SD_Assignment 4

Steve Dunn

July 13, 2016

Prof. Andy Catlin IS 643 - Special Topics: Recommender Systems Department of Data Analytics, City University of New York

For this project I am using Content-based Filtering which involves analyzing an item that may be similar to other items another user have selected or recommended based on Genres or other specific features. I am using a subset of the Movie Lens data set to recommend movies by using Hierarchial clustering.Clustering is an unsupervised learning method where the goal is to segment data into similar groups.The data set was split into 10 clusters this allows for grouping of the movies across multiple genres seeing that some movies have multiple genres . The movies are recommended not based on the rating but by similarity of genres.

Looking a the structure of the dataset I can see that the data for the Genres contain 0's and 1's which indicates that a particular Movie belongs to a particular Genre from this observation it can be seen that some movies have multiple genres.

```
str(movies)
```

```
## 'data.frame':    1664 obs. of  20 variables:
##  $ Title      : Factor w/ 1664 levels "'Til There Was You (1997)",..: 1525 618 555 594 344 1318 1545
##  $ Unknown    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Action     : int  0 1 0 1 0 0 0 0 0 0 ...
##  $ Adventure  : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ Animation  : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ Childrens  : int  1 0 0 0 0 0 0 1 0 0 ...
##  $ Comedy     : int  1 0 0 1 0 0 0 1 0 0 ...
##  $ Crime      : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ Documentary: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Drama      : int  0 0 0 1 1 1 1 1 1 1 ...
##  $ Fantasy    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ FilmNoir   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Horror     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Musical    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Mystery    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Romance    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ SciFi      : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ Thriller   : int  0 1 1 0 1 0 0 0 0 0 ...
##  $ War        : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ Western    : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
head(movies)
```

```
##                                                Title Unknown Action
## 1                                  Toy Story (1995)       0      0
## 2                                  GoldenEye (1995)       0      1
## 3                                 Four Rooms (1995)       0      0
## 4                                 Get Shorty (1995)       0      1
## 5                                    Copycat (1995)       0      0
## 6 Shanghai Triad (Yao a yao yao dao waipo qiao) (1995)       0      0
```

```
##   Adventure Animation Childrens Comedy Crime Documentary Drama Fantasy
## 1         0         1         1      1     0           0     0       0
## 2         1         0         0      0     0           0     0       0
## 3         0         0         0      0     0           0     0       0
## 4         0         0         0      1     0           0     1       0
## 5         0         0         0      0     1           0     1       0
## 6         0         0         0      0     0           0     1       0
##   FilmNoir Horror Musical Mystery Romance SciFi Thriller War Western
## 1        0      0       0       0       0     0        0   0       0
## 2        0      0       0       0       0     0        1   0       0
## 3        0      0       0       0       0     0        1   0       0
## 4        0      0       0       0       0     0        0   0       0
## 5        0      0       0       0       0     0        1   0       0
## 6        0      0       0       0       0     0        0   0       0
```

