Pset3-1

yunzhi wang

1/24/2018

For this assignment, I intend to do this analysis for the bank to detect people who are not creditworthy. I assume the categorical variables that are important to my analysis is Purpose, saving accounts/bonds, personal status/sex, housing, occupation.

As I load the raw data, I have to set the categotical variables into 1,2,3,4, excluding 0, in case of the disruption to LCA analysis. Furthermore, from the categorical variables we can find out that only purpose has 0 inside. Other categotical variables have numeric numbers starting from 1.

```
mydata.cate <- as.data.frame(mydata[,c(5,7,10,16,18)])
#mydata.cate <- data.frame(lapply(mydata[,c(5,7,10,16,18)], as.character),
stringsAsFactors=TRUE)
mydata.cate[,1] <- as.numeric(mydata.cate[,1]) +1
head(mydata.cate)
```

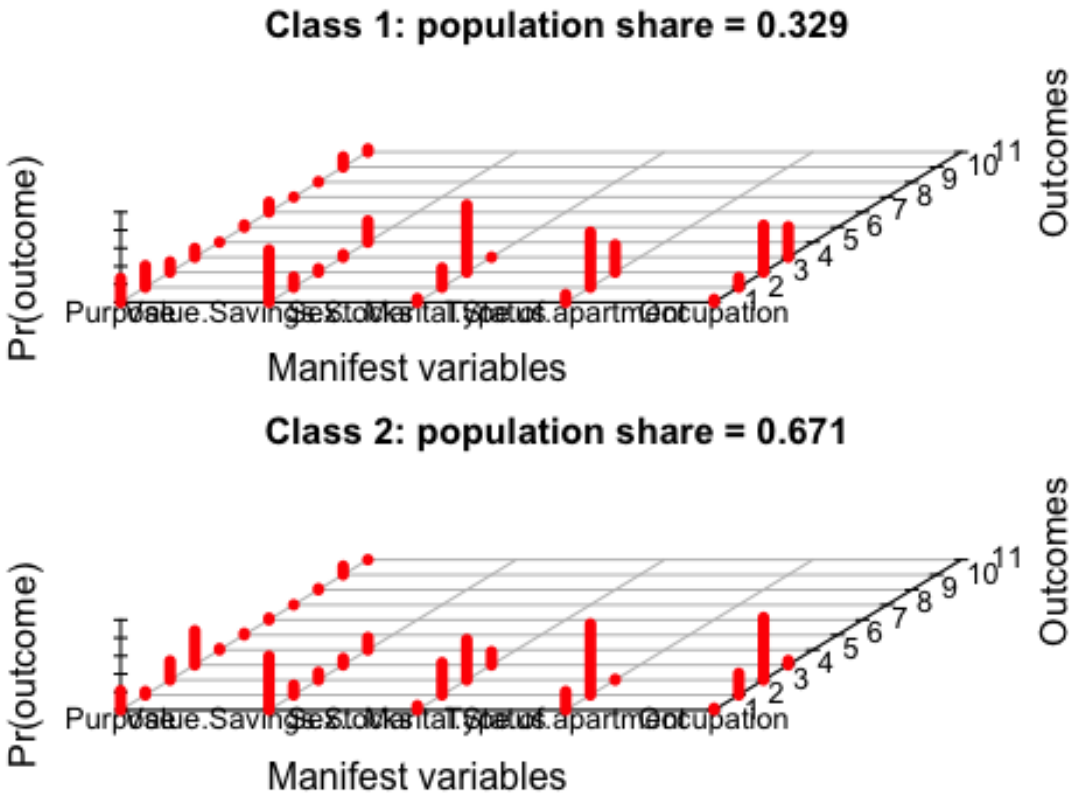
```
##   Purpose Value.Savings.Stocks Sex...Marital.Status Type.of.apartment
## 1       3                    1                    2                 1
## 2       1                    1                    3                 1
## 3      10                    2                    2                 1
## 4       1                    1                    3                 1
## 5       1                    1                    3                 2
## 6       1                    1                    3                 1
##   Occupation
## 1          3
## 2          3
## 3          2
## 4          2
## 5          2
## 6          2
```

```
#split the dataset into training and test parts
data.split <- sample(1:nrow(mydata.cate),size = 0.632 * nrow(mydata.cate))
data.train <- mydata.cate[data.split,]
data.test <- mydata.cate[-data.split,]
head(data.train)
```

```
##     Purpose Value.Savings.Stocks Sex...Marital.Status Type.of.apartment
## 414      10                    3                    3                 2
## 442      10                    1                    3                 2
## 411       4                    2                    2                 2
## 993       2                    5                    3                 3
## 753       7                    4                    3                 3
```

```
## 265       4                    1                    2                 2
##     Occupation
## 414          3
## 442          3
## 411          3
## 993          3
## 753          3
## 265          3
```

2.1 determine 2,3,...k clusters/solutions class = 2




```
## [1] 7360.72
```

Interpretation:Class1 shares 78.3% of population and class2 accounts for 21.7%.

