Topic 7: Unsupervised Learning

Jonathan Ibifubara Pollyn

GitHub Repository: https://github.com/JonathanPollyn/Machine-Learning-for-Data-Science

College of Science, Engineering and Technology, Grand Canyon University

DSC-540: Machine Learning for Data Science

Dr. Aiman Darwiche

November 15, 2021

Introduction

A cluster is a collection of data items that have been grouped due to commonalities. You'll set a target number, k, for the number of centroids you'll require in your dataset. A centroid is an imaginary or real point representing the cluster's center. The in-cluster sum of squares is reduced to assign each data point to one of the clusters. In other words, the K-means algorithm finds k centroids and then assigns each data point to the cluster with the fewest centroids possible. The means in K-means refers to data averaging or determining the centroid. The K-means technique in data mining starts with the first group of randomly picked centroids, used as the beginning points for each cluster. It then performs iterative (repetitive) calculations to optimize the centroids' placements.

In this report, k-means was used to answer some question about the data is pulled from Kaggle from the education, universities, and college tag. The data frame has 777 observations and 18 variables. I choose the dataset since it is suitable for practice cluster analysis, data visualization, management, and predictions. Question that the k-means will be answering are which university have the highest graduation rate? Are private universities preferred to public university? How is the private university when compared to the public university?

Application

The data is split into two university groups private and public university, figure 1 shows that description of the data. We can see that the graduation rate has a mean of 65.5 with a standard deviation of 17.2. The minimum graduation rate is 10 and a maximum of 118. There is 25 percent graduation rate of 53. The full-time undergraduates and part-time undergraduates are very far apart. The full-time undergraduates are recording a mean of 3699.9 while the part-time undergraduate is recording 855.3.

	apps	accept	enroll	top10perc	top25perc	f_undergrad	p_undergrad	grad_rate
count	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.00000
mean	3001.638353	2018.804376	779.972973	27.558559	55.796654	3699.907336	855.298584	65.46332
std	3870.201484	2451.113971	929.176190	17.640364	19.804778	4850.420531	1522.431887	17.17771
min	81.000000	72.000000	35.000000	1.000000	9.000000	139.000000	1.000000	10.00000
25%	776.000000	604.000000	242.000000	15.000000	41.000000	992.000000	95.000000	53.00000
50%	1558.000000	1110.000000	434.000000	23.000000	54.000000	1707.000000	353.000000	65.00000
75%	3624.000000	2424.000000	902.000000	35.000000	69.000000	4005.000000	967.000000	78.00000
max	48094.000000	26330.000000	6392.000000	96.000000	100.000000	31643.000000	21836.000000	118.00000

Figure 1: Data description

Figure 2 shows the graduation rate among the university, and we can see that the private university tends to attract higher graduation rates when compared to the public universities. This result clearly indicates that private universities are preferred to public universities since they have a higher graduation rate.

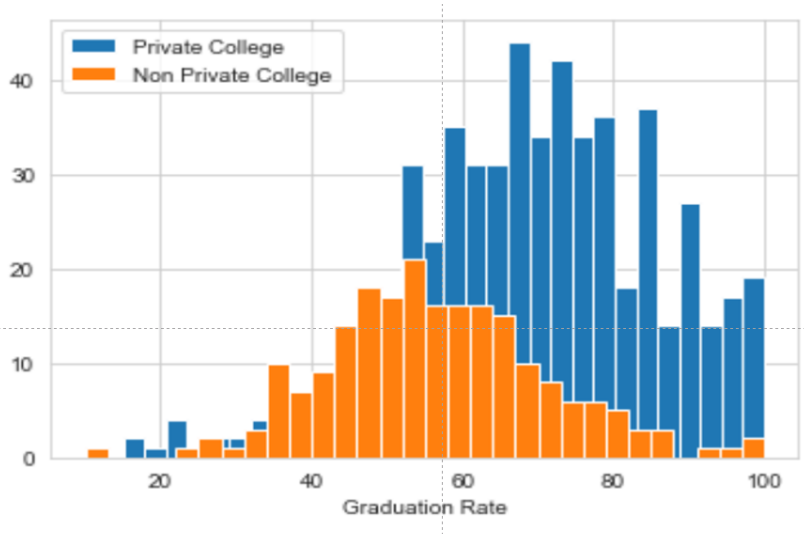

Figure 2: University Graduation Rate

We can see in figure 3 that there are more clusters for the private university (cluster 1) than the public universities (cluster 2). We can also see that the full-time undergraduate has clusters that are more spread out while the part-time undergraduate is close together. This shows that the centroid of the two clusters is different with the part-time undergraduate having a closer centroid compared to the full time. See figure 4.

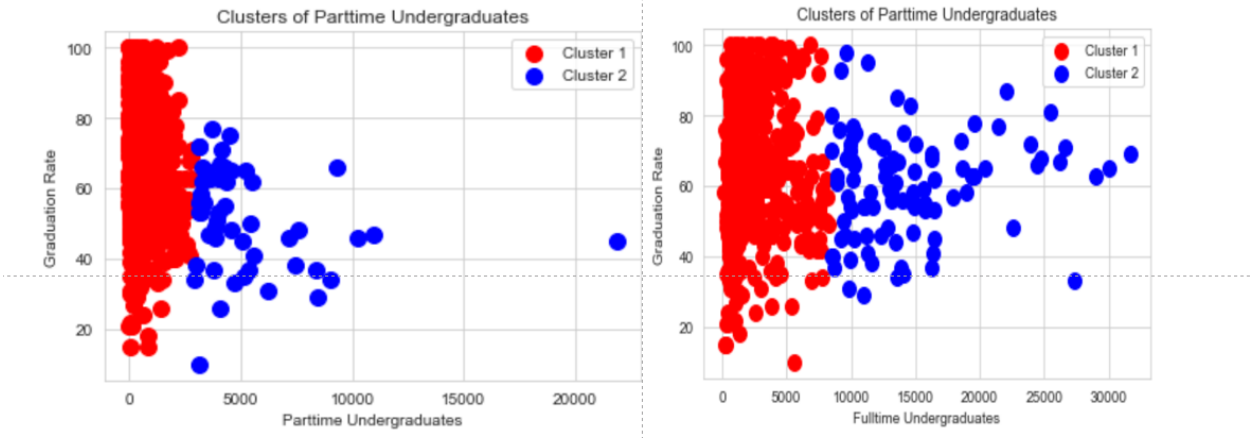

Figure 3: Cluster of Part-time and Full-time Undergraduates

```
array([[ 548.70661157,   66.44352617],
       [5219.7254902 ,   51.15686275]])
```

```
array([[ 2129.10014728,    66.20765832],
       [14583.35714286,    60.12244898]])
```

Figure 4: Centroid of Part-time and Full-time

The ethics of modern research include data protection and privacy. When organizing one's project activities, several difficulties must be addressed and resolved first. Establishing ethical and legal issues correctly allows handling data efficiently and effectively in the future. Adhering to ethical and legal terms is a basic assumption for a modern researcher's work. What questions should be asked and answered when planning a project, and what shared consensus should be reached? Is it necessary for responders to give their informed consent? What kind of consent will

be required, and in what form? Must obtain data subjects' agreement to store and share data. How should the identities of participants be protected if necessary? It is essential to understand the data used and ensure all the required steps to protect sensitive information.

References List

Ethical aspects of data management. (n.d.). Open Science Support Centre. https://openscience.cuni.cz/OSCIEN-84.html

Garbade, M. J. (2018, September 12). *Understanding k-means clustering in machine learning | by dr. michael j. garbade | towards data science*. Medium. https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1