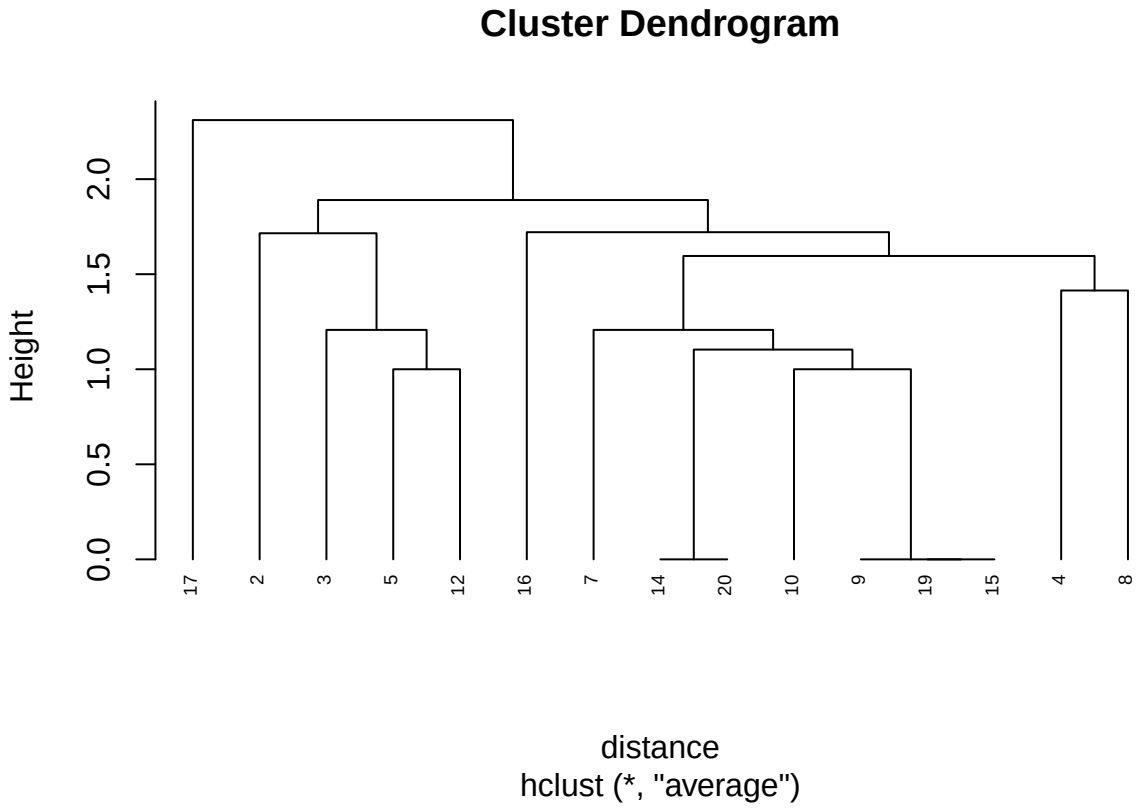
create a subset from the data and calculate the distance matrix which Computes the distance between each pairs of points per cluster. I utilized the dist function for this purpose.The most popular distance methos is Euclidean distance .After calculating the distance I used a dendogram to represent the data.I looked at a subset of the data as looking at the entire data was not very clear

```
set.seed(1306)
sammovies <- movies[sample(2:20, 15),] # sample from the data set

distance<- dist(sammovies[2:20], method="euclidean")
distance
```