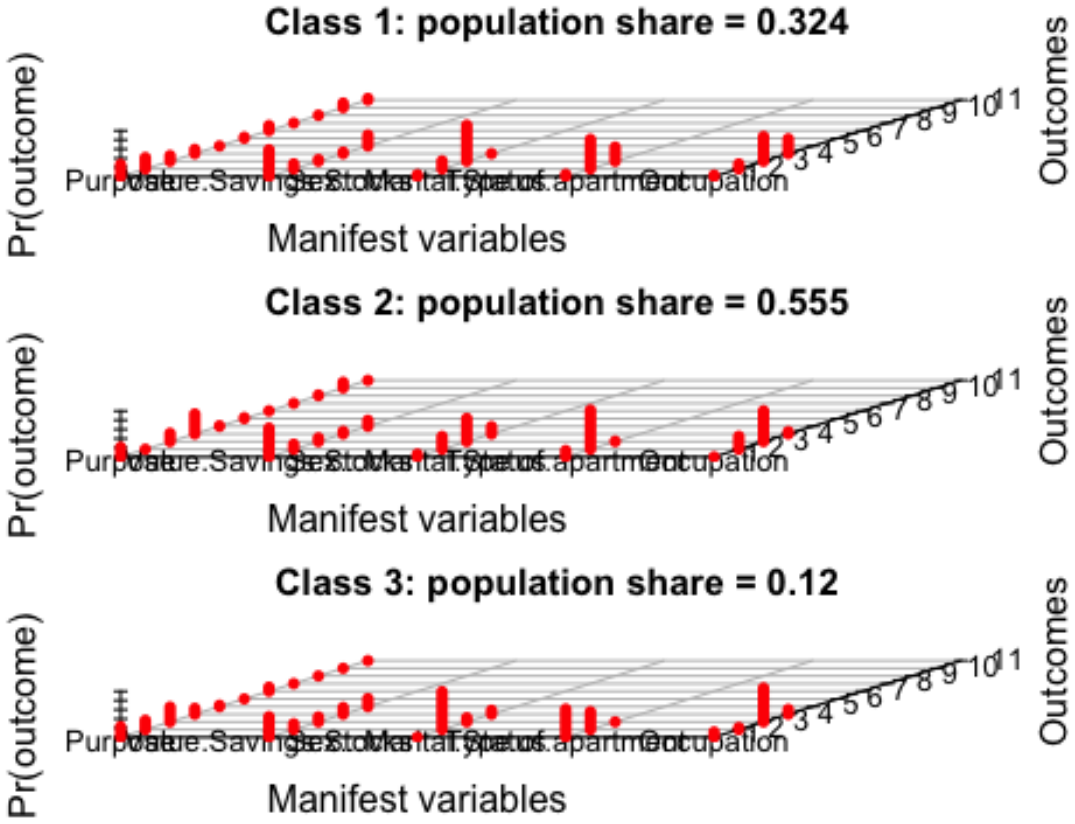
For class1, customers are highly likely to be not-single-female or single male employees equipped with some skills. They have high probability of buying new cars,furnitures and home appliances and owning their residence.

For class2, customers are very likely to be single male professionals with high titles (skilled and highly-qualified employee). They have high prbability of buying cars and having their own apartments.

This model has AIC value = 7366.658, which is a little bit large. Further decisions will be made when looking at other results.

class size = 3

```
results.3 =
poLCA(f1,data.train,nclass=3,maxiter=1000,nrep=500,tol=.001,verbose=FALSE,
graphs=TRUE)
```




```
results.3$aic

## [1] 7334.049
```

Interpretation: Class 1 share 13% popultion, class 2 shares 61.3% population, class 3 share 25.6 population.

For class1, the customers are mostly not-single-female with high ranks and skills, who own their apartments. They have high probability of buying cars.

