HW GMM

Group F

1/22/2020

Problem 1

Reading in the dataset - tradeshow.csv

```
tds <- read.csv("/Users/shreyashiganguly/Documents/Northwestern_MSiA/Winter
2020/Data Mining/HW2/tradeshow.csv")
colnames(tds) <- c("buy","social","educ")
```

Part(a) - KMeans clustering

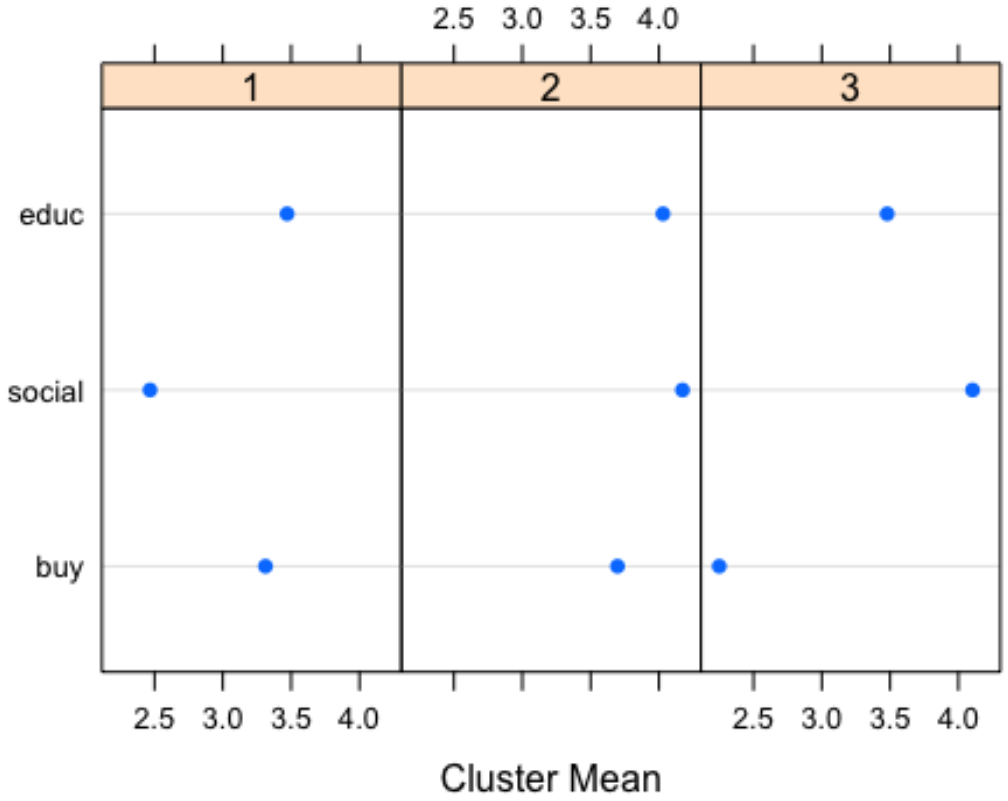
```
set.seed(12345)
fit.tds3 = kmeans(tds, 3, nstart=100)

#Cluster sizes, means, RMSE
summary(fit.tds3)
```

```
##     n  Pct  buy social educ   RMSE
## 1 169 0.38 3.31   2.47 3.47 0.6076
## 2 170 0.38 3.70   4.17 4.03 0.5578
## 3 106 0.24 2.25   4.10 3.48 0.6534
##   445 1.00 3.21   3.51 3.69 0.6006
## SSE= 478.3618 ; SSB= 467.9116 ; SST= 946.2733
## R-Squared =  0.4944782
## Pseudo F =  216.1721
```

```
plot(fit.tds3)
```

```
## Loading required package: lattice
```