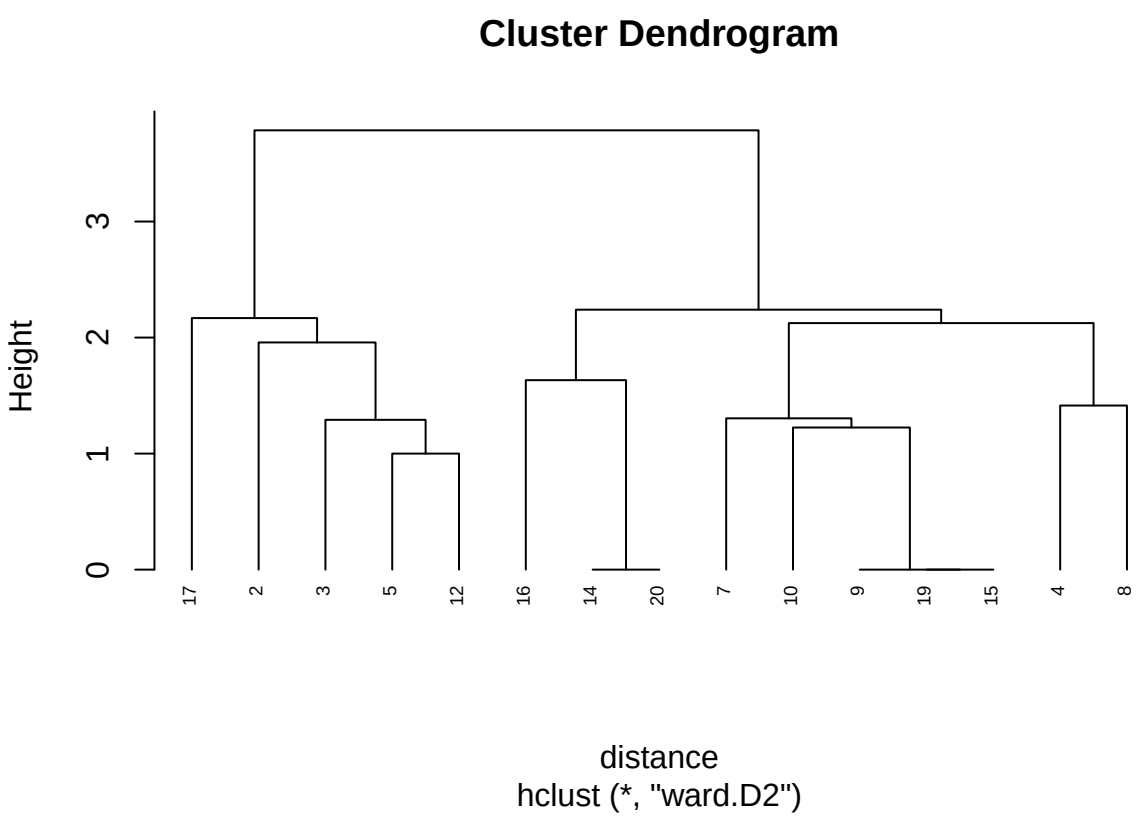
```
##          16        5       10        3       12       14        7       20
## 5  2.236068
## 10 2.000000 1.732051
## 3  1.732051 1.414214 1.732051
## 12 2.000000 1.000000 2.000000 1.000000
## 14 1.414214 1.732051 1.414214 1.732051 2.000000
## 7  2.000000 1.732051 1.414214 1.732051 2.000000 1.414214
## 20 1.414214 1.732051 1.414214 1.732051 2.000000 0.000000 1.414214
## 17 2.236068 2.000000 2.645751 2.000000 1.732051 2.645751 2.645751 2.645751
## 4  1.732051 2.000000 1.732051 2.000000 2.236068 1.732051 1.732051 1.732051
## 19 1.732051 1.414214 1.000000 1.414214 1.732051 1.000000 1.000000 1.000000
## 2  2.236068 2.000000 2.236068 1.414214 1.732051 2.236068 2.236068 2.236068
## 15 1.732051 1.414214 1.000000 1.414214 1.732051 1.000000 1.000000 1.000000
## 8  1.732051 2.000000 1.732051 2.000000 2.236068 1.732051 1.732051 1.732051
## 9  1.732051 1.414214 1.000000 1.414214 1.732051 1.000000 1.000000 1.000000
##          17        4       19        2       15        8
## 5
## 10
## 3
## 12
## 14
## 7
## 20
## 17
## 4  2.000000
## 19 2.449490 1.414214
## 2  2.000000 2.000000 2.000000
## 15 2.449490 1.414214 0.000000 2.000000
```

```
## 8  2.449490 1.414214 1.414214 2.449490 1.414214
## 9  2.449490 1.414214 0.000000 2.000000 0.000000 1.414214
```

```
cluster<- hclust(distance, method="average")

#The vertical lines represents the distance
plot(cluster, hang=-1,cex = 0.6)
```




```
clustMovies = hclust(distance, method = "ward.D2")
#rect.hclust(hclust(distances, method = "ward.D2"),h=3)

plot(clustMovies, hang=-1,cex = 0.6)
```