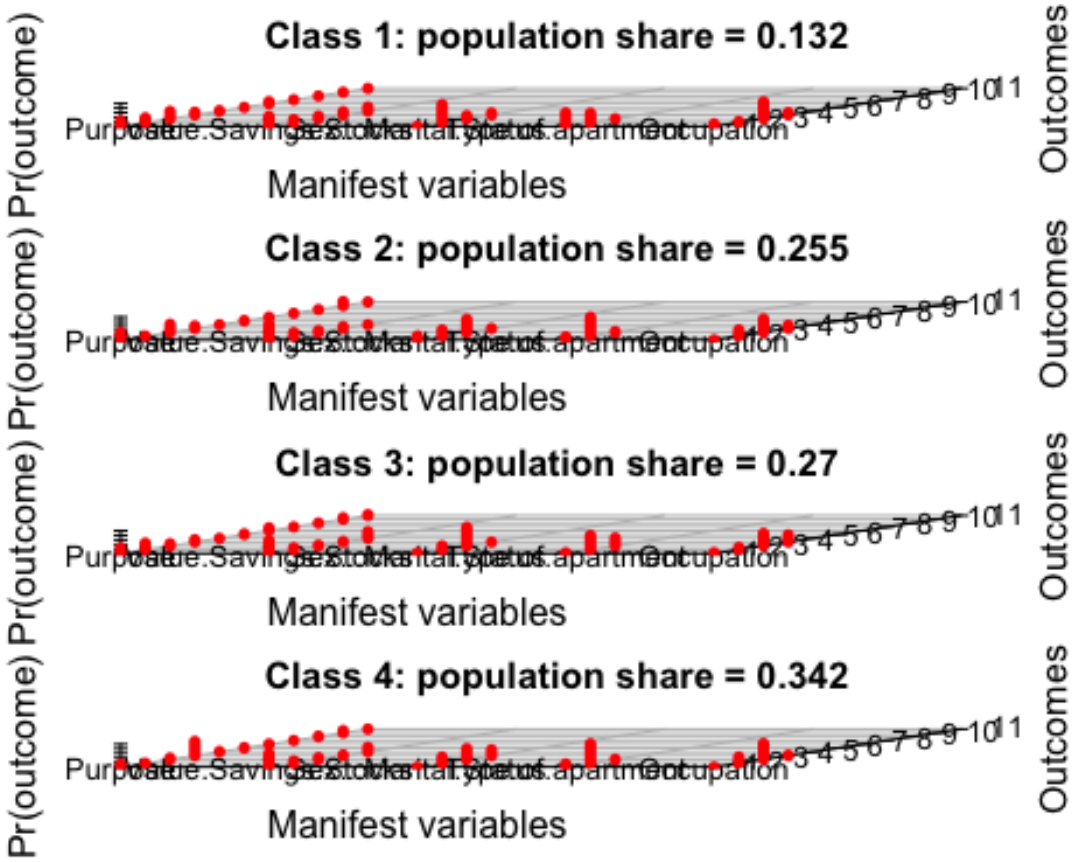
For class2, the customers are skilled single male, who have their own apartments. They have high probability of buying cars and radio/televisions.

As for Class3, the customers in this class are mostly single male, who have their apartments and skilled/high-ranking. They have high probabiliyt of buying cars.

We can tell that the class2 and class3 nearly share the same composition of population, which shows that there is no need to split them. The AIC value = 7351.192, which is also pretty similar to the two-class model showed above.

