# Analysis of Clustering Study on Universities Data

## Removing all records with missing measurements from the dataset

I have done this on MATLAB by using “ismissing” function. The function returns the value 1 if the given cell has a missing value.

## Normalizing the Measurements, Finding the Most Fitting Clustering Method and Applying It.

After normalizing the the measurements, I run the algorithm for different k values with 10 replicates.

## Finding the Best k Value for k-Means Algorithm.

I run k-means algorithm for k values from 20 to 50 and for each of them, I kept an average silhouette value. Silhouette is a measure to test the performance of the clustering algorithms. Higher the Silhouette value, better the result. The best k value in my algorithm is between 2-4 in the runs I took.

## Comparison of the Clusters and Statistics of Clusters.

Best k value is 3 in the run I decided to investigate. Here is the information on the values of different attributes for each cluster in this run.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Cluster Number** | **Number of Points** |  | Math SAT | Verbal SAT | ACT | # appli. rec'd | # appl. accepted | # new stud. enrolled | % new stud. from top 10% | % new stud. from top 25% | # FT undergrad | # PT undergrad |
| **1** | 114 | Max | 0,55 | 0,61 | 0,75 | 0,17 | 0,22 | 0,18 | 0,52 | 0,86 | 0,21 | 0,19 |
| Mean | 0,39 | 0,44 | 0,55 | 0,03 | 0,04 | 0,05 | 0,20 | 0,43 | 0,05 | 0,02 |
| Min | 0,02 | 0,03 | 0,25 | 0,00 | 0,00 | 0,00 | 0,00 | 0,13 | 0,01 | 0,00 |
| **2** | 65 | Max | 0,98 | 0,93 | 1,00 | 0,42 | 0,49 | 0,51 | 0,98 | 1,00 | 0,47 | 0,14 |
| Mean | 0,64 | 0,67 | 0,75 | 0,07 | 0,07 | 0,09 | 0,48 | 0,75 | 0,08 | 0,01 |
| Min | 0,47 | 0,48 | 0,60 | 0,01 | 0,01 | 0,02 | 0,20 | 0,46 | 0,02 | 0,00 |
| **3** | 27 | Max | 0,69 | 0,66 | 0,80 | 0,38 | 0,57 | 0,86 | 0,61 | 0,94 | 1,00 | 1,00 |
| Mean | 0,50 | 0,49 | 0,60 | 0,18 | 0,24 | 0,39 | 0,27 | 0,58 | 0,47 | 0,17 |
| Min | 0,34 | 0,40 | 0,35 | 0,06 | 0,08 | 0,19 | 0,13 | 0,38 | 0,21 | 0,03 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| in-state tuition | out-of-state tuition | room | board | add. fees | estim. book costs | estim. personal $ | % fac. w/PHD | stud./fac. ratio | Graduation rate |
| 0,59 | 0,58 | 0,48 | 0,52 | 0,32 | 1,00 | 0,60 | 0,90 | 0,30 | 1,00 |
| 0,32 | 0,33 | 0,23 | 0,27 | 0,07 | 0,21 | 0,19 | 0,58 | 0,14 | 0,47 |
| 0,01 | 0,08 | 0,07 | 0,06 | 0,00 | 0,07 | 0,03 | 0,00 | 0,05 | 0,06 |
| 0,78 | 0,77 | 0,53 | 0,58 | 0,42 | 0,58 | 0,31 | 0,95 | 0,19 | 0,84 |
| 0,57 | 0,57 | 0,30 | 0,33 | 0,08 | 0,21 | 0,14 | 0,79 | 0,11 | 0,65 |
| 0,09 | 0,22 | 0,09 | 0,13 | 0,01 | 0,09 | 0,05 | 0,54 | 0,03 | 0,31 |
| 0,51 | 0,50 | 0,51 | 0,51 | 0,24 | 0,34 | 0,47 | 0,92 | 0,23 | 0,66 |
| 0,08 | 0,24 | 0,21 | 0,24 | 0,10 | 0,22 | 0,28 | 0,78 | 0,17 | 0,43 |
| 0,01 | 0,12 | 0,08 | 0,10 | 0,00 | 0,00 | 0,08 | 0,63 | 0,09 | 0,25 |

By looking at these statistics, we can say that the schools in Cluster 1 have low “# applications rejected”, “# of applications accepted”, “# of new students enrolled”, “# of FT Undergrad”, “# of PT Undergrad”, “add. fees”, “student/facilities ratio”; high “% fac. w/PHD” and average for the rest of the attributes.

For schools in Cluster 2, we can say that they have low “# applications rejected”, “# of applications accepted”, “# of new students enrolled”, “# of FT Undergrad”, “# of PT Undergrad”, “add. fees”, “estimated book costs”, “estimated personal spending”, “student/facilities ratio”; high “Math SAT”, “Verbal SAT”, “ACT”, “% new stud. from top 25%”,” % fac. w/ PHD” and average for the rest of the attributes.

For schools in Cluster 3, we can say that they have low “in-state tuition”, “add. fees”, “estimated book costs”, “student/facilities ratio”; high “ACT”, “% new stud. from top 25%”,” % fac. w/ PHD” and average for the rest of the attributes.

## Including the Categorical Variables that are not Included in the First Analysis

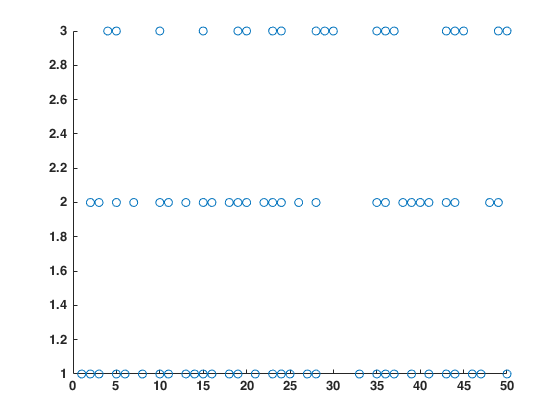
Percentages of private schools in each cluster is as the following for my run:

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster** | 1 | 2 | 3 |
| **Percentage** | 0,85088 | 0,86842 | 0,65789 |

There are 832 private schools and 470 public schools. So, it is understandable that private schools are dominating the clusters. However, there is no clusters specifically for public schools and this tells us that Private/Public property does not affect our clustering algorithm as expected – since we did not use that in our data – and does not have a visible correlation with the other attributes.

The scatterplot of clusters of schools in each state is as the following:

|  |  |  |  |
| --- | --- | --- | --- |
| Math SAT | Verbal SAT | ACT | # PT Undergrad |
| 0.6390 | 0.6735 | 0.7538 | 0.0115 |



In this plot, the values on x axis are states and the values on y axis are clusters. As it can be seen here, the schools in the same state can be in different clusters and state variable does not have a visible effect.

## Filling the Missing Values for Universities.

We have the information on the centers of clusters. If we use the existing information of Tufts University to find the cluster it tends to be in, we can estimate the missing values of Tufts University by taking value of the centroid of that cluster on that attribute. In my run, when using Euclidean Distance, closest centroid to Tufts University is centroid of cluster 2, so for the missing values, I used the centroid’s values.