K means Clustering

Observe change in learning motivation & grouping students

Eddie Lin

2018/03/17

Project Description

K means is a commonly used technique in statistics and machine learning. It is an efficient way to clustering all data point in our dataset. It is unsupervised despite we have to assign (and try) cluster numbers to achieve an optimal outcome. The logic of K means is to create different cluster of data points, where all data points in a particular cluster will be near each other and not the others in other clusters. In this project, I will use K means to analyze student motivation and students' background. Major research question include:

1. How do we use K means to categorize students' learning motivation

2. How to use K means to group students in a class

Tools & Data

- R
- R packages: dplyr, tidyr, ggplot2
- Data source: a self-report class survey containing (1)students' learning motivation throughout 5 weeks' class and (2) students' background info. and preference.

Data Description

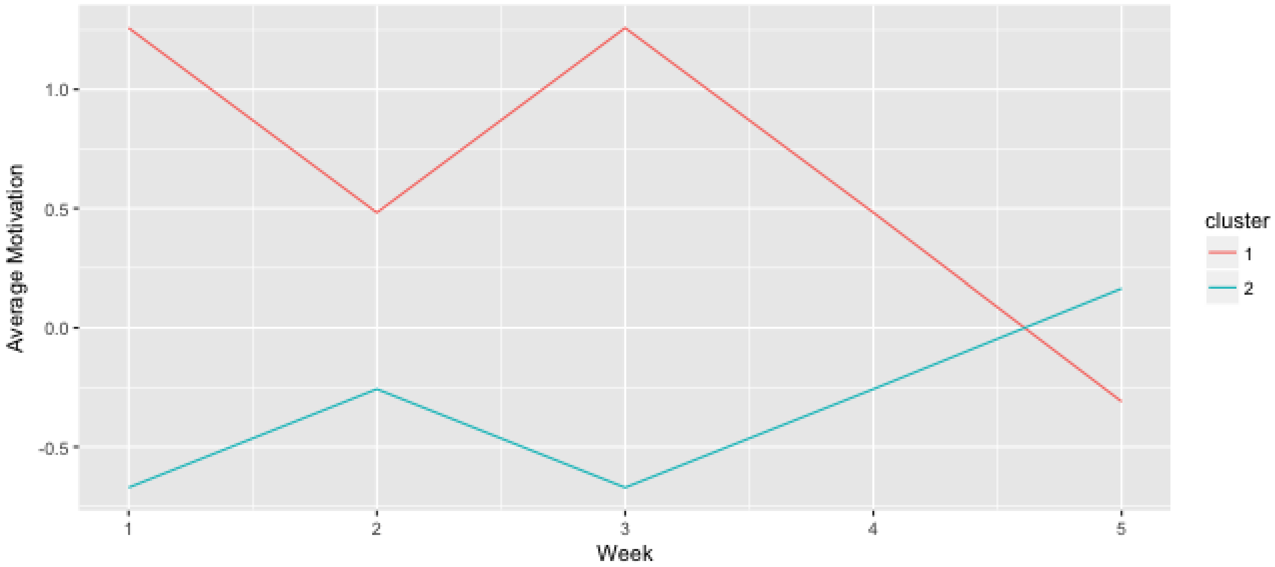
The data set has been anonymized for privacy reason. Questions are divided into 2 parts: learning motivation & student info. Their specific questions are as follows:

- Part 1:
  - how motivated are you about learning in this class this week (week 1~5, 5-points Likert scale)
- Part 2:
  - Q1 : Have you ever owned a cat?
  - Q2 : Do you pronounce "gif", with a J (j-iff) or a G (g-iff)?
  - Q3 : How many months have you lived in New York City?
  - Q4 : How many siblings (brothers/sisters) do you have?
  - Q5 : How many times do you play sport each week?
  - Q6 : How many miles do you travel from home to school?
  - Q7 : Estimate how many of your friends own Android phones
  - Q8 : How many movies have you seen in the cinema this year?
  - Q9 : How many classes are you taking this semester?
  - Q10 : How many states have you visited in the US?
  - Q11 : What city/town did you grow up in?
  - Q12 : What state/province did you grow up in?
  - Q13 : What country did you grow up in?

Motivation variation throughout the 5 weeks

- Let's plot the motivation scores in the 2 cluster throughout 5 weeks

```
K5 <- tidyr::gather(K4, "week", "motivation", 1:5)
K6 <- K5 %>% group_by(week, cluster)
K6 <- summarise(K6, avg = mean(motivation))
K6$week <- as.numeric(K6$week)
K6$cluster <- as.factor(K6$cluster)
ggplot(K6, aes(week, avg, colour = cluster)) + geom_line() + xlab("We
```



Try 3 clusters

- It takes a bit of hacking to get the "right" number of clusters in K means
- This means for some data set, there may not be a correct number of clusters
- Sometimes cluster number can be inferred based on similar research or some practical experience from the experts _ We will take a exploratory step to use **cluster** = **3**

Try 3 clusters

```
fit.g3 <- kmeans(K3, 3)
K4.g3 <- data.frame(K3, fit.g3$cluster)
names(K4.g3) <- c("1", "2", "3", "4", "5", "cluster")
K5.g3 <- tidyr::gather(K4.g3, "week", "motivation", 1:5)
K6.g3 <- K5.g3 %>% group_by(week, cluster)
K6.g3 <- summarise(K6.g3, avg = mean(motivation))
K6.g3$week <- as.numeric(K6.g3$week)
K6.g3$cluster <- as.factor(K6.g3$cluster)
ggplot(K6.g3, aes(week, avg, colour = cluster)) + geom_line() + xlab
```

