Project 1 Statistics 6620

Edward Moradian

29 April 2018

## Introduction and Project Proposal: clustering with k-means

We want to use k-means to cluster students into two distinct educational statuses: above average educational status and below average educational status. Educational status is being defined as a student’s performance in school, involvement in educational activities, parent’s educational background and social life. An above average educational status student is one that exhibits favorable outcomes in these characteristics, whereas a below average educational status student is one that exhibits unfavorable outcomes in these characteristics.

The dataset used was the Student Performance Data Set that was taken from UCI’s machine learning repository. The dataset includes information about students at two Portuguese schools and was collected using school reports and questionaires. Student performance was measured in their mathematics classes.

## Step 1: Load the data

## Warning: package 'readr' was built under R version 3.4.3

## Parsed with column specification:  
## cols(  
## .default = col\_character(),  
## age = col\_integer(),  
## Medu = col\_integer(),  
## Fedu = col\_integer(),  
## traveltime = col\_integer(),  
## studytime = col\_integer(),  
## failures = col\_integer(),  
## famrel = col\_integer(),  
## freetime = col\_integer(),  
## goout = col\_integer(),  
## Dalc = col\_integer(),  
## Walc = col\_integer(),  
## health = col\_integer(),  
## absences = col\_integer(),  
## G1 = col\_integer(),  
## G2 = col\_integer(),  
## G3 = col\_integer()  
## )

## See spec(...) for full column specifications.

## Step 2: Exploring and preparing the data

We are keeping only the attributes that are relevant to this study. Factors with binary yes and no answers are being converted to the numeric type (i.e., 1 for no and 2 for yes).

Twenty-four variables are being examined for each student. These variables in order are age, mother’s education level (Medu), father’s education level (Fedu), home to school travel time (traveltime), weekly study time (studytime), number of past class failures (failures), extra educational support (schoolsup), family educational support (famsup), extra paid classes within the course subject (paid), extracurricular activities (activities), attended nursery school (mursery), higher education interest (higher), internet access at home (internet), involvement in a romantic relationship (romantic), quality of family relationships (famrel), free time after school (freetime), going out with friends (goout), workday alcohol consumption (Dalc), weekend alcohol consumption (Walc), current health status (health), the number of school absences (absences), first period grade (G1), second period grade (G2) and final grade (G3).

There are 395 observations of these twenty-four variables.

## Classes 'tbl\_df', 'tbl' and 'data.frame': 395 obs. of 24 variables:  
## $ age : int 18 17 15 15 16 16 16 17 15 15 ...  
## $ Medu : int 4 1 1 4 3 4 2 4 3 3 ...  
## $ Fedu : int 4 1 1 2 3 3 2 4 2 4 ...  
## $ traveltime: int 2 1 1 1 1 1 1 2 1 1 ...  
## $ studytime : int 2 2 2 3 2 2 2 2 2 2 ...  
## $ failures : int 0 0 3 0 0 0 0 0 0 0 ...  
## $ schoolsup : num 2 1 2 1 1 1 1 2 1 1 ...  
## $ famsup : num 1 2 1 2 2 2 1 2 2 2 ...  
## $ paid : num 1 1 2 2 2 2 1 1 2 2 ...  
## $ activities: num 1 1 1 2 1 2 1 1 1 2 ...  
## $ nursery : num 2 1 2 2 2 2 2 2 2 2 ...  
## $ higher : num 2 2 2 2 2 2 2 2 2 2 ...  
## $ internet : num 1 2 2 2 1 2 2 1 2 2 ...  
## $ romantic : num 1 1 1 2 1 1 1 1 1 1 ...  
## $ famrel : int 4 5 4 3 4 5 4 4 4 5 ...  
## $ freetime : int 3 3 3 2 3 4 4 1 2 5 ...  
## $ goout : int 4 3 2 2 2 2 4 4 2 1 ...  
## $ Dalc : int 1 1 2 1 1 1 1 1 1 1 ...  
## $ Walc : int 1 1 3 1 2 2 1 1 1 1 ...  
## $ health : int 3 3 3 5 5 5 3 1 1 5 ...  
## $ absences : int 6 4 10 2 4 10 0 6 0 0 ...  
## $ G1 : int 5 5 7 15 6 15 12 6 16 14 ...  
## $ G2 : int 6 5 8 14 10 15 12 5 18 15 ...  
## $ G3 : int 6 6 10 15 10 15 11 6 19 15 ...

