STAT40740 Assignment 2

Darren Gilligan ID: 14206345

2 April 2016

1. In R, write your own function which answers the question: are the two communities significantly different with respect to the characteristics of the properties available for sale? [10 marks]

```
prices <- read.csv('./prices(1).csv')



prices.ftest <- function (prices) {
  mp.prices <- prices[prices$setArea == 'MP',-1]
  pa.prices <- prices[prices$setArea == 'PA',-1]

  mu.pa <- as.matrix(sapply(pa.prices,mean))
  mu.mp <- as.matrix(sapply(mp.prices,mean))


  mp.prices <- as.matrix(mp.prices)

  pa.prices <- as.matrix(pa.prices)

  S1 <- cov(pa.prices)
  S2 <- cov(mp.prices)
  n1 <- dim(pa.prices)[1]
  n2 <- dim(mp.prices)[1]
  p = 4

  S <- (n1 - 1) * S1 + (n2 - 1) * S2 / (n1 + n2 - 2)

  D_sqrd = t(mu.pa - mu.mp) %*% solve(S) %*% (mu.pa - mu.mp)
  T_sqrd = ((n1 * n2) * D_sqrd) / (n1 + n2)

  F = ((n1 + n2 - p - 1) * T_sqrd) / ((n1 + n2 - 2) * p)

  F_crit = qf(.95, df1 = p, df2 = n1 + n2 - p - 1)

  if (F > F_crit)
    return ("Reject null hypothesis for alpha = 0.05")
  else
    return ("Fail to reject null hypothesis and accept alternate hypothesis for alpha = 0.05")
}
prices.ftest(prices)
```

```
## [1] "Fail to reject null hypothesis and accept alternate hypothesis for alpha = 0.05"
```

Our answer here is the two communities **do not** significantly differ with respect to the characteristics of the properties available for sale.

2 (a) Load the voting data into R. As the data are binary in nature, which of the clustering methods that we have seen to date could be used to cluster the TDs? Apply your chosen clustering method to the voting data, detailing any decisions you make in the process. How many clusters of TDs do you think are present?

First I decided what distance measure to use. In this case measuring a binary distance is best as euclidean distance doesn't make sense for binary data as binary distances don't fit onto the continuous euclidean space. I used simple matching on binary data as the distance measure

Then I selected the cluster type. K-means generally using euclidean distance for continuous data so we will choose hierarchical clustering.

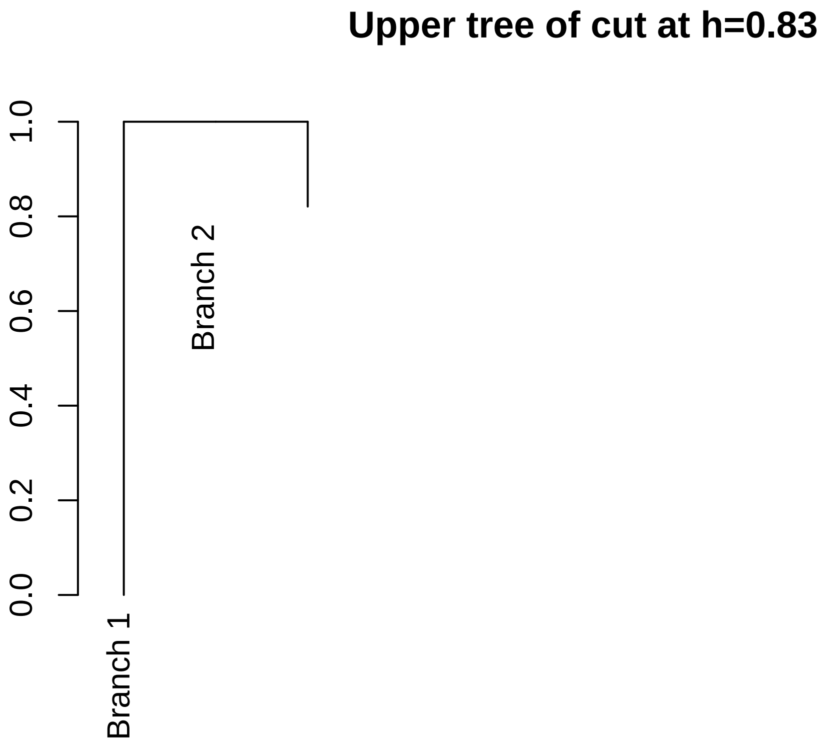
I did not choose single linkage as it joins clusters by the shortest link – poor at discerning between poorly separated clusters. This results in long string like clusters.