class size = 4

```
results.4 = poLCA(f1,data.train,nclass=4,nrep=500,tol=.001,verbose=FALSE,
graphs=TRUE, maxiter = 1000)
```




```
results.4$aic

## [1] 7332.578

results.4$probs

## $Purpose
##                Pr(1)        Pr(2)      Pr(3)     Pr(4)        Pr(5)
## class 1:  0.2432468 2.050483e-01 0.32330709 0.1182654 2.496414e-02
## class 2:  0.3296029 7.392170e-07 0.34972284 0.1132140 2.812923e-43
## class 3:  0.2514041 2.861385e-01 0.09318191 0.1046263 1.217642e-51
## class 4:  0.1367484 3.234149e-28 0.06324489 0.6237444 1.808737e-02
##                  Pr(6)        Pr(7) Pr(8)        Pr(9)       Pr(10)
## class 1:  2.083356e-31 8.516830e-02     0 1.284054e-39 3.256322e-22
## class 2:  1.736252e-02 2.777467e-10     0 1.355366e-97 1.862744e-01
```

```
## class 3:  2.641474e-02 1.250375e-01     0 5.668209e-03 7.015285e-02
## class 4:  2.166499e-02 1.168798e-02     0 2.324964e-02 1.015723e-01
##                 Pr(11)
## class 1:  2.397235e-151
## class 2:   3.822533e-03
## class 3:   3.737594e-02
## class 4:   5.281362e-65
##
## $Value.Savings.Stocks
##               Pr(1)      Pr(2)      Pr(3)      Pr(4)        Pr(5)
## class 1:  0.4368870 0.13098512 0.10924890 0.14287480 1.800042e-01
## class 2:  0.6968632 0.14863069 0.07224772 0.08225810 3.259343e-07
## class 3:  0.5687862 0.08252390 0.03149155 0.02820120 2.889971e-01
## class 4:  0.5913989 0.09702103 0.08245680 0.03227389 1.968493e-01
##
## $Sex...Marital.Status
##                 Pr(1)     Pr(2)     Pr(3)        Pr(4)
## class 1:  6.800153e-54 0.7846291 0.1323844 8.298654e-02
## class 2:  1.614080e-01 0.2696412 0.5689508 5.284326e-08
## class 3:  1.741529e-02 0.1652894 0.8172953 3.788425e-52
## class 4:  1.028700e-20 0.2714445 0.4741175 2.544380e-01
##
## $Type.of.apartment
##                Pr(1)     Pr(2)        Pr(3)
## class 1:  0.59188882 0.4081112 5.961640e-28
## class 2:  0.13548918 0.8353930 2.911783e-02
## class 3:  0.02800841 0.6250670 3.469246e-01
## class 4:  0.13365442 0.8568655 9.480122e-03
##
## $Occupation
##                 Pr(1)     Pr(2)     Pr(3)      Pr(4)
## class 1:  1.065887e-01 0.0436713 0.7458173 0.10392271
## class 2:  3.848837e-16 0.2718630 0.5837834 0.14435356
## class 3:  3.569263e-02 0.1000322 0.5490953 0.31517988
## class 4:  8.268439e-83 0.2746005 0.7013187 0.02408088
```

Interpretation: Class1 share 32.8% popultion, class 2 shares 16.1% population, class3 share 36.2% population while class 4 shares 14.9%.

In class1, the customers are single skilled male, who have their own apartments. They have high probabilities of buying new cars and furnitures. It's hard to tell their occupation.

Class2, the customers are single male, who have their apartments. They have high probabilities of buying used cars.It's hard to tell their occupation.

Class3, the customers are skilled single male, who have their apartments. They have high probabilities of buying radio/televisions.

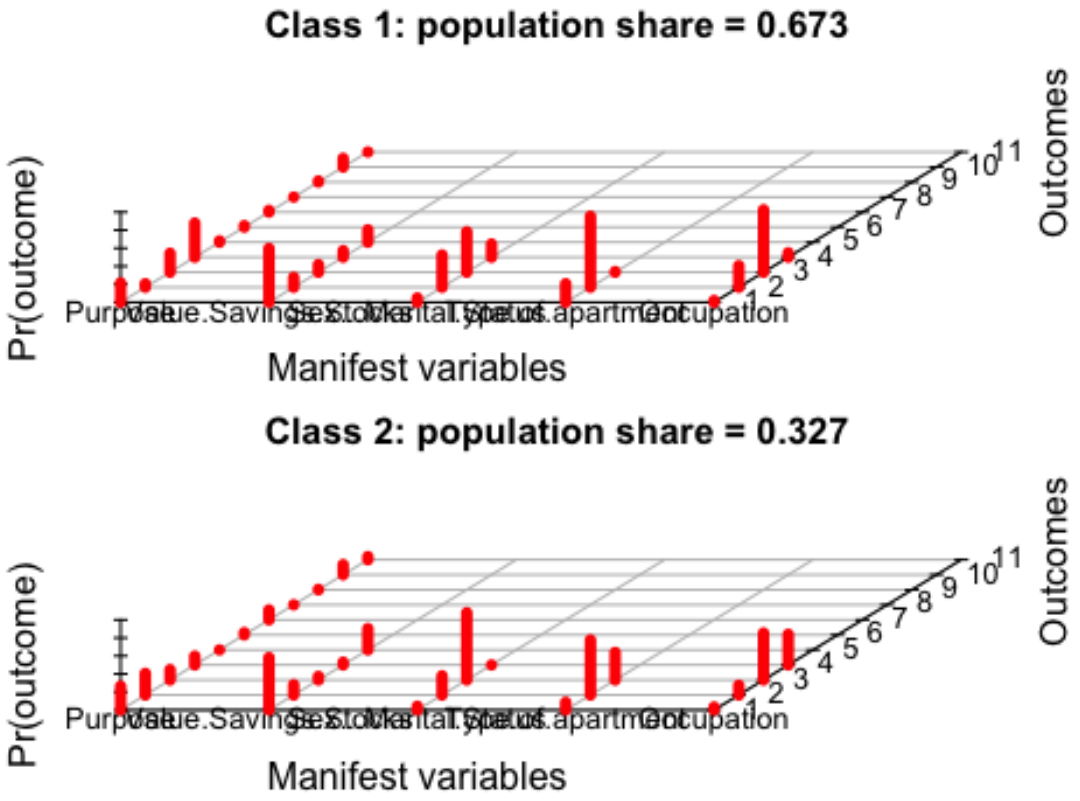
Class4, the customers are skilled not-single-female who rent their places. They have high probability of buying furniture and radio/televisions.