## Step 3: Training a model on the data

We are choosing a k value of 2 becuase we want to have two different clusters - one for above average educational status and one for below average educational status.

We are using the set seed function to initialize R’s random number generator to a specific sequence so that our k-means algorithm returns the same resulting clusters each time it is run.

We are also applying z-score standardization to the dataset to scale the dataset.

## Step 4: Evaluating model performance

# 

# look at the size of the clusters

The size of the clusters is sufficiently alike in comparison of frequencies.

## [1] 146 249

# look at the cluster centers

The clusters that have been created are consistent with the objective of this study. We have two clusters of students that fall into our defined categories - above average educational status (categorized by cluster 2) and below average educational status (categorized by cluster 1).

## age Medu Fedu traveltime studytime failures  
## 1 0.3507979 -0.4467684 -0.4100131 0.3198998 -0.3850080 0.5914003  
## 2 -0.2056887 0.2619606 0.2404093 -0.1875718 0.2257477 -0.3467648  
## schoolsup famsup paid activities nursery higher  
## 1 -0.03775285 -0.02033477 -0.3144178 -0.15453613 -0.1874008 -0.3933647  
## 2 0.02213621 0.01192320 0.1843574 0.09061155 0.1098816 0.2306476  
## internet romantic famrel freetime goout Dalc  
## 1 -0.2311367 0.09005948 -0.04482666 0.1825729 0.3082734 0.3212050  
## 2 0.1355259 -0.05280596 0.02628390 -0.1070508 -0.1807547 -0.1883371  
## Walc health absences G1 G2 G3  
## 1 0.3429914 0.14313011 0.14250165 -0.8041513 -0.8598878 -0.8426157  
## 2 -0.2011114 -0.08392368 -0.08355518 0.4715104 0.5041912 0.4940638

By looking at the below table, we can see that the above average educational status students score higher than the below average educational status students in every measure related to a favorable educational status. We can see that the below average educational status students score higher than the above average educational status students in every measure related to an unfavorable educational status.

## AboveAverageEducationalStatus   
## [1,] "Mother's Educational Level"   
## [2,] "Father's Educational Level"   
## [3,] "Weekly Study Time"   
## [4,] "Extra Educational Support"   
## [5,] "Family Educational Support"   
## [6,] "Extra Paid Classes Within the Course Subject"  
## [7,] "Extracurricular Activities"   
## [8,] "Attended Nursery School"   
## [9,] "Higher Education Interest"   
## [10,] "Internet Access at Home"   
## [11,] "Quality of Family Relationships"   
## [12,] "First Period Grade"   
## [13,] "Second Period Grade"   
## [14,] "Final Grade"   
## BelowAverageEducationalStatus   
## [1,] "Age"   
## [2,] "Home to School Travel Time"   
## [3,] "Number of Past Class Failures"   
## [4,] "Involvement in a Romantic Relationship"  
## [5,] "Going out with Friends"   
## [6,] "Free Time After School"   
## [7,] "Workday Alcohol Consumption"   
## [8,] "Weekend Alcohol Consumption"   
## [9,] "Current Health Status"   
## [10,] "The Number of School Absences"   
## [11,] "NA"   
## [12,] "NA"   
## [13,] "NA"   
## [14,] "NA"

In checking to see if we get the same results with a different seed for random number generation, we do get the same results. The only difference is the labels have switched between cluster 1 and cluster 2, but the values remain the same.