I decided to use average linkage as a good universal linkage.

Then to decide then how many clusters I looked for a relatively wide range of distances over which the number of clusters in the solution does not change.

In this case the biggest y value range of which the numbers of clusters did not change is for k=2 which is between 1.0 and 0.8. Although this is the biggest range it prob is not too useful as we only have 2 clusters. The next choice isn't as obvious so I decided to stick with the simple choice of k=2.

```r
load("~/Documents/edu/master/mulivariate_ana/asg2/2016_First6Votes_PresentAbsent.Rdata")
set.seed(123)
myvotes = votes-1
votes.hclust.average <- hclust(dist(myvotes,method='binary' ), method="average")
votes.hclust.complete <- hclust(dist(myvotes,method='binary' ), method="complete")
hcd <- as.dendrogram(votes.hclust.average)

hcl = cutree(votes.hclust.average, k = 2)
plot(cut(hcd, h = 0.83)$upper, main = "Upper tree of cut at h=0.83")
```



2 (b) Latent class analysis (LCA) can be thought of as a model-based approach to clustering when the recorded data are binary in nature. The 2011 Journal of Statistical Software paper poLCA: An R Package for Polytomous Variable Latent Class Analysis by Linzer and Lewis is posted on Blackbord. (Note that Linzer is famous as the 'stats man who predicted Obama's win'.) Read the Linzer and Lewis (2011) paper, in particular the examples, to gain some background to and understanding of Latent Class Analysis. (You can ignore the latent class regression parts of the paper.) Familiarise yourself with the poLCA function in R by reading its help file.

Use the poLCA function in R to cluster the TDs based on their voting data. Detail any decisions you make in the process. How many clusters of TDs do you think are present now? On what basis did you make this decision? Include any output or plots which you use to motivate your decision. [15 marks]

First I wanted to check the model fit for various numbers of classes. So I tried chooses number of classes in the range of 2:10 and recorded the nclass with the minimum aic and bic. This number of classes which minimized one or both of these values would be the best model fit.

To avoid local minima I set nrep =10. This will make sure we genuinely get a correct model fitting at the correct global maximum for log likelihood. I choose maxiter to be 3000 so we run the algorithm a maximum of 3000 time to try and converge.

By doing this my number of classes appears to be 4 or 5. The minimum bic was 4 and minimum aic was 5.

```r
#install.packages('poLCA')

library(poLCA)
```

```
## Loading required package: scatterplot3d
## Loading required package: MASS
```