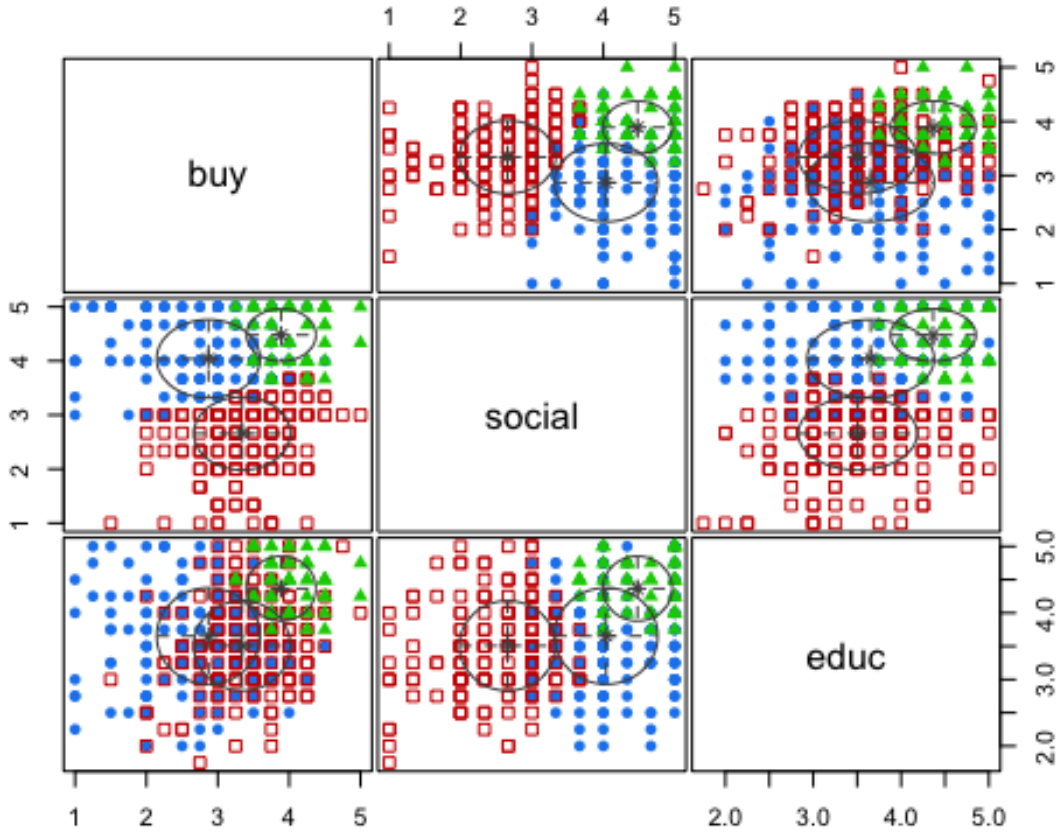
#

Cluster Descriptions

- Non social - here to educate themselves and buy some
- Ambitious - here to do everything
- Non buyer - here to network and educate themselves, not to buy

Part(b) - Gaussian Mixture (VII)

```
fit.tds.gmm = Mclust(tds,G=3,modelNames="VII")
plot(fit.tds.gmm, what = "classification")
```




```
fit.tds.gmm$parameters$pro
```

```
## [1] 0.4395187 0.4265838 0.1338976
```

```
fit.tds.gmm$parameters$mean
```

```
##             [,1]     [,2]     [,3]
## buy    2.868500 3.339815 3.888867
## social 4.038811 2.657831 4.478752
## educ   3.652959 3.506621 4.365602
```

```
sqrt(fit.tds.gmm$parameters$variance$sigmasq)
```

```
## [1] 0.7245096 0.6759707 0.4863842
```

Observations

- Though the three clusters have the same descriptions, there is better distinction in their values now
  - Cluster 1 - Non buyer
  - Cluster 2 - Non social
  - Cluster 3 - Ambitious

