CMPINF 2100

Final project proposal

Instructions

Previous assignments and the final project proposal

- You have been required to look through the provided final projects.
- You have investigated their data sets by applying some of the skills we have discussed in lecture.
- The final project proposal formalizes those discussions and actions into a single report and asks you to try CLUSTERING to reveal potentially hidden patterns within the data.

The final project proposal and the remainder of the semester

- The second half of the semester is dedicated to understanding predictive modeling.

- **In the proposal you must think through what quantity you wish to predict and state if you are working with a regression or classification problem.**

- You do **NOT** have to state the models you will use in your project.
  - We will discuss multiple model types in the second half of the semester.

Major areas of the proposal

- Part A) Motivation
- Part B) Predictive modeling discussion
- Part C) Exploratory Data Analysis
- Part D) Cluster analysis

Part A) Motivation

- Why did you choose this final project?
  - What interested you about it?
  - Was it the application in particular?
  - Was it the aspects of the data – e.g. regression/classification, data size, etc…
- How interested are you in this project?
  - Could you not decide between two of the listed projects?
- **Answer this question after going through Parts B), C), D):**
  - After exploring the data and working through how you would model the data in more detail, would you prefer to work on a different project?
  - Do you find this application not interesting? Is it confusing? Is the application just not something you want to work on?

Part B) Predictive modeling discussion

- Do not state the specific models you will use, but you **MUST** state:
  - If you are working on a REGRESSION or CLASSIFICATION problem.
  - Which variables are inputs?
  - Which variables are responses/outputs/outcomes/targets?
    - Did you need to DERIVE the responses of interest by SUMMARIZING the available data?
    - If so, what summary actions did you perform?
  - Which variables are identifiers and should **NOT** be used in the models?

- **After completing Parts C) and D):** Discuss which of the inputs you think influence the response based on your exploratory visualizations.
  - Describe which exploratory visualization helped you identify potential input-to-output relationships.
  - If you are not sure which inputs seem to influence the response, it is ok to say so.

Part B) Predictive modeling discussion - HELP

- Module 1's Course Goals item included the "Data Science Supervised Learning Overview" which discussed the differences between REGRESSION and CLASSIFICATION.

Part C) Exploratory Data Analysis – general requirements

- You must read in the data associated with your project.
- You must perform the **ESSENTIAL** exploration activities:
  - Display the number of rows and columns
  - Display the columns names and their associated data types
  - Display the number of missing values for each column
  - Display the number of unique values for each column
- You must state if you want to effectively treat a numeric column as a non-numeric for exploration purposes.
- You do NOT need to display the COUNTS for categorical variables because you will visually the variables. However, you may display/print COUNTS if it helps you.

Part C) Exploratory Data Analysis – general requirements

- You must visualize the **MARGINAL** distributions for **ALL** variables in your data.
  - Continuous variables: Histograms or Density plots
  - Categorical variables: Bar charts

- You must decide appropriate visualizations to show:
  - Categorical-to-categorical relationships (COMBINATIONS)
  - Categorical-to-continuous relationships
  - Continuous-to-continuous relationships

- You must decide appropriate visualizations to show if RELATIONSHIPS change across GROUPS.
  - This is especially important to continuous-to-continuous relationships.
  - Do not forget though that categorical-to-continuous relationships can also be GROUPED BY a secondary categorical variable!

Part C) Exploratory Data Analysis – specific requirements

- If you are working on a REGRESSION problem, you must:
  - Visualize the relationship between the continuous response and the continuous inputs using:
    - Scatter plots and trend plots
    - Group the relationships by categorical variables (if appropriate)
  - Summarize the response for each unique value of the categorical inputs using:
    - Boxplots, Violin plots, and Point plots

- If you are working on a CLASSIFICATION problem, you must:
  - Visualize conditional distributions of the continuous inputs GROUPED BY the response (outcome) unique values.
  - Visualize relationships between continuous inputs GROUPED BY the response (outcome) unique values.
  - Visualize the counts of combinations between the response (outcome) and categorical inputs.