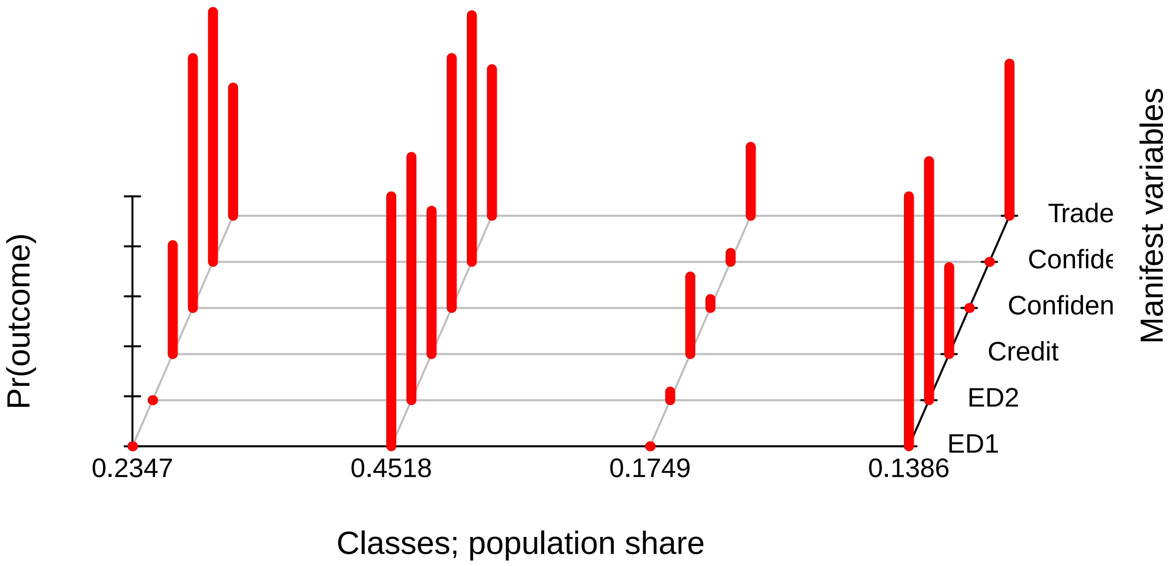
```r
f <- cbind(ED1, ED2, Credit, Confidence1, Confidence2, Trade) ~ 1


for (i in 2:6)
{
  fit <- poLCA(f, votes, nclass = i, maxiter = 3000, nrep = 10,verbose=FALSE)

  if (i ==  2){
    min_aic_index=2
    min_bic_index=2
    min_aic = fit$aic
    min_bic = fit$bic
  }
  if(min_aic >fit$aic ) {
    min_aic<- fit$aic
    min_aic_index <- i
  }
  if (min_bic >fit$bic ) {
    min_bic<- fit$bic
    min_bic_index <- i
  }

}
```

I chose my number of classes to be 4. So the next step was to look at the Predicted class memberships (by modal posterior prob.) versus the Estimated class population shares. If these values match it again suggests that we have a good modal fit. The Estimated class population corresponds to $\hat{p_r}$ in our modal which is the prior estimate based proportion of our data assigned to each class and our predicted class memberships are based on calculating the modal posterior probability $\hat{P}(r_i; Y_i)$.Congruence between these two sets of population shares often indicates a good fit of the model to the data.

Below we see a good matching for 4 classes.

Estimated class population shares			
0.4518	0.1386	0.1749	0.2347
Predicted class memberships (by modal posterior prob.)			
0.4518	0.1386	0.1747	0.2347

```r
#install.packages('poLCA')
set.seed(123)
lca_cl<- poLCA(f, votes, nclass = 4, maxiter = 3000, nrep = 10,graphs=TRUE,verbose=FALSE)
```



So looking at the graphs can analyze in many ways. For example We can see that for confidence votes almost all TDS in groups who have 14% and 18% of the TD population are absent but the other 2 groups are almost always present. For ED votes groups the 14% and 24% groups are almost always absent while other 2 groups are always present. The 45% group is the largest group of TDS whoattend most votes in high numbers. Credit votes have low turnout but appears that all groups have the simliar levels of turnout.

2 (c) Compare the resulting LCA clustering to the resulting clustering from part (a).

Our method here will be to create a table of our results which has a single vector which will basically associate each TD with a class.

```r
library(e1071)
library('mclust')
```

```
## Warning: package 'mclust' was built under R version 3.2.4
```

```
## Package 'mclust' version 5.2
## Type 'citation("mclust")' for citing this R package in publications.
```

```
# compare the cuttree at k=2 to the k=4 LCA
tab <- table(as.vector(hcl),lca_cl$predclass)
classAgreement(tab)
```

```
## $diag
## [1] 0.2349398
##
## $kappa
## [1] -0.02628761
##
## $rand
## [1] 0.4395765
##
## $crand
## [1] 0.111278
```

```
adjustedRandIndex(as.vector(hcl),lca_cl$predclass)
```

```
## [1] 0.111278
```

First we can see our rand index was 0.4395. A high agreement value would be 1 so our result is not very high and suggests general disagreement between the clusters.

We also should look at adjusted rand index which will take into account agreement due to chance which inflates the rand index value.

