A place to learn about statistics and data science

Wayne Tai Lee's Github Page

Project 3: Final Project

Overview

The final project in this class is an open-ended **data mining** project. Key deliverables are

- A **written report**, in a PDF format. There is no page limit but you must address all of the requirements below and produce a cohesive report.
- A critique written about **another group's** final project, your final project can be critiqued by multiple teams but there is no requirement that your project must be summarized.
- Your code on Github.
- You **can** have a partner for the final project.

Written report

- The project must perform a form of data mining:
    - Generate a non-obvious insights from the data that can be leads for further investigations (similar to the observation step in the scientific method). The most common way to do this is by merging datasets from different sources.
    - A metric or algorithm that quickly filters or sorts large amounts of data for people.
    - You should avoid projects where the end goal is a predictive tasks (too easy) or inferential statistics (too hard).
    - Please talk to me if you're not sure that your project would be considered data mining or not.
- You must identify an audience and articulate the potential value for that audience in the report.
    - Mining implies value is discovered and so you must articulate what value is generated for what audience in your project. Your audience can be an individual or an institution.
        - You should have some level of detail like "recent college grad" or "think tank for public policy on housing" rather than a generic person/institution.
    - Value can be realized in the form of saved time, decreased uncertainty, or increased reward. I'm open to hearing other forms of value definition.
- You should use at least 2 datasets or a single large dataset for your project.
    - If you have two datasets you should articulate how the datasets complement each other, for example
        - One dataset could provide complementing features (e.g. one dataset may have the activities and another may have the user demographic data).
        - One dataset could provide more resolution (e.g. detailed surveys tend to be done less frequently so they're often complemented with surveys that are less detailed but collected more frequently).
        - One dataset could provide more data from a different population (e.g. a system change has made historical patent data to be stored in a different system).
        - You should not call a dataset collected from a different time point but from the same source as a different dataset.
    - For single datasets that are truly large or complex, you should articulate why it is considered large/complex relative to the standards your audience is used to.
    - You must perform some sort of exploratory data analysis that checks the completion and/or quality of the data.
    - You must identify the source of the data and summarize the datasets.

        - Ideally, the data source should come from the entity that manages the data. Management is defined by the entity responsible if the data has quality issues. For example, data aggregators like Kaggle do not manage the data it posts on its challenges but Twitter manages Twitter data. Some exceptions exist like platforms like NYC OpenData allow you to report data quality issues and will be considered a manager of the data.
- You must use an algorithm to explore the relationship between different features in the data similar to Project 1.
    - Please have at least one graphical summary that highlights how your algorithm is working (e.g. the correlation plot from HW1, the loadings plot from PCA, etc).
    - You must quantitatively evaluate how your algorithm fits to the data. For example, correlations are meaningful if they're close to -1 or 1, clustering and supervised learning methods all have different metrics associated with them.
    - Please engineer at least one feature and evaluate its usefulness in the context of your project.
- Please verify the results from your mining whether is due to chance or likely a real pattern.
    - For example, you may have to subset the data to see if the result is due to an outlier. You can also quote external sources to help understand the data.
    - Please make sure you quantitatively address the question "if you had a different dataset, how robust are the results from the algorithm?"
        - Please be sure to articulate how high uncertainty may affect your conclusions (even if your results are robust).
    - Please note that the algorithm above should be a quick filtering where the second step is a close examination of the result.
- Introduction and Conclusion (5 pts)
    - Please talk about any "iterations" that you had to perform from the start of the project to the end. Please have at least one paragraph that discusses the potential for data-snooping for your project.
    - Your introduction and conclusion should be written at the level so a peer outside of this class can understand your intent and findings.

Critique requirements

To complete this portion, I recommend you talk to each other **before** the full project is finished.

- What was the initial motivation for tackling the project?
- What datasets were used?
- What aspect of the project is considered a **data-mining** and what is discovered?
- Is there anything you would have done differently? For example
    - Used different datasets
    - Used different models to explore the data
    - Generated a different feature
    - Arrived at a different conclusion from the given data. (These are simply suggestions, you team only has to come up with one thing)

Github Code

- You should have a README page
    - A at most 5 sentence summary of the project
    - An general explanation of the different files/folders if someone were to replicate your study.
- No data should be on Github but your README should explain how the data can be obtained.
- Your written report and code should both exist on Github.
    - You do not have to check in the PDF version of your report on Github but your Rmarkdown or LaTeX file should exist on Github.
    - If you use Word as a text editor. Make sure you have a link from your README to this document.