HW9_lda_Yuefei_Chen

Yuefei Chen

2024-04-24

Model development

Running the following code, we build a linear discriminant analysis model to classify rent house data. Its explanatory variables "area", "rooms", "bathroom", "parking spaces", "hoa", "property tax", "fire insurance". The predictor variable is "furniture". This model attempts to find the best linear combination to distinguish between different groups of furniture.

```
library(MASS)
library(ggplot2)
library(memisc)
```

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

```
##
## Attaching package: 'memisc'
```

```
## The following object is masked from 'package:ggplot2':
##
##     syms
```

```
## The following objects are masked from 'package:stats':
##
##     contr.sum, contr.treatment, contrasts
```

```
## The following object is masked from 'package:base':
##
##     as.array
```

```
library(ROCR)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:memisc':
##
##     collect, recode, rename, syms
## The following object is masked from ’package:MASS’:
##
##     select

## The following objects are masked from ’package:stats’:
##
##     filter, lag

## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```
library(klaR)
library(readr)
house_data <- read_csv("Dataset/Rent_House_random_200_multi_regression.csv")
```

```
## Rows: 200 Columns: 11

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): floor, animal, furniture
## dbl (8): area, rooms, bathroom, parking_spaces, hoa, rent_amount, property_t...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
house_data <- house_data[, c(1:4, 6:11)]
house_data <- house_data[,-c(5)]
str(house_data)
```

```
## tibble [200 x 9] (S3: tbl_df/tbl/data.frame)
##  $ area          : num [1:200] 120 45 50 35 204 177 15 70 180 180 ...
##  $ rooms         : num [1:200] 3 1 2 1 4 3 1 2 3 4 ...
##  $ bathroom      : num [1:200] 4 1 1 1 4 3 1 2 3 4 ...
##  $ parking_spaces: num [1:200] 3 1 1 0 2 4 0 1 2 2 ...
##  $ furniture     : chr [1:200] "not furnished" "furnished" "not furnished" "not furnished" ...
##  $ hoa           : num [1:200] 1350 3000 226 260 0 2700 0 1800 700 2600 ...
##  $ rent_amount   : num [1:200] 5600 5520 750 1400 3440 6900 1200 4200 2700 2000 ...
##  $ property_tax  : num [1:200] 560 0 0 0 100 509 0 250 175 584 ...
##  $ fire_insurance: num [1:200] 71 70 10 18 62 89 16 55 40 26 ...
```

```
r <- lda(formula = furniture ~ ., data = house_data)
head(r$class)
```

```
## NULL
```

```
summary(r)
```

```
##         Length Class  Mode
## prior   2      -none- numeric
## counts   2      -none- numeric
## means   16      -none- numeric
## scaling  8      -none- numeric
## lev      2      -none- character
## svd      1      -none- numeric
## N        1      -none- numeric
## call     3      -none- call
## terms    3      terms  call
## xlevels  0      -none- list
```

Model Acceptance

In this model, we can see that the first linear discriminant explains all the between-group variance in the house data. Therefore, the model can be used to analyze the house data.

```r
r$svd
```

```
## [1] 5.232996
```

```r
(prop = r$svd^2/sum(r$svd^2))
```