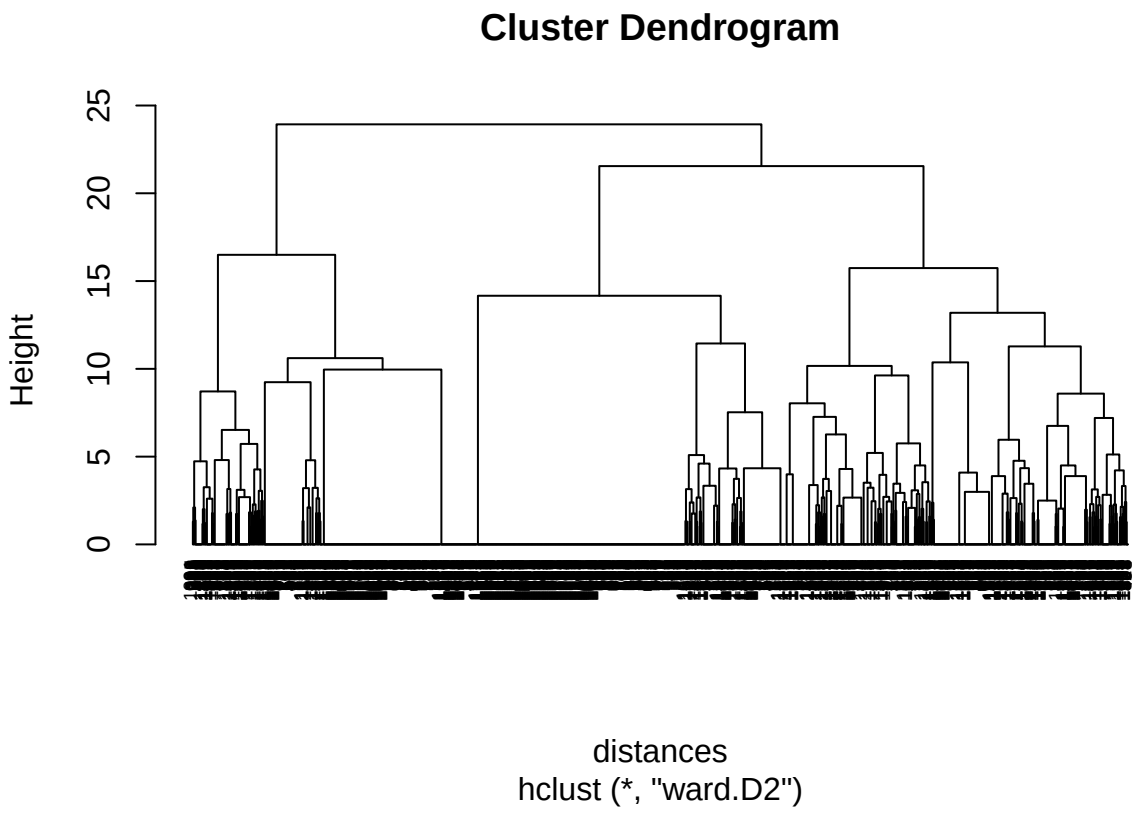



```
#cuts the tree in to groups of 10
cutree(clustMovies,10)
```

```
## 16  5 10  3 12 14  7 20 17  4 19  2 15  8  9
##  1  2  3  4  2  5  6  5  7  8  3  9  3 10  3
```

using differnt method to dispay the Dendogram

```
# Calculate distances between genre features:
distances = dist(movies[2:20], method = "euclidean")

clustMovies = hclust(distances, method = "ward.D2")

plot(clustMovies, hang=-1,cex = 0.6)
```




```
#Label each movie in clusters with k=10 clusters
clustGrps = cutree(clustMovies, k = 10)
```

Calculate average for Comedy and Romance genre within each cluster

```
tapply(movies$Comedy, clustGrps, mean)
```

```
##          1          2          3          4          5          6
## 0.34645669 0.12790698 0.11111111 0.12195122 0.00000000 0.22033898
##          7          8          9         10
## 1.00000000 0.09166667 0.83809524 0.10280374
```

```
tapply(movies$Horror, clustGrps, mean)
```

```
##          1          2          3          4          5          6
## 0.00000000 0.06201550 0.07407407 0.02439024 0.00000000 0.00000000
##          7          8          9         10
## 0.00000000 0.01666667 0.00000000 0.54205607
```

