Assignment Instructions: Assignment 4

Purpose

The purpose of this assignment is to use k-Means for clustering.

Directions

The dataset on American College and University Rankings contains information on 1302 American colleges and universities offering an undergraduate program. For each university, there are 17 measurements, including continuous measurements (such as tuition and graduation rate) and categorical measurements (such as location by state and whether it is a private or public school).

Note that many records are missing some measurements.

- Remove all records with missing measurements from the dataset.
- For all the continuous measurements, run K-Means clustering. Make sure to normalize the measurements. How many clusters seem reasonable for describing these data? What was your optimal K?
- Compare the summary statistics for each cluster and describe each cluster in this context (e.g., "Universities with high tuition, low acceptance rate…").
- Use the categorical measurements that were not used in the analysis (State and Private/Public) to characterize the different clusters. Is there any relationship between the clusters and the categorical information?
- What other external information can explain the contents of some or all of these clusters?
- Consider Tufts University, which is missing some information. Compute the Euclidean distance of this record from each of the clusters that you found above (using only the measurements that you have). Which cluster is it closest to? Impute the missing values for Tufts by taking the average of the cluster on those measurements.

File Attached: Universities.csv

Learning Outcomes

The assignment will help you with the following course outcomes:

1. Think critically about how to use machine learning algorithms to solve a given business problem.
2. Know how to formulate business problems and identify relevant data to use in modeling frameworks.
3. Know how to evaluate the appropriateness and estimate the performance of using k-Means for a given task.
4. Know how to use software tools (such as R) effectively to implement k-Means.
5. Foster the communication and presentation of statistical results and inferences.

Requirements

All due dates are included in the Assignment Schedule.

General Submission Instructions

All work must be your own. Copying other people's work or from the Internet is a form of plagiarism and will be prosecuted as such.

- If you are using R, upload a pdf/html document that you "knitted" from the RMD file to your git repository. If using Jupyter notebooks, the notebook should contain the code, output, and documentation. Name your file Username_#.ext, where Username is your Kent State User ID (the part before @), and # is the Assignment number. In this case, 4.

Provide the link to your git repository in Blackboard Learn for the assignment.