Project Description for Unsupervised Learning Class with PCA Analysis (Engineers and Business Audience)

Valle Varo

January 2024

1. Teamwork: This project will be conducted in pairs.
2. Presentation: As part of the project, you will be required to prepare a presentation with slides. This presentation will serve as a means to communicate your findings, insights, and the overall journey of your analysis.
3. Jupyter Notebook: In addition to the presentation, you will need to submit a Jupyter notebook containing all the code and technical details of your analysis. The notebook should be well-structured and include comments and explanations to ensure that your approach is clear and reproducible.
4. Data Sharing: One crucial aspect of this project is data sharing. You and your partner are responsible for providing the dataset you will be using for the analysis. Make sure to choose a dataset that is clean, well-documented, and suitable for unsupervised learning techniques. When submitting your project, ensure that you include the dataset, either in the form of a downloadable link or as an attachment.

Introduction

- Begin with a brief introduction to the project.
- Present the real-world problem or business challenge that you aim to address.
- Clearly articulate how unsupervised learning, specifically cluster analysis, and/or Principal Component Analysis (PCA) can provide valuable insights or solutions to this problem.
- State the overarching objective of the analysis and the specific problem you intend to tackle.

Data Overview

- Describe the data sources you will be using for the analysis.
- Provide details about the data, including its nature, format, and any relevant preprocessing steps.
- Highlight that the clustering analysis will focus on numerical variables but leave room to explore the behavior of dichotomous or nominal variables if it adds value.
- Specify that a minimum of 4 numerical variables should be employed, with a recommended maximum of 10 variables.
- Emphasize the importance of data integrity, including handling missing values and outliers.

Data Exploration and Feature Engineering (if applicable)

- Consider conducting an initial data exploration, including a statistical summary of key variables.
- Visualize the data with tools like histograms, box plots, scatter plots, or other relevant graphs.
- Discuss any insights gained from the initial exploration.
- Discuss any insights gained from the initial exploration.
- Mention the possibility of creating new variables through feature engineering, with examples relevant to the dataset.
- Explain how such transformations are carried out and why they are beneficial.
- Clearly define which variables will be used for the final clustering and whether any dichotomous/nominal variables will be considered.

Principal Component Analysis (PCA)

- Introduce PCA as a dimensionality reduction technique.
- Explain the rationale behind using PCA, such as reducing multicollinearity or simplifying the dataset.
- Describe the steps involved in conducting PCA, including standardization and eigenvalue decomposition.
- Highlight that PCA will be used to reduce the number of variables while preserving as much variance as possible.
- Specify the number of principal components to be retained based on explained variance or other relevant criteria.

Cluster Analysis

- Prior to conducting cluster analysis, perform a preliminary data analysis to better understand the dataset's characteristics.
- Visualize the data in ways that help identify patterns, similarities, or groupings among the data points.
- Describe the tools and techniques you will employ for cluster analysis, including hierarchical clustering and K-means.
- Justify the selection of a particular clustering approach based on the results of initial exploratory analysis or domain knowledge.

Analysis Results, Cluster Interpretation, and Business Insights

- Present the outcomes of the dimensionality reduction technique chosen (PCA, clustering, other).
- Describe the identified clusters and provide insights into their characteristics.
- Utilize statistical summaries, centroid profiles, heatmaps, or any other relevant visualization tools to aid in cluster interpretation.
- If appropriate, link the clusters to additional dichotomous/nominal variables that were not initially used for clustering.
- Assign meaningful labels or identifiers to each cluster and explain how they relate to the business problem at hand.
- Propose potential actions or strategies for addressing the business challenge based on the cluster insights.

Final Conclusion and Potential Extensions

- Summarize the key findings and insights derived from the analysis, including both clustering and PCA results.
- Reflect on how the project has addressed the initial problem or challenge.
- Mention any possible extensions to the project, such as integrating the results with predictive algorithms, if relevant data is available.

Bibliography and Appendices (if necessary)

Include a list of references and any supplementary materials or code used in the analysis.