check to see which cluster the movie Clerks and Phycho are in and similar movies to them that would be in the same or similar Genres

```
subset(movies, Title=="Toy Story (1995)")
```

```
##              Title Unknown Action Adventure Animation Childrens Comedy
## 1 Toy Story (1995)       0      0         0         1         1      1
##   Crime Documentary Drama Fantasy FilmNoir Horror Musical Mystery Romance
## 1     0           0     0       0        0      0       0       0       0
##   SciFi Thriller War Western
## 1     0        0   0       0
```

```
cluster2 = subset(movies, clustGrps==1)
cluster2$Title[1:10]
```

```
##  [1] Toy Story (1995)
##  [2] Babe (1995)
##  [3] Free Willy 2: The Adventure Home (1995)
##  [4] Santa Clause, The (1994)
##  [5] Lion King, The (1994)
##  [6] Mask, The (1994)
##  [7] Free Willy (1993)
##  [8] Home Alone (1990)
##  [9] Aladdin (1992)
## [10] Snow White and the Seven Dwarfs (1937)
## 1664 Levels: 'Til There Was You (1997) ... Zeus and Roxanne (1997)
```

```
subset(movies, Title=="Psycho (1960)")
```

```
##             Title Unknown Action Adventure Animation Childrens Comedy
## 185 Psycho (1960)       0      0         0         0         0      0
##     Crime Documentary Drama Fantasy FilmNoir Horror Musical Mystery
## 185     0           0     0       0        0      1       0       0
##     Romance SciFi Thriller War Western
## 185       1     0        1   0       0
```

```
cluster3 = subset(movies, clustGrps==3)
cluster3$Title[1:10]
```

```
##  [1] Four Rooms (1995)
##  [2] Taxi Driver (1976)
##  [3] Disclosure (1994)
##  [4] Dolores Claiborne (1994)
##  [5] Firm, The (1993)
##  [6] Blade Runner (1982)
##  [7] So I Married an Axe Murderer (1993)
##  [8] Silence of the Lambs, The (1991)
##  [9] Diabolique (1996)
## [10] Lone Star (1996)
## 1664 Levels: 'Til There Was You (1997) ... Zeus and Roxanne (1997)
```

