DATA 319: Model-based and Data-based Methods for Data Analytics

Fall 2024

WASHINGTON STATE UNIVERSITY
Data Analytics

Final Project

This document provides details and guidelines for the team project, which represents 25% of your final grade. The purpose of this project is to give you an opportunity synthesize the material that we have covered in the course and give you an opportunity to perform multivariate analysis on a real dataset of your choice.

1 Project Overview

In teams of 3-5 students, you will select a dataset, determine relevant research questions, and address them using the multivariate techniques we have discussed in class. Specifically, you will be asked to select a topic of interest to your team and to formulate research questions that can be answered by the methods from our class. I am happy to consult with your team to help find relevant datasets (several potential sources are listed below) or to narrow down specific questions to study. In terms of analysis, you will need to apply at least two of the methods we have covered in class to your dataset(s), as well as doing some comparison between your data and the multivariate normal distribution. The two main outputs from the project will be a research report and a team presentation, also described in more detail below.

2 Project Components

As the main goal of the project is to perform multivariate analysis, you will need to select and apply two of the topics that we have covered in the course to apply to your dataset. These high-level topics include those on the list below. Note that for many of these topics we have discussed several algorithms or specific techniques, and you don't need to perform them all for your dataset, but you should justify your choice if you make one. For example, if you choose to study dimension reduction, you might choose to implement PCA, or MDS, or Factor Analysis, etc. depending on the properties of your dataset, in which case you should explain in your report the factors that led to your decision.

- Multivariate normal modeling
- Normality testing
- Hypothesis testing
- Dimension reduction
- Classification (or other supervised learning)
- Clustering (or other unsupervised learning)

It is natural that the choice of methods will inform the potential questions that you can answer with the data and that your topics of interest will help determine which methods are most useful. While this doesn't need to be an iterative process, the two sets of decisions do not need to be made completely independently and one of the evaluation criteria will be whether the chosen methods are sufficient to provide reasonable answers to the research questions. Thus, you should spend some time thinking about the project as a whole and not just as a set of unrelated computations.

In addition to the two chosen methods each project must also perform a couple of common analyses. First, a correlation matrix for all of the numerical columns in the dataset should be constructed and analyzed, with any interesting or surprising features discussed in the report. Second, at least three of the columns should be selected for comparison with a multivariate normal distribution. The selected columns should be tested for normality and you should construct a similarly-sized set of points from a multivariate normal distribution with parameters μ and Σ matched to those of the chosen columns. Your observations comparing the actual data to the synthetic points should also be included in the report.

3 Data Sources

Below are some potential sources for datasets, although you should also feel free to select your own from another source. Note that many of these resources provide additional information about the data beyond the specific values and you should incorporate a discussion of the relevant pieces of this information (how the data was gathered, what was the original purpose of the data, what cleaning steps were performed before the data was uploaded, etc.) into your report and presentation.

- https://archive.ics.uci.edu/ml/datasets.php
- https://www.usa.gov/statistics
- https://www.nass.usda.gov/
- https://cseweb.ucsd.edu/~jmcauley/datasets.html
- http://www.cs.toronto.edu/~roweis/data.html
- http://www.cs.cmu.edu/~epxing/Class/DS/projects.html
- https://datasetsearch.research.google.com/
- https://www.kaggle.com/datasets
- https://snap.stanford.edu/data/

4 Evaluation

You team will summarize and describe your work in a final report. The report and presentation will be evaluated on the criteria listed below with a particular focus on technical soundness and creativity:

- **Research Questions:** Are the questions interesting, clearly stated, and specific?
- **Data Selection:** Is the chosen dataset a reasonable option for addressing the questions?
- **Methodology:** Are (at least) two multivariate methods specified and applied correctly? Is the normality testing and synthetic data comparison performed correctly?
- **Analysis:** Are the methods that we used to analyze the data appropriate and carried out correctly? Is the analysis thorough and logically conducted?
- **Conclusions:** Do the final conclusions provide satisfactory answers to the stated questions? Are the conclusions supported by the analysis that was performed and presented?
- **Reproducibility:** Is the code used to analyze the data correct and easily interpretable?
- **Visualizations:** Are the visualizations used to represent the analysis effective and complete?
- **Ethical considerations:** Are the potential ethical consequences discussed and mitigated?
- **Presentation:** Is the final report well organized and neat? Is the team presentation effective and informative? Are the plots designed with care for the presentation formats, including color choices, appropriate labelling, and other aspects of good visualization design?

4.1 Team Contract

By August 30 your team will need to submit a completed team contract (the template will be provided) which outlines the roles and responsibilities of each team member, frequency and logistics of meetings etc. The team contract is worth 1 point.

4.2 Project Proposal

By November 1 your team will need to submit a 1-2 page project proposal that summarizes your planned analysis and expected timeline. The project proposal is worth 2 points and it should include:

- A brief overview of your project idea and motivation (at most two paragraphs).
- A brief description of the dataset you have selected.
- Full statements of the research questions you intend to answer.
- The choice of multivariate methods you intend to implement and a brief justification for their relevance to the research questions.
- Any initial observations you have made about the data.
- A statement of expected individual contributions for each team member. It is ok to separate different aspects of the project based on the skills and interests of individuals in the team but this should be explicitly stated and everyone should agree to the plan. Additionally, everyone should be involved in writing up the results and preparing the presentation.

4.3 Project Report

Your final report should thoroughly describe your experiences and analysis, as though you were reporting the results of this project to a manager or supervisor. The document should be submitted as a .pdf file on Canvas **by the day of the project presentation** (*TBD later*). The following information should be contained in the report, as well as appropriate figures from the analysis incorporated into the markdown file:

- Describe the dataset and why you selected it for this project.
- Describe any processing problems you identified with the data and how you overcame those issues.
- Describe your research questions and why the data is a good choice to answer them.
- Describe the results of your exploratory analysis and what preliminary conclusions you were able to draw based on this analysis.
- Describe how you selected the methodology for your analysis of the big question and the pros and cons of that method and any alternative methods you considered.
- Describe your final conclusions based on your analysis and support them with analytics on your dataset.
- Describe any additional analyses that you would have liked to carry out and any additional data that would have been needed in order to extend your analysis.

4.4 Presentation

On the day of the project presentation (*TBD later*) your group will have 20-30 minutes (depends on the number of groups) to present the results of your project. Each member of the group should be speaking for at least two minutes. You will not be able to discuss every analysis and computation that you performed over the month that you were working on the project. Instead, you will need to prioritize the important pieces of your analysis decide how to efficiently present the data, question, and conclusions. Your presentation will be followed by a Q&A session.

The project report and the presentation are graded together at the maximum of 12 points.

4.5 Peer Evaluation

By December 13 you should provide the feedback on the performance and effort of your teammates using the provided form. The peer evaluation is worth 10 points, out of which 2 points are given for the merely submission of the completed form, and the remaining 8 points will be determined based on the feedback of your teammates. Note that this is an individual assignment.