## [1] 249 146

## age Medu Fedu traveltime studytime failures  
## 1 -0.2056887 0.2619606 0.2404093 -0.1875718 0.2257477 -0.3467648  
## 2 0.3507979 -0.4467684 -0.4100131 0.3198998 -0.3850080 0.5914003  
## schoolsup famsup paid activities nursery higher  
## 1 0.02213621 0.01192320 0.1843574 0.09061155 0.1098816 0.2306476  
## 2 -0.03775285 -0.02033477 -0.3144178 -0.15453613 -0.1874008 -0.3933647  
## internet romantic famrel freetime goout Dalc  
## 1 0.1355259 -0.05280596 0.02628390 -0.1070508 -0.1807547 -0.1883371  
## 2 -0.2311367 0.09005948 -0.04482666 0.1825729 0.3082734 0.3212050  
## Walc health absences G1 G2 G3  
## 1 -0.2011114 -0.08392368 -0.08355518 0.4715104 0.5041912 0.4940638  
## 2 0.3429914 0.14313011 0.14250165 -0.8041513 -0.8598878 -0.8426157

Another attempt at running a different seed shows the same result as before. The model returns the same results with a different seed. In conclusion, our model is not sensitive to the randomly-chosen cluster centers.

## [1] 249 146

## age Medu Fedu traveltime studytime failures  
## 1 -0.2056887 0.2619606 0.2404093 -0.1875718 0.2257477 -0.3467648  
## 2 0.3507979 -0.4467684 -0.4100131 0.3198998 -0.3850080 0.5914003  
## schoolsup famsup paid activities nursery higher  
## 1 0.02213621 0.01192320 0.1843574 0.09061155 0.1098816 0.2306476  
## 2 -0.03775285 -0.02033477 -0.3144178 -0.15453613 -0.1874008 -0.3933647  
## internet romantic famrel freetime goout Dalc  
## 1 0.1355259 -0.05280596 0.02628390 -0.1070508 -0.1807547 -0.1883371  
## 2 -0.2311367 0.09005948 -0.04482666 0.1825729 0.3082734 0.3212050  
## Walc health absences G1 G2 G3  
## 1 -0.2011114 -0.08392368 -0.08355518 0.4715104 0.5041912 0.4940638  
## 2 0.3429914 0.14313011 0.14250165 -0.8041513 -0.8598878 -0.8426157

## Step 5: Improving model performance

The cluster assignment for the first ten students is accurate based on the studytime, failures, absences, G1, G2 and G3 variables.

## # A tibble: 10 x 7  
## cluster studytime failures absences G1 G2 G3  
## <int> <int> <int> <int> <int> <int> <int>  
## 1 2 2 0 6 5 6 6  
## 2 2 2 0 4 5 5 6  
## 3 2 2 3 10 7 8 10  
## 4 1 3 0 2 15 14 15  
## 5 1 2 0 4 6 10 10  
## 6 1 2 0 10 15 15 15  
## 7 1 2 0 0 12 12 11  
## 8 2 2 0 6 6 5 6  
## 9 1 2 0 0 16 18 19  
## 10 1 2 0 0 14 15 15

The cluster assignment is accurate given the means of the six variables. The above average educational status students scored higher in weekly study time, lower in the number of past class failures, lower in the number of school absences, higher in the first period grade, higher in the second period grade and higher in the final grade. The below average educational status students scored lower in weekly study time, higher in the number of past class failures, higher in the number of school absences, lower in the first period grade, lower in the second period grade and lower in the final grade.

