CIS4930: Introduction to Multimodal Machine Learning in Python
Individual Coding Assignment 01
Due Feb 07, 2023, 11:59 pm

1. **Python Fundamentals (20 points).** For this question, use the data in "articleInfo.csv" and "authorInfo.csv."
   Bibliometric analysis uses statistical methods to analyze books, articles, and other publications. It enables us to unpack the evolutionary nuances of a specific field while shedding light on the emerging areas in that field.

   Imagine you are interested in the current state of a research topic, "How have VR (virtual reality)/AR (augmented reality) techniques been used for education?" and you have collected the basic information from journal articles in this field in the past 5 years. "articleInfo.csv" includes the following information about each article: "*Title*, *Year*, *Author Number*, *Citation*, *Source*, *Abstract*, and *Type*" "authorInfo.csv" includes the following information about each author: "*Author Name*, *Author Affiliation*, *Country*, and *h-index*."

   Merge "articleInfo.csv" and "authorInfo.csv" into one data frame based on "Article No.", fill all empty cells with the value of 0, and answer the following questions.

   1. Plot the *yearly_publication* figure, in which the x-axis is the *year*, the y-axis is the number of articles published during that *year*.

   2. Plot the *yearly_citation* figure, in which the x-axis is the *year*, the y-axis is the total number of citations during that *year*.

   3. Plot the figure of the number of publications across countries. You may use any available python libraries, such as *pygal_maps_world*, *geopandas*, or others.

   4. What are the top 5 institutions that have the most published articles in this area?

   5. Who are the top 5 researchers that have the most h-index in this area?

2. **Regression (40 points).** For this question, use the data in "data.csv" file.

   Task-oriented dialogue systems (e.g., Siri, Alexa, Google Now/Home, Cortana, etc.) are computer programs that use conversation with users to help complete tasks, give directions, control appliances, find restaurants, or make calls.

   Imagine you are conducting a study to evaluate the system usability of Siri for purchasing flight tickets, and 60 people have participated in this study. Also, assume that you also collect the following information from the users during your study:

   - ***Purchase:*** Whether the customer purchased a ticket or not by using Siri. (1: Yes, 0: No)
   - ***SUS:*** System Usability Survey scores, which the users filled out after interacting with Siri. (See this link for more info)
   - ***ASR_Error:*** Number of times Siri fails to recognize the user's speech.
   - ***Intent_Error:*** Number of times the system failed to classify the user's intention/speech act.
   - ***Duration:*** Total duration (seconds) of the dialogue between Siri and the user.
   - ***Gender:*** Gender of the user (0: Female, 1: Male)

   Train a **Regression** model in which: the **independent** variables (inputs) are "***ASR_Error***," "***IntentError***," "***Duration***," "***Gender***," and "***Purchase***"; the **dependent** variable (output) is "***SUS***., and anwer the following questions:

   1. Show the statistical results of your trained regression model.

   2. What features are significant? What features are insignificant?

   3. Were the results what you expected? Explain why or why not, for each feature.

   4. What does the model suggest is the most influential factor on SUS? Explain what tells you this is the most influential factor statistically.

   5. What are the potential reasons for these factor(s) being significant predictors of SUS?

3. **Classification (40 points).** For this question, use the same dataset in the above question. Train a **Classification** model in which: the **independent** variables (inputs) are "***ASR_Error***," "***IntentError***," "***Duration***," and "***Gender***,"; the **dependent** variable (output) is "***Purchase***." Use the evaluation metrics we introduced in class to compare the performance of the following four machine learning classification algorithms: (1) Logistic Regression, (2) SVM, (3) Naive Bayes, and (4) Random Forest.