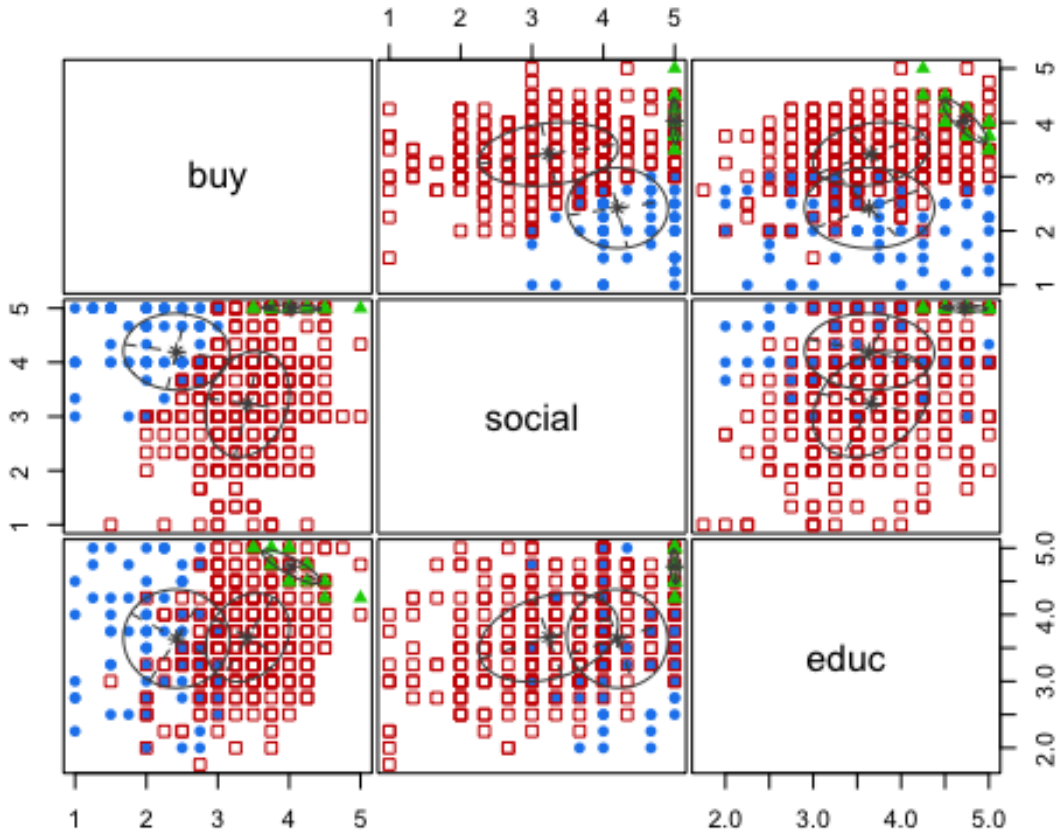
- K-means churned out almost equal sized clusters. However GMM has made the 'Ambitious' cluster almost one-third the size of the other two clusters. This solution makes more sense as there must be only a handful of 'ambitious' people, intuitively.
- Both K-Means and GMM clusters have the same ordering of within cluster variances Ambitious < Non Social < Non Buyer K-Means gives very small difference in the RMSE values, GMM depicts larger differences
- Number of variance parameters estimated = 3

Part (c) - Gaussian Mixture (VVE)

```
fit.tds.gmm1 = Mclust(tds,G=3)
fit.tds.gmm1$parameters$mean
```

```
##              [,1]     [,2]     [,3]
## buy    2.420752 3.417206 4.030997
## social 4.195782 3.231188 4.999999
## educ   3.639689 3.657019 4.721012
```

```
plot(fit.tds.gmm1, what = "classification")
```




```
fit.tds.gmm1$parameters$variance$sigma
```

```
## , , 1
##
##                 buy      social         educ
## buy     0.553753113 -0.01041070 -0.006476061
## social -0.010410703  0.50153507 -0.017107224
## educ   -0.006476061 -0.01710722  0.546659790
##
## , , 2
##
##               buy    social       educ
## buy    0.34510202 0.1169826 0.08611469
## social 0.11698262 0.9540970 0.19174011
## educ   0.08611469 0.1917401 0.44021197
##
## , , 3
##
##                 buy       social         educ
## buy     0.167057935 -0.003870663 -0.088292319
## social -0.003870663  0.004554370 -0.003208518
## educ   -0.088292319 -0.003208518  0.065283215
```

Observations

- Though the three clusters have the same descriptions, there is better distinction in their values now
  - Cluster 1 - Non buyer
  - Cluster 2 - Non social
  - Cluster 3 - Ambitious
- Variance Model = VVE (BIC largest?) Class-conditional distributions : Variable volume, variable shape, variable orientation (classification plot?)
- Number of variance parameters estimated = 18

Part (d) - Solution Preferred

GMM with VVE model - smaller uncertainties