Part D) Cluster analysis

- This is an initial attempt at cluster analysis to help you explore the data.

- You will also perform cluster analysis for the final project submission.
  - The approach you use for the final project might be different from what you try now, that is ok!
  - This is just an initial attempt. See what you learn! If it does not seem to reveal anything “useful”…that is ok! That’s part of the learning process!

Part D) Cluster analysis

- You will **NOT** apply cluster analysis using ALL variables.

- Cluster analysis is **UNSUPERVISED**. It does NOT make distinctions between INPUTS and OUTPUTS.

- However, the projects are **PREDICTIVE ANALYTICS** problems with INPUTS and OUTPUTS. Therefore, you will need to decide which class of variables you will focus on in your Cluster analysis.

Part D) Cluster analysis – select variables

- You **MUST** state which class of variables you used in your cluster analysis. For example, you could:

- Cluster using CONTINUOUS inputs.
- Cluster using CONTINUOUS responses (if you have multiple outputs).
- Cluster using multiple summary statistics derived from inputs.
  - For example, if your data includes MANY categorical variables consider GROUPING BY them and SUMMARIZING the OTHER inputs.
    - Summary statistics to consider: number of unique values, mean, standard deviation, median, etc…
    - If you do this, do NOT start by creating COMBINATIONS of ALL categorical inputs. Select a few which provide many COMBINATIONS.
- Cluster using multiple summary statistics derived from the outputs.
  - For example, if your data includes MANY categorical variables consider GROUPING BY them and SUMMARIZING the outputs.
    - Summary statistics to consider: number of unique values, mean, standard deviation, median, etc…
    - If you do this, do NOT start by creating COMBINATIONS of ALL categorical inputs. Select a few which provide many COMBINATIONS.

Part D) Cluster analysis - considerations

- How many variables are you using relative to the number of observations?
- Are the variables you are using “Gaussian-like” or are the distributions “odd looking”?
  - You must visualize the MARGINAL histograms of the variables IF you are using summary statistics derived from INPUTS or OUTPUTS. If you are using the INPUTS or OUTPUTS directly, you already visualized the MARGINAL distributions.
- Are the variables you are using highly correlated to each other?
  - You must visualize the relationships between the variables you are using. If you are using the INPUTS/OUTPUTS directly then you already visualized those relationships.
- How should you handle missing values?
  - If you drop all rows with at least one missing value, how many observations would you still have (thus, how many complete cases do you have)?

Part D) Cluster analysis - requirements

- You must use KMeans to execute the Cluster analysis.
- You must use PCA to support the cluster analysis visualizations.

- First, use 2 clusters and:
  - Count the number of observations per identified cluster. Are the clusters balanced?
  - Visualize the cluster results between 2 of the variables used for the cluster analysis.
  - Visualize the cluster results using PCA. Is the visualization easier to interpret that the previous one?

- Then, identify the optimal number of clusters, and rerun KMeans for the optimal number.
  - Count the number of observations per identified cluster.
  - Visualize the cluster results between 2 of the variables used for the cluster analysis.
  - Visualize the cluster results using PCA. Is the visualization easier to interpret that the previous one?

Part D) Cluster analysis - interpretation

- **Do the OPTIMAL number of clusters ALIGN with CATEGORIES of KNOWN grouping variables in your data?**
  - If you are working on a classification problem, are the identified clusters consistent with the outcome categories?
  - If you are working on a regression problem, are the identified clusters consistent with categories of a categorical input?

- **What are the CONDITIONAL distributions of the variables you used for clustering GIVEN the identified clusters?**
  - Use appropriate visualizations to study the CONDITIONAL distributions of the variables used for clustering GROUPED BY the identified clusters.
  - ***HINT:*** you are creating CONDITIONAL distributions using the DERIVED cluster groupings rather than using a "real" categorical variable.
  - Describe the clusters based on the CONDITIONAL distributions.

Proposal submission

- You should include all text, discussion, code, tables, and graphics in a Jupyter notebook.

- You must submit both the source .ipynb file and the rendered HTML report.

- Submit the proposal via Canvas just like the other homework assignments.