Assignment 3 - Part I

Matthew Dunne

July 15, 2018

Data: Choosing Variables

We will load the data, take out the variables not suitable to LCA (continuous variables), create a new column so as to make all variables readable for the poLCA function, and then choose which variables to use in our model.

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

```
## Loading required package: ggplot2
```

The variables we base the model upon are:

-**Payment Status of Previous Credit**: (1=no credits taken, 2=all paid back duly, 3=existing credits paid duly until now, 4=past delay in paying, 5=critical account)
-**Occupation**: (1=unemployed/unskilled-non-resident, 2=unskilled-resident, 3=skilled employee/official, 4=management/self-employed/highly qualified employee/officer)
-**Purpose**: (1=car (new), 2=car (used), 3=furniture/equipment, 4=radio/television, 5=domestic appliances, 6=repairs, 7=education, 8=(vacation - does not exist?), 9=retraining, 10=business, 11=others)

Running LCA Models

We will run several LCA models, using 2 to 6 latent classes, and save the AIC and BIC of each in order to plot them.

```
require(poLCA)
```

```
## Loading required package: poLCA
```

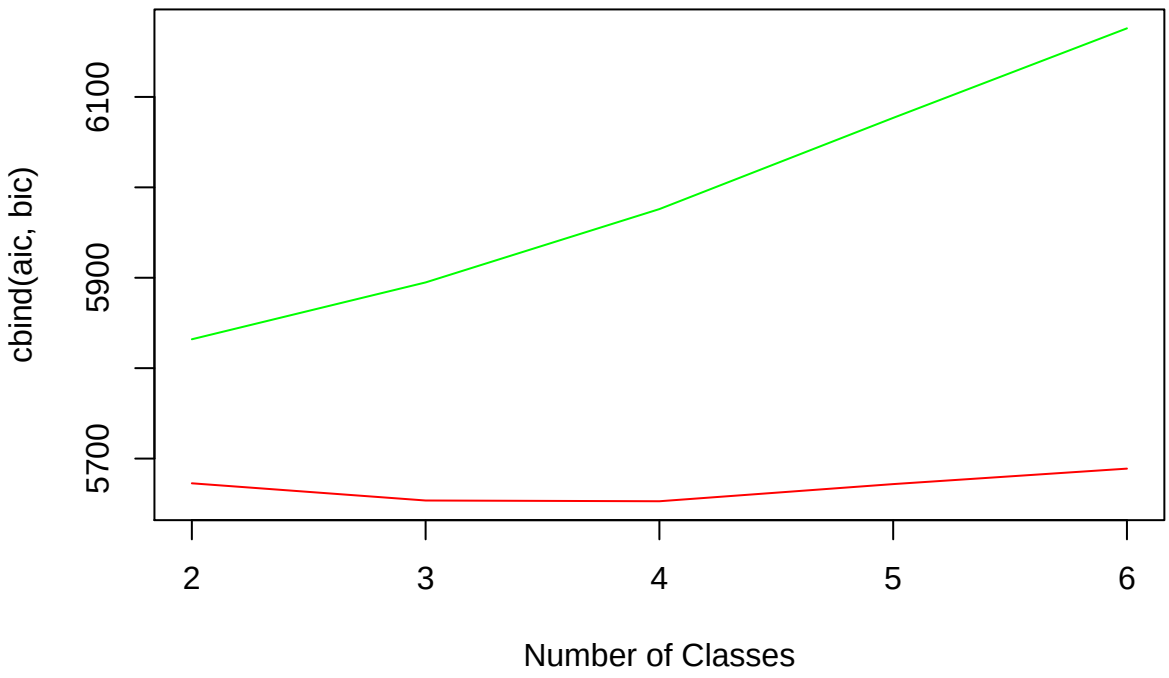
```
## Loading required package: scatterplot3d
```

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

```
clusters<-2:6
aic<-c()
bic<-c()
set.seed(1234)
for (num in clusters){
  results=poLCA(f1,train,nclass=num,nrep=10,tol=.001,verbose=FALSE)
  aic<-append(aic, results$aic)
  bic<-append(bic, results$bic)
}
```

When we plot the AIC and BIC there is a "kink" in the AIC (red) at 3 and 4 as well as in the BIC (green), albeit more pronounced at 3 than at 4.

```
indices<-c(2,3, 4, 5,6)
matplot(indices, cbind(aic,bic),type="l",col=c("red","green"),lty=c(1,1), xlab = "Number of Classes")
```