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

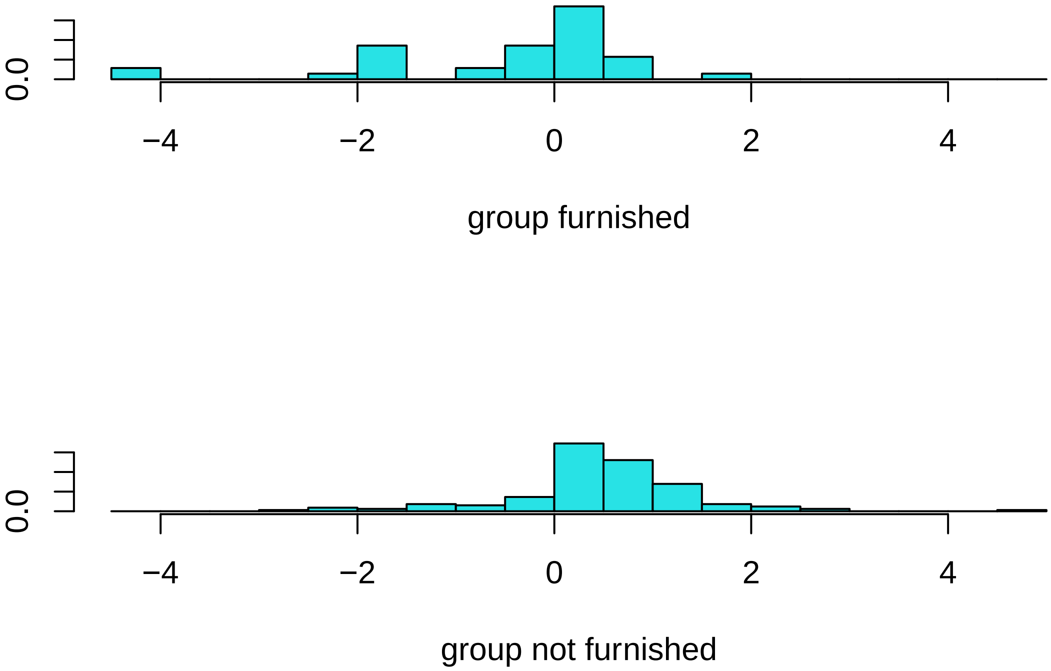
Residual Analysis

Since this model is a classification model, we focus on the posterior value of the model. The following code is to train the new model r3 and the model is used to test the model and display the predicted result and posterior probability. The plots of r1 and r3 shows how the model distinguishes between different furniture categories on training data

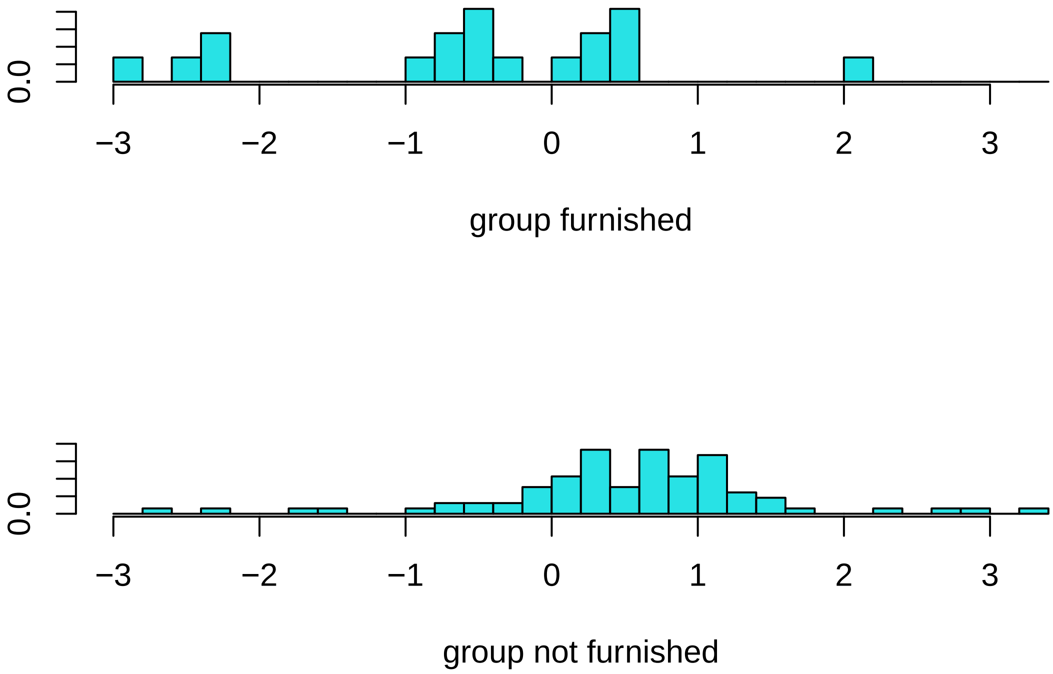
```r
r2 <- lda(formula = furniture ~ ., data = house_data, CV = TRUE)
head(r2$posterior, 3)
```

```
##   furnished not furnished
## 1 0.2904855     0.7095145
## 2 0.4439871     0.5560129
## 3 0.0793424     0.9206576
```

```r
plot(r)
```



```
train <- sample(1:200, 100)
r3 <- lda(furniture ~ .,
          house_data,
          prior = c(1,1)/2,
          subset = train)
plot(r3)
```



```r
plda = predict(object = r3, # predictions
               newdata = house_data[-train, ])
head(plda$class)
```

```
## [1] furnished     not furnished furnished     not furnished furnished
## [6] not furnished
## Levels: furnished not furnished
```

```r
head(plda$posterior, 6) # posterior prob.
```

```
##   furnished not furnished
## 1 0.7353088    0.26469119
## 2 0.4115428    0.58845721
## 3 0.7373188    0.26268122
## 4 0.3638763    0.63612371
## 5 0.9939096    0.00609038
## 6 0.4253634    0.57463659
```

```r
head(plda$x, 3)
```

```
##          LD1
## 1 -0.9599192
## 2  0.3359594
## 3 -0.9696453
```