In this case our adjusted rand index is 0.111278 which is small and suggests there is a lot of disagreement between the clusters and our rand index result was indeed inflated due to chance. Even then the rand index was small so our conclusion here is there is strong disagreement between the clusters.

Mining Dail Attendance Records Suggests Strategic Absenteeism

Normally, bedtime reading of the Dail Attendance records would quicken the pulse of few more than Vinny Browne, however, applying a data mining technique known as Latent Cluster Analysis to said records, illuminates a conspiracy even Dan Brown would be proud of.

The freemasonry at play here involves attendance patterns of the TDs for specific votes between January 14th 2016 and January 21st 2016. We used data mining to ask ourselves the question: "Do TDs strategically miss or attend specific votes?" Now, fear not dear readers, that we suggest that the Shinners only turn up at the sound of 1 tap dripping or that FF were all nursing themselves after the Christmas drink-in. No, no, it appears in fact that TDs actually think (shock horror) for themselves whether to turn up or not independent of party affiliation.

For those of you wise enough to puzzle through Brother William Of Baskerville's aedificium I can provide some detail on the algorithm of choice. Latent Class Analysis, as it's title suggests, is a statistical technique for the analysis of multivariate categorical data. It is an unsupervised clustering algorithm that can be applied to classify sample data into a specified number of classes. It does not automatically determine the number of latent classes but can tell us for a specified number of classes how well the model fits the data. Basically, observations with similar sets of responses will tend to cluster within the same latent classes.

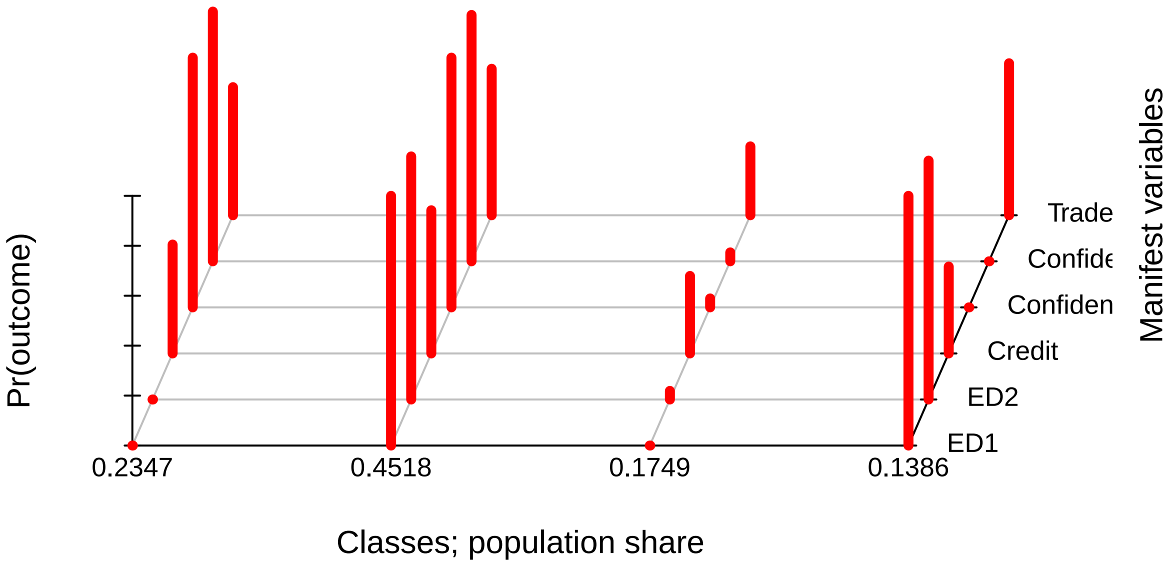
If we have J outcome variables and R latent classes then for a categorical variable j in a particular latent class r, the probability of outcome k will be π_{rjk}. This is the class-conditional probability where within each class, within each variable, $\sum_{k=1}^{K} \pi_{rjk} = 1$. We use p_r to denote the proportion of population in class r where $\sum_{r=1}^{R} p_r = 1$ and we refer to this as the prior probability of an observation i being in latent class r. Denote Y_ijk as the observed values of the J manifest variables such that $Y_{ijk} = 1$ if oberservation i gives the kth response to the jth variable, and $Y_{ijk} = 0$ otherwise, where j = 1. . . J and k = 1. . . Kj.

