number of classes, they recalculated the centroids to calculate the mean of each dimension of all data items. Through the analysis, Guo found that:

> "the return of our constructed portfolio is significantly better than the market index. In most cases, the return rate of the portfolio is more than twice the return rate of the market index" (Guo, 5).

This is a phenomenal return on back tested data. However, based on previous summaries, this data could be overfitted to preform well on past data. A true test for this method of stock analysis would be testing the data live on real time market data. This would prove, by either losing or gaining money, that the analysis worked.

Guo, Y. (2020). Stock trading based on principal component analysis and clustering analysis. *IOP Conference Series: Materials Science and Engineering*, *740*(1), 012129. https://doi.org/10.1088/1757-899x/740/1/012129

Extra Credit (10 points)

An academic paper from a conference or Journal will be posted to the Homework 4 content section of D2L. Review the paper and evaluate their usage of FA and Latent Dirichlet Allocation (the other LDA). In particular address the following: **(See article on The analysis of customer satisfaction factors which influence chatbot acceptance in Indonesia)**

- What is the application of this paper?

  The application of this paper is to use natural language processing to determine people in indonesia are satisfied with chatbots and if they are accepted in Indonesia.

• What is the research question the authors wish to answer in this paper?

"This study aims to determine the extent of the customer satisfaction factors that successfully influence chatbot acceptance in Indonesia" (Sanny et al., 1225)

• What is Natural Language Processing (NLP) and what can we learn from it?

NLP is the use of AI to comprehend contents of written or typed documents allowing for the AI to extract information withing said documents. The AI can also categorize and orgaize documents allowing for a better understanding of documents.

• How does this paper utilize FA and LDA in Natural Language Processing?

This study uses an exploratory FA, utilizing KMO and Bartletts test of sphericity. However no mentions of LDA utilization in the article appeared.

• What are the results and conclusions from this paper?

Use factors such as usefulness, brand image, personality, and ease of use when building a chatbot in Indonesia.

• What other areas or fields do you think would benefit from LDA?

LDA can be used in facial recognition softwares. It can eliminate useless pixels of data allowing for the software to calculate facial recognition based on useful pixels.

• What other thoughts do you have on topic modeling, NLP, and LDA?

LDA and NLP could be used in my area of interest for algorithmic trading. NLP could be used to track recently published news articles and calculate the sentiment of the stock based on said article and then LDA can process the data resulting in a more significant measure of sentiment around a stock.