The AIC value =7350.637, which is the lowest among models but still pretty close to the previous models.

All in all, the saving accounts/ bonds doesn't differ from class to class. The seperation into 3 and 4 classes doesn't bring more meaningful insight or model accuracy enhancement. To mantain model simplicity, 2-class LCA model would be chosen to divide the market segments, and I will also use this two-class model to validate the testing dataset.

3. Perform Holdout validation of LCA

```
results.valid <- poLCA(f1, data.train, nclass=2, maxiter=1000,
graphs=TRUE,nrep=500,tol=1e-10, na.rm=TRUE,verbose=FALSE,
probs.start=results.2$probs)
```




```
results.valid$aic

## [1] 7360.717

results.valid$probs

## $Purpose
##               Pr(1)      Pr(2)     Pr(3)      Pr(4)        Pr(5)
## class 1:  0.2148399 0.03871375 0.2100205 0.37848701 1.411541e-02
## class 2:  0.2642336 0.23942106 0.1146612 0.09721786 2.178266e-26
##                Pr(6)      Pr(7) Pr(8)       Pr(9)     Pr(10)       Pr(11)
```

```
## class 1:  0.01629560 0.01947887     0 0.011987359 0.09606158 2.713185e-18
## class 2:  0.02451653 0.10979497     0 0.004371303 0.11195603 3.382742e-02
##
## $Value.Savings.Stocks
##               Pr(1)     Pr(2)      Pr(3)      Pr(4)     Pr(5)
## class 1:  0.5968774 0.1093726 0.08454520 0.07246320 0.1367416
## class 2:  0.5812853 0.1136083 0.03896242 0.02995304 0.2361909
##
## $Sex...Marital.Status
##                Pr(1)     Pr(2)     Pr(3)        Pr(4)
## class 1:  0.04784899 0.3576082 0.4486836 1.458593e-01
## class 2:  0.04185400 0.2125930 0.7455530 2.941624e-34
##
## $Type.of.apartment
##                Pr(1)     Pr(2)       Pr(3)
## class 1:  0.20536533 0.7876451 0.006989609
## class 2:  0.08556364 0.6098497 0.304586662
##
## $Occupation
##                Pr(1)     Pr(2)     Pr(3)      Pr(4)
## class 1:  0.01935168 0.2393475 0.6923863 0.04891455
## class 2:  0.03273640 0.1075771 0.5204070 0.33927947
```

Interpretation:

Class1 shares 21.6% of population and class2 shares 78.4%.

In class1, customers are highly likely to be single skilled and high-ranking male, who have high prbability of buying cars and owning their residence.

For class2, customers are highly likely to be not-single-female or single male employees with skills. They have high probability of buying new cars and home appliances and owning their apartments.

The AIC value = 7366.65, which is the same as the training model.

Comparison between the train and holdout model results of two-class lca models: The training model predicted the validation dataset well. There are almost no differences in class sizes (21.9% and 78.1% for train, 21.6% and 78.2% for validate). And both train and holdout have the same segmentation.

4. Comment on goodness, stability, interprebility and adequacy of model solutions.

Despite having different class sizes, they have approximately the same maximum loglikelihood, AIC and BIC criterion, the three models have equal goodness of fit.

As for interpretibility, since the increase of class size does not generate meaniingful insights about market segmentation, therefore LCA model with 2 classes is chosen as our final model for holdout.

As I compare the train and holdout model results of two-class LCA models, the class sizes and segmentation is the same, which built up the stability of my model.

As all the models above have large AIC and BIC creterion, I speculate these classifications are not optimal. To improve the market segmentation with more sense, probably more features should be involved.

5. Compare K-means solution in Asisgnment 2 with LCA solution.

In LCA model, I used categorical features, and I classified the population into two classes by personal status/sex, occupation and purposes. And one class has doubled the size of the other class. Furthermore, the AIC does not indicate a good classification and more models need to be considered to find better solutions.

For K-means solution, only numerical features are used. I classified popultion into three classes based on the age and durations. The K-means and K-o-means give me a more evenly splitted class sizes than LCA models.