Capstone Report

Fábio Franco Costa

November 4, 2015

Title

Do similar users rates places equally?

Introduction

When you are using apps like YELP, you get ratings based on the average reviews from all users. In that situation I may get good ratings for a place that I would not really appreciate, because not every user are similar to others. My problem is testing if clusters of similar users rate places equally, so that I can use the average rate of this clusters to predict how I would rate some new place.

Methods and Data

For our analysis, the first step was to decide a clustering strategy. We decided to use the features in the user database as our characteristics. We did some transformations and end up with the following structure:

```
## 'data.frame':    50000 obs. of  25 variables:
##  $ user_id            : chr  "Kb2FOnGteVLhN0ZhsmgqNw" "YlhgtRS9yArNolg8j_tpgw" "UBPKjtrReZ01gOD3953T
##  $ fans               : int  0 5 0 0 0 0 1 2 1 0 ...
##  $ average_stars      : num  3.67 3.38 4.5 3 4.33 3 5 4.26 3.59 3.67 ...
##  $ votes.funny        : int  1 269 0 4 0 9 0 3 24 1 ...
##  $ votes.useful       : int  2 293 1 5 6 15 1 17 125 2 ...
##  $ votes.cool         : int  2 206 0 2 2 1 0 5 25 1 ...
##  $ compliments.profile: num  0 4 0 0 0 0 0 0 0 0 ...
##  $ compliments.cute   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ compliments.funny  : num  0 32 0 0 0 0 0 0 1 0 ...
##  $ compliments.plain  : num  0 29 0 0 0 0 0 0 3 0 ...
##  $ compliments.writer : num  1 11 0 0 0 0 0 0 1 0 ...
##  $ compliments.note   : num  0 22 0 0 2 1 0 0 4 0 ...
##  $ compliments.photos : num  0 11 0 0 0 0 0 1 0 0 ...
##  $ compliments.hot    : num  0 65 0 0 0 0 0 2 0 0 ...
##  $ compliments.cool   : num  0 101 0 0 0 0 0 0 4 0 ...
##  $ compliments.more   : num  0 3 0 0 0 0 0 0 0 0 ...
##  $ compliments.list   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ reviews            : int  3 233 4 2 10 7 1 36 97 3 ...
##  $ friends_no         : int  0 120 0 1 0 0 9 21 22 0 ...
##  $ yelping_months     : num  39 103 28 17 27 65 37 27 84 46 ...
##  $ C1                 : int  13 11 13 13 13 7 13 13 10 13 ...
##  $ C2                 : int  7 5 7 7 7 13 7 7 3 7 ...
##  $ C3                 : int  11 9 11 11 11 15 11 11 7 11 ...
##  $ C4                 : int  14 5 14 14 14 15 14 14 7 14 ...
##  $ C5                 : int  10 1 10 10 10 4 10 10 12 10 ...
```

We clustered the data using K means, for 15 clusters using the following command:

```
cl_users <- kmeans(mtx_user, clusters_no,
                   iter.max = 1000, nstart = 5)
```

After that we used Hypothesis Test to check if different groups have different average rates for the same business. We used a 90% confidence interval to our analysis because we didn't want to impose a very strict test.

Results

We found that the

Discussion

- Explain how you interpret the results of your analysis and what the implications are for your question/problem.