So the choice for number of classes in our model is between three and four. The overall BIC is lower for the three class model so we will choose that one.

```
set.seed(1234)
model=poLCA(f1,train,nclass=3,nrep=10,tol=.001,verbose=FALSE)
```

Testing the Chosen Model

We will perform the holdout validation by entering the probabilities from our chosen model into a LCA model with the test data, using the same number of classes.

```
set.seed(1234)
validation=poLCA(f1,test,nclass=3,nrep=10,tol=.001,verbose=FALSE, probs.start = model$probs)
```

Does the model work well? A good LCA model is a stable model, meaning the model should perform similarly on train and test data.

We will look at two criteria: (1) similarity of relative class sizes and (2) stability of conditional probabilities.

Similarity of Class Sizes

We can get the class sizes (proportions) with $P.

For the training data, the probabilities are:

```
model$P
```

```
## [1] 0.79824985 0.12832237 0.07342779
```

and on the test data:

```
validation$P
```

```
## [1] 0.3927997 0.4403771 0.1668232
```

As we can see, these are not very stable in terms of class size/proportions.

Note: AIC and BIC are not for comparing train to test and validating results. They are used in choosing your model when you are building it on the training data.

Stability of Conditional Probabilities

We can get the estimated class-conditional response probabilities with $probs.

```
model$probs
```

```
## $Payment.Status.of.Previous.Credit
##                       1          2          3            4            5
## class 1:  7.231132e-07 0.04075148 0.61640890 5.305450e-03 0.3375334475
## class 2:  3.331966e-01 0.03451260 0.01931076 6.126919e-01 0.0002882214
## class 3:  4.027273e-02 0.13869756 0.48310987 5.016605e-18 0.3379198434
##
## $Occupation
##                       1            2            3          4
## class 1:  1.382242e-02 2.445432e-01 6.594945e-01 0.08213987
## class 2:  8.057277e-37 1.709465e-01 6.956926e-01 0.13336094
## class 3:  2.193871e-01 1.842447e-23 5.627007e-18 0.78061286
##
## $Purpose
##               Pr(1)      Pr(2)        Pr(3)        Pr(4)         Pr(5)
## class 1:  0.2280533 0.08491099 1.999708e-01 3.262237e-01  1.641040e-02
## class 2:  0.1909503 0.06624901 1.253687e-01 1.860756e-01 7.367782e-161
## class 3:  0.2999510 0.38138559 1.434276e-17 3.200867e-13  1.615347e-02
##                 Pr(6)      Pr(7) Pr(8)       Pr(9)     Pr(10)
## class 1:  0.015240517 0.05871602     0 0.005782159 0.06469206
## class 2:  0.044749342 0.03241336     0 0.013484207 0.32767406
## class 3:  0.009034006 0.04434695     0 0.010853168 0.06650194
##                  Pr(11)
## class 1:  2.397523e-245
## class 2:   1.303544e-02
## class 3:   1.717739e-01
```

and on the test data:

```
validation$probs
```

```
## $Payment.Status.of.Previous.Credit
##                       1            2         3            4         5
## class 1:  5.966383e-74 3.203679e-02 0.7212611 3.333654e-02 0.2133655
## class 2:  1.207397e-11 9.253272e-02 0.3884801 1.973271e-01 0.3216601
## class 3:  1.598499e-01 9.011407e-16 0.4532408 4.256482e-05 0.3868667
##
## $Occupation
##                       1          2         3            4
## class 1:  1.203841e-67 0.21425474 0.7856574 8.789125e-05
## class 2:  4.766435e-50 0.05102536 0.5706079 3.783668e-01
## class 3:  5.994371e-02 0.31992199 0.6200858 4.848499e-05
##
## $Purpose
```

```
##                Pr(1)        Pr(2)        Pr(3)        Pr(4)        Pr(5)
## class 1:  0.08395241 5.247206e-07 2.857241e-01 5.335539e-01 1.697218e-02
## class 2:  0.25402737 2.041262e-01 1.841612e-01 1.371998e-01 1.990104e-38
## class 3:  0.61036094 6.058678e-02 3.596583e-06 3.147118e-06 3.005696e-29
##                 Pr(6)        Pr(7) Pr(8)        Pr(9)       Pr(10)
## class 1:  2.634017e-04 3.749424e-02     0 3.387483e-02 0.0009369964
## class 2:  4.324511e-33 5.738733e-02     0 1.333330e-20 0.1544160175
## class 3:  1.792109e-01 1.242946e-06     0 1.637010e-04 0.1496435047
##                Pr(11)
## class 1:  7.227456e-03
## class 2:  8.682021e-03
## class 3:  2.614075e-05
```