weighting according to clusters for variable within the cluster

```
binsp = split(movies[2:20], clustGrps)
lapply(binsp, colMeans)
```

```
## $`1`
##     Unknown      Action   Adventure   Animation   Childrens      Comedy
## 0.000000000 0.070866142 0.338582677 0.307086614 0.889763780 0.346456693
##       Crime Documentary       Drama     Fantasy    FilmNoir      Horror
## 0.007874016 0.000000000 0.157480315 0.165354331 0.000000000 0.000000000
##     Musical     Mystery     Romance       SciFi    Thriller         War
## 0.149606299 0.000000000 0.031496063 0.062992126 0.007874016 0.000000000
##     Western
## 0.000000000
##
## $`2`
##     Unknown      Action   Adventure   Animation   Childrens      Comedy
## 0.000000000 0.732558140 0.333333333 0.011627907 0.011627907 0.127906977
##       Crime Documentary       Drama     Fantasy    FilmNoir      Horror
## 0.031007752 0.000000000 0.147286822 0.000000000 0.000000000 0.062015504
##     Musical     Mystery     Romance       SciFi    Thriller         War
## 0.000000000 0.007751938 0.046511628 0.317829457 0.298449612 0.038759690
##     Western
## 0.100775194
##
## $`3`
##     Unknown      Action   Adventure   Animation   Childrens      Comedy
##  0.01234568  0.03703704  0.00000000  0.00000000  0.01234568  0.11111111
##       Crime Documentary       Drama     Fantasy    FilmNoir      Horror
##  0.09876543  0.00000000  0.31481481  0.00617284  0.14197531  0.07407407
##     Musical     Mystery     Romance       SciFi    Thriller         War
##  0.00000000  0.35185185  0.06790123  0.03703704  0.79012346  0.00000000
##     Western
##  0.00000000
##
## $`4`
##     Unknown      Action   Adventure   Animation   Childrens      Comedy
##  0.00000000  0.17073171  0.01219512  0.00000000  0.00000000  0.12195122
##       Crime Documentary       Drama     Fantasy    FilmNoir      Horror
##  0.97560976  0.00000000  0.48780488  0.00000000  0.00000000  0.02439024
##     Musical     Mystery     Romance       SciFi    Thriller         War
##  0.00000000  0.01219512  0.07317073  0.01219512  0.32926829  0.00000000
##     Western
##  0.00000000
##
## $`5`
##     Unknown      Action   Adventure   Animation   Childrens      Comedy
##           0           0           0           0           0           0
##       Crime Documentary       Drama     Fantasy    FilmNoir      Horror
##           0           0           1           0           0           0
##     Musical     Mystery     Romance       SciFi    Thriller         War
##           0           0           0           0           0           0
##     Western
##           0
##
```

```
## $`6`
##     Unknown      Action   Adventure   Animation   Childrens      Comedy
##  0.00000000  0.22033898  0.03389831  0.00000000  0.00000000  0.22033898
##       Crime Documentary       Drama     Fantasy    FilmNoir      Horror
##  0.00000000  0.01694915  0.59322034  0.00000000  0.00000000  0.00000000
##     Musical     Mystery     Romance       SciFi    Thriller         War
##  0.00000000  0.00000000  0.20338983  0.03389831  0.05084746  1.00000000
##     Western
##  0.01694915
##
## $`7`
##     Unknown      Action   Adventure   Animation   Childrens      Comedy
##   0.0000000   0.0000000   0.0000000   0.0000000   0.0000000   1.0000000
##       Crime Documentary       Drama     Fantasy    FilmNoir      Horror
##   0.0000000   0.0000000   0.2372263   0.0000000   0.0000000   0.0000000
##     Musical     Mystery     Romance       SciFi    Thriller         War
##   0.0000000   0.0000000   0.0000000   0.0000000   0.0000000   0.0000000
##     Western
##   0.0000000
##
## $`8`
##     Unknown      Action   Adventure   Animation   Childrens      Comedy
## 0.000000000 0.116666667 0.000000000 0.000000000 0.000000000 0.091666667
##       Crime Documentary       Drama     Fantasy    FilmNoir      Horror
## 0.016666667 0.000000000 0.666666667 0.000000000 0.008333333 0.016666667
##     Musical     Mystery     Romance       SciFi    Thriller         War
## 0.000000000 0.000000000 1.000000000 0.000000000 0.091666667 0.008333333
##     Western
## 0.000000000
##
## $`9`
##     Unknown      Action   Adventure   Animation   Childrens      Comedy
##  0.00000000  0.03809524  0.00952381  0.00000000  0.01904762  0.83809524
##       Crime Documentary       Drama     Fantasy    FilmNoir      Horror
##  0.00000000  0.00000000  0.08571429  0.00000000  0.00000000  0.00000000
##     Musical     Mystery     Romance       SciFi    Thriller         War
##  0.35238095  0.00000000  0.75238095  0.00952381  0.00952381  0.00952381
##     Western
##  0.00000000
##
## $`10`
##     Unknown      Action   Adventure   Animation   Childrens      Comedy
##  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000  0.10280374
##       Crime Documentary       Drama     Fantasy    FilmNoir      Horror
##  0.00000000  0.45794393  0.07476636  0.00000000  0.00000000  0.54205607
##     Musical     Mystery     Romance       SciFi    Thriller         War
##  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000
##     Western
##  0.00000000
```

Adding Context to the recommender by recommending movies with similar genes allows for a possible wider choice in selection as opposed to UBCF where the recommendation is based on other user who rated similar movies. If we look at Toy Story I can see that it spans Animation,Childrens and Comedy genres and gives a recommendation for Mask, The (1994) that would not normally be classified as animation or comedy.