The parameters estimated by the latent class model are p_r and π_{rjk}.So given $\hat{p_r}$ and $\hat{\pi_{rjk}}$, the posterior probability that each individual belongs to each class, conditional on the observed values of the manifest variables can be calculated using Bayes' formula:

$\hat{P}(r|Y_i) = \frac{\hat{p_r} f(Y_i;\hat{\pi_r})}{\sum_{q=1}^{R} \hat{p_q} f(Y_i;\hat{\pi_q})}$

where $f(Y_i;\pi_r)$ is the probability that an individual i in class r produces a particular set of J outcomes on the manifest variables, assuming conditional independence of the outcomes Y given class memberships.

So applying this knowledge to our TDs voting data Let's make a plot of the latent classes.



We determined that 4 classes is the best fit to the data and we can see the $\hat{p_r}$ values of each latent class on the X-axis. So For example the 0.4518 latent class shows 45% of all TDs in the Dail and 0.1386 group shows 14% of all TDS in the Dail. The height of the bar is the $\hat{pi_{rjk}}$ and in this graph indicates the probability of attendance for the TDs in that that latent class. Finally the Z-axis shows each Vote type. So for example we can see that the 14% latent class has extremely low probability of turning up to Confidence1 votes. In fact it is close to zero.

The most striking observation is that TDS almost always to miss or attend **both** ED (Emergency Department) votes and they almost always miss or attend **both** Confidence votes. Surely if attendance was random then TDs would miss, for example, a Confidence1 vote but then attend a Confidence2 vote. From all 166 TDS only 3 TDs attended just 1 of the Confidence Votes and only 4 TDs attended just 1 of the ED votes.

All the other TDs **chose** to attend both votes, perhaps due to interest in the topic, or they **chose** to miss both, perhaps due to lack of interest in the topic or indeed could it be they have a personal view on the topic that disagrees with party HQ. In that case perhaps they cannot bring themselves to vote in line with party HQ and so would rather abstain.

Analyzing each latent class individually then.

- Class 1 (24% of all TDs) never attend ED votes but always attend Confidence votes. Trade and Credit votes are less extreme. There seems about 30% probability of this group attending Credit votes but 50% attending Trade.
- Class 2 (45% of all TDs) almost always attend ED votes and Confidence votes. But once again, Trade and Credit votes are less extreme. There seems about 50% probability of this group attending Trade or Credit votes.
- Class 3 (18% of all TDs) almost never attend ED nor Confidence votes. Yet again Trade and Credit votes are less extreme. There seems about 25% probability of this group attending Trade or Credit votes.
- Class 4 (14% of all TDs) always attend ED votes but never attend Confidence votes. But once again, Trade and Credit votes are less extreme. There seems about 50% probability of this group attending Trade and 25% probability of attending Credit votes. . . .

So if there was a Confidence vote scheduled for next week our latent classes predict everybody from Class 1 and Class 2 would turn up and nobody from Class 3 and Class 4 would turn up. The results seem too extreme to ignore.

Also when we look at party information we see that each class contains a mix of parties. This would suggest that missing votes is not a party strategy but rather a personal one. We can see this from a count of people in Class 4. This class misses Confidence votes. It includes the following TD counts: 3 Labour; 6 Fianna Fail; 7 Fine Gael; 4 Independents; 1 Sinn Fein.

Ok so perhaps Sinn Fein are the party which seems to buck the trend by having almost all members attend both Confidence and ED votes. Could that suggest the desire to agree with the party line. A certain zealotry even. But with only 14 TDs that coherence may change with bigger numbers.

Overall we can say with high confidence that most TDs are choosing to be absent from certain votes independent of what party. Perhaps this suggests that people who choose to be absent from a certain vote have more in common with each other than their own parties. For anyone interested in a forming a new party perhaps LCA clustering can be a start on who to talk with.