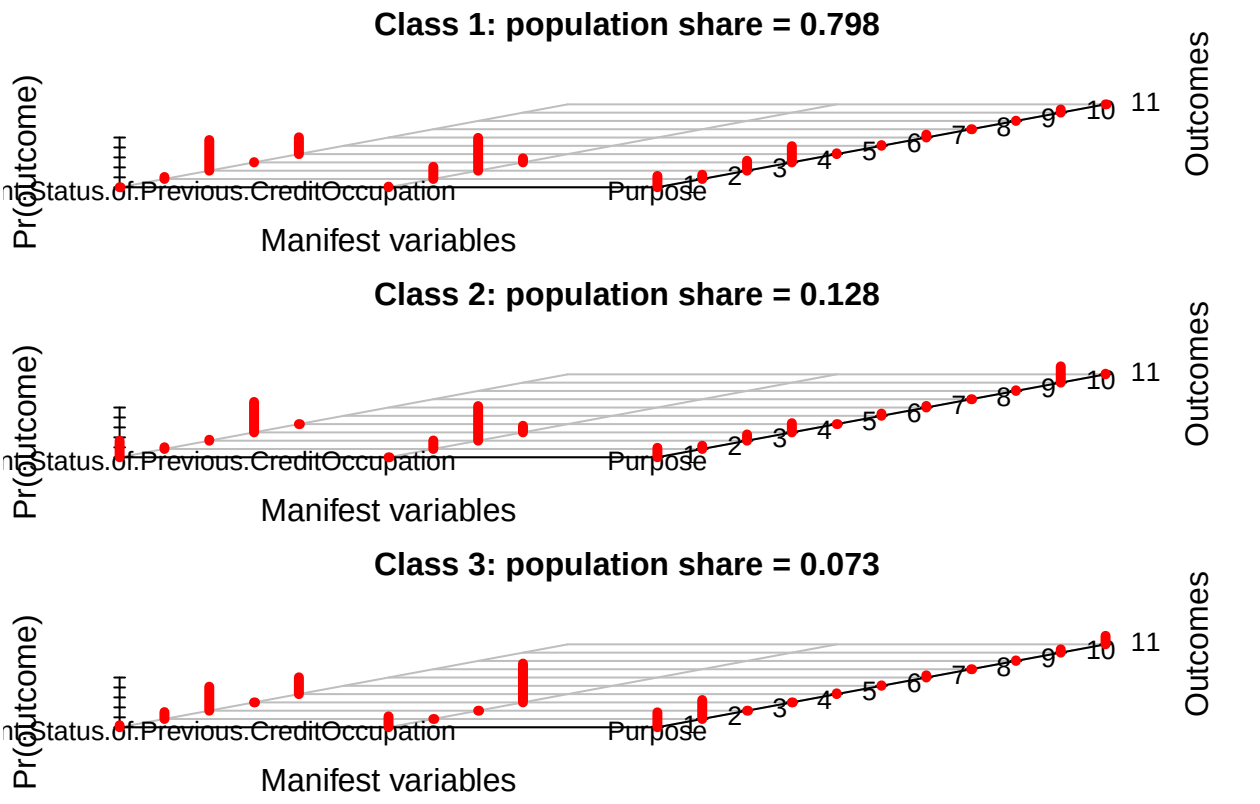
The conditional class probabilities are stable from train to test only for some combinations of Class and variable (e.g. Class 3 and Payment Status of Previous Credit). But overall the probabilities do not appear stable. These probabilities are plotted in the next section.