## cluster studytime  
## 1 1 2.224900  
## 2 2 1.712329

## cluster failures  
## 1 1 0.07630522  
## 2 2 0.77397260

## cluster absences  
## 1 1 5.040161  
## 2 2 6.849315

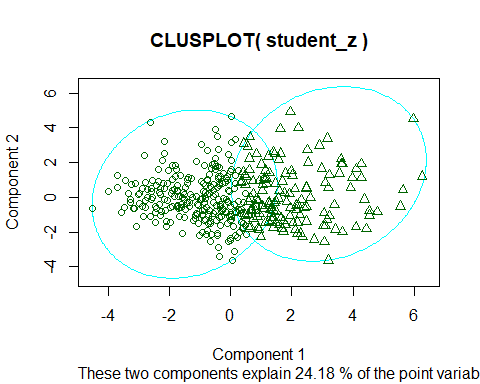
## cluster G1  
## 1 1 12.473896  
## 2 2 8.239726

## cluster G2  
## 1 1 12.610442  
## 2 2 7.479452

## cluster G3  
## 1 1 12.678715  
## 2 2 6.554795

## Conclusions

By clustering with k-means, we were able to cluster students into two groups - above average educational status and below average educational status. This would be useful to any education sector to help identify students who may need to be targeted with educational aid in the form of after-school programs, counseling, or extracurricular activities. If the data set had a student identification number, this can be used to identify students to target in these aids. Many schools already identify students as having a satisfactory standing or not. A k-means clustering algorithm could be used as an objective way in identifying students as having a satisfactory standing or not. The algorithm’s specifications can be changed to include different features of a student’s educational background as necessary. The k-value can be changed as to possibly group students in a way to find students of very high educational status and students of very low educational status with a k-value of 5.

A cluster plot summarizes visually the clustering that happened into the two groups - above average educational status and below average educational status. 

## 

## APPENDIX

library(readr)  
student <- read\_delim("student-mat.csv", ";", escape\_double = FALSE, trim\_ws = TRUE)

student <- student[c(3,7,8,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33)]  
student$schoolsup <- as.numeric(as.factor(student$schoolsup))  
student$famsup <- as.numeric(as.factor(student$famsup))  
student$paid <- as.numeric(as.factor(student$paid))  
student$activities <- as.numeric(as.factor(student$activities))  
student$nursery <- as.numeric(as.factor(student$nursery))  
student$higher <- as.numeric(as.factor(student$higher))  
student$internet <- as.numeric(as.factor(student$internet))  
student$romantic <- as.numeric(as.factor(student$romantic))  
str(student)

student\_z <- as.data.frame(lapply(student, scale))  
set.seed(2345)  
student\_clusters <- kmeans(student\_z, 2)

student\_clusters$size

student\_clusters$centers

AboveAverageEducationalStatus <- c("Mother's Educational Level","Father's Educational Level","Weekly Study Time","Extra Educational Support","Family Educational Support","Extra Paid Classes Within the Course Subject","Extracurricular Activities","Attended Nursery School","Higher Education Interest","Internet Access at Home","Quality of Family Relationships","First Period Grade","Second Period Grade","Final Grade")  
BelowAverageEducationalStatus <- c("Age","Home to School Travel Time","Number of Past Class Failures","Involvement in a Romantic Relationship","Going out with Friends","Free Time After School","Workday Alcohol Consumption","Weekend Alcohol Consumption","Current Health Status","The Number of School Absences","NA","NA","NA","NA")  
cbind(AboveAverageEducationalStatus,BelowAverageEducationalStatus)

set.seed(66666665)  
student\_clusters <- kmeans(student\_z, 2)  
student\_clusters$size  
student\_clusters$centers

set.seed(5555)  
student\_clusters <- kmeans(student\_z, 2)  
student\_clusters$size  
student\_clusters$centers

student$cluster <- student\_clusters$cluster  
student[1:10,c("cluster","studytime","failures","absences","G1","G2","G3")]

aggregate(data=student,studytime ~ cluster,mean)  
aggregate(data=student,failures ~ cluster,mean)  
aggregate(data=student,absences ~ cluster,mean)  
aggregate(data=student,G1 ~ cluster,mean)  
aggregate(data=student,G2 ~ cluster,mean)  
aggregate(data=student,G3 ~ cluster,mean)

clusplot(student\_z, student\_clusters$cluster)