Clustering based on students' interest & preference

- Another thing we can do with K means in education is to group students based on their responses to questions in a survey

- We will do this with part of the Part 2 survey questions and `recode` the factor features into numeric features

```
DF1 <- read.table("cluster-class-data.csv", sep = ",", header  = TRUE
names(DF1) <- c("studnetID", "ownCat", "jORgif", "monNYC", "numSib",

DF.ANSWER <- DF1[, 2:11]
DF.REGION <- DF1 [, 12:14]
DF.ANSWER$ownCat <- as.character(DF.ANSWER$ownCat)

# transfer facotral variables
DF.ANSWER$ownCat <- ifelse(DF.ANSWER$ownCat == "Yes" & !is.na(DF.ANSW
DF.ANSWER$jORgif <- ifelse(DF.ANSWER$jORgif == "j-iff" & !is.na(DF.AN
DF.ANSWER <- scale(DF.ANSWER)
```

Run K means with 6 clusters

```
library(cluster)
set.seed(123456)
ANSWER.CLUSTERS <- kmeans(DF.ANSWER, 6)
ANSWER.CLUSTERS$cluster
```

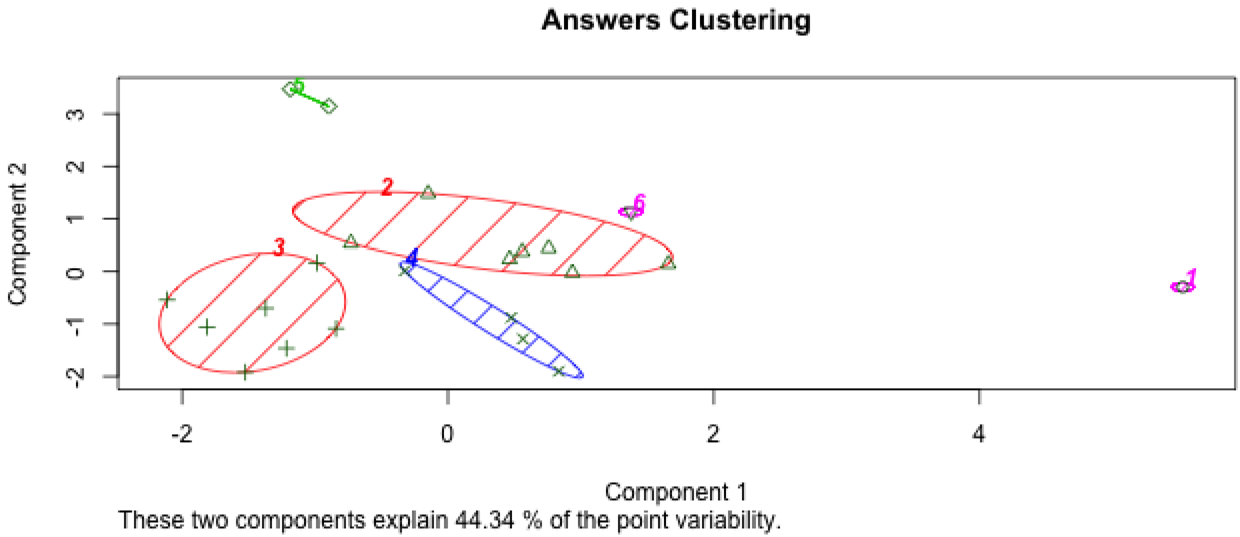
```
##  [1] 3 2 2 6 2 1 4 3 3 2 3 2 5 2 5 2 4 4 3 3 3 4
```

```
ANSWER.FINAL <- data.frame(DF.ANSWER, cluster = ANSWER.CLUSTERS$clust
ANSWER.CLUSTERS$size
```

```
## [1] 1 7 7 4 2 1
```

Run K means with 6 clusters

```
clusplot(DF.ANSWER, ANSWER.CLUSTERS$cluster,lines = 0, labels =5, co
```



Findings & Summary

- In this project, we use students' self-report motivation score and their responses about their interests and preference to group them.
- We found that during the 5 weeks' class, there is high/ low motivation cluster, as well as, another cluster that is somewhere in between.
- In practice, we can further investigate the attributes of these different clusters of students. We can also do this by conducting a focus group/individual interview to know how they felt in each week 's class
- We also use a number of interest & preference question to create student group (clusters) that may be useful for group activities.

Limitations & Suggestions

No data analytics is perfect, I came up with a few thoughts and make some suggestions in the followings:

- K means is useful, but the machine clusters people only based on numbers and mean distance. So we shouldn’t be surprised if some grouping is strange is a bit strange in human eyes.
- The number of clusters is less about being precise and takes a bit luck with some theoretical foundation to get if right. It depends on the purpose of data analytics to settle on an optimal point where clustering result makes the most sense
- K means, like some other algorithms, may be subject to outliers of data points especially when there is a systematic bias. It is useful to visualize data points and think about what to do with outliers before putting them into the algorithm. Also, don’t forget to **scale()** it.