Interpreting the Model and Similarity to K-Means Clustering Solution

Interpreting the Model

Let us look at the probabilities for the model based on the training data:

```
plot(model)
```



One might describe the classes as follows:

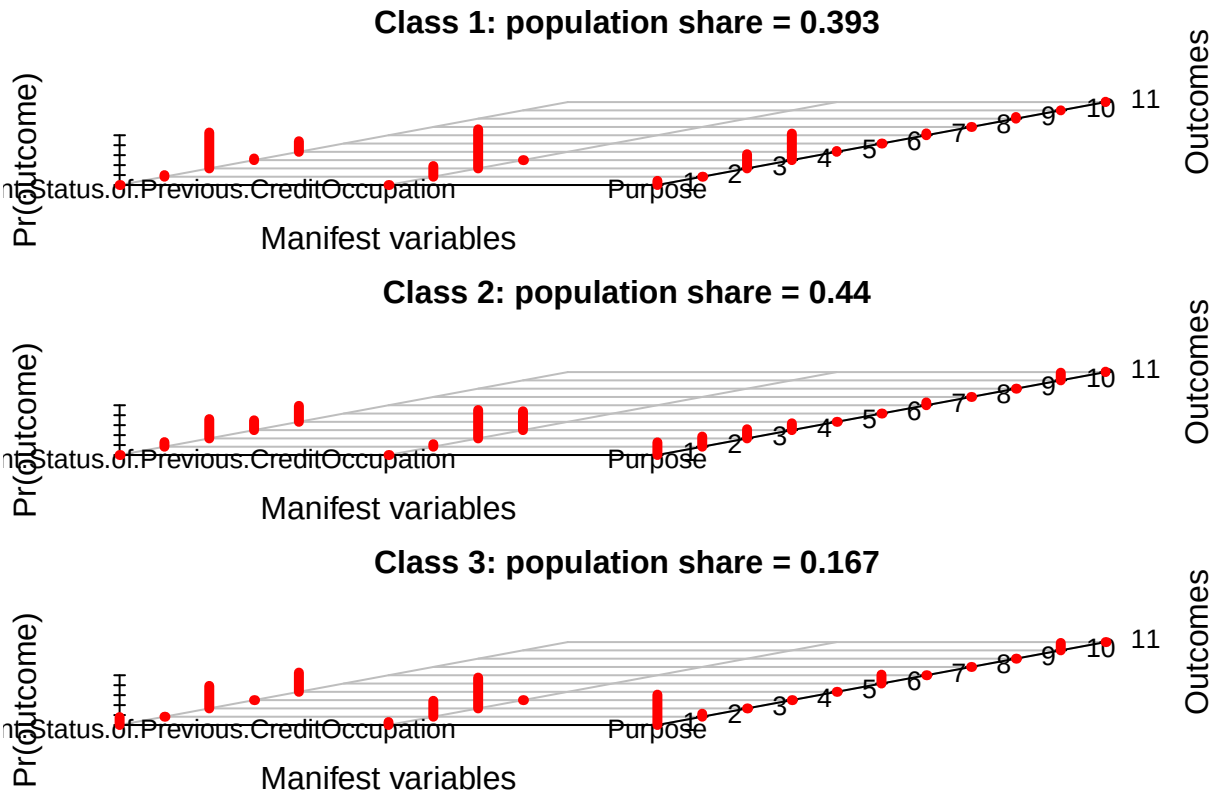
Class 1: (1) skilled employees (2) with either fair or terrible credit history (3) taking out loans predominantly for consumer goods

Class 2: (1) skilled employees (2) with spotty credit history (3) taking out business loans and some consumer goods

Class 3: (1) management/self-employed/highly qualified employee/officer (and some unemployed) (2) with varying credit history (3) taking out a loan for cars and other purposes

And let us look at the probabilities derived from the test data:

```
plot(validation)
```



The classes as applied to the test data have a somewhat different delineation. Class 1 is similar to Class 1 the the training data with better credit history and loans geared more toward furniture and electronics. Class 2 has little discernible pattern. Class 3 is clearly for car loans.

Comparison to the K-Means Solution in Assignment 1

We used a K-means model in Assignment 1 and so used continuous variables (Duration, Age, Amount) instead of categorical variables (Credit History, Occupation, Age). This makes it somewhat difficult to compare how these models classify borrowers as we are using different standards to do it. I see little overlap between the K-means and LCA analysis, especially as it is difficult to see a correlation between the clusters/classifications, i.e. older borrowers does not necessarily mean skilled employees and business loans do not necessarily equate to high amount loans.

The K-means model we used was far more consistent from training to test data than our LCA model is here.