Prediction

The data will be predicted in the model and the predicted first linear discriminant scores of the rent house are as follows.

```
r <- lda(furniture ~ .,
          house_data,
          prior = c(1,1)/2,)
prop.lda = r$svd^2/sum(r$svd^2)
plda <- predict(object = r,
                newdata = house_data)
dataset = data.frame(furniture = house_data[,"furniture"],lda = plda$x)
dataset$LD1
```

```
##   [1] -0.594789014 -1.601636788  0.934746798  0.093943933  1.744852303
##   [6] -0.606047030  0.133442620 -0.495942922  1.051692543  1.076099666
##  [11]  0.579509450  0.087465978  0.637485115  0.066311574  0.962403508
##  [16]  1.256685547 -4.272375039 -0.187511726 -0.304903986  0.232610047
##  [21]  0.887570670  0.665947645  1.239917737  2.349280403  0.855164083
##  [26]  0.300916484  0.159012483  0.350086839  0.533528473  0.147816237
##  [31]  0.098701403  0.796721987  0.299154147  0.807285598  0.327601142
##  [36] -1.812005038 -1.227534289  1.498608653 -0.015236887  1.651579464
##  [41]  0.410160506 -0.085081430  0.342958748  0.234905868  1.121264910
##  [46]  1.927662360 -2.327813152  1.460123850 -4.284490951  0.294766917
##  [51]  0.903543188  0.530036017  0.869912069  0.770052175  4.515935419
##  [56]  0.794919856  0.133335759  0.407141678  0.929742645  0.920853458
##  [61]  0.440916320 -0.244668483 -0.725607410  0.177404074  0.894757098
##  [66]  1.123532935  0.672261151  0.253089664 -0.147572890  0.465474925
##  [71]  0.917785604  1.873454008  2.597762427  0.189566851  0.219187443
##  [76]  0.403188414  0.551166362  0.280656127 -0.402475366  1.216608885
##  [81]  1.469761850  0.610292468  1.241726506  0.755132507  1.200623511
##  [86]  0.867618894 -1.878964072  0.300787325  0.777600059  0.362592129
##  [91]  1.781260982  0.466857279  2.083169107 -0.169919926  0.509785284
##  [96] -1.339240117  0.443401570  0.350092530  0.486677163 -1.155895879
## [101] -1.968991345 -0.115325502  0.955886614  0.196518065  0.027081589
## [106]  0.292127293 -1.694632396  0.670629423 -0.120634231  0.748515947
## [111]  0.048800836  0.349973359  0.033092343 -0.110111334 -0.106776440
## [116]  0.559699761 -1.065921917  0.781555848 -2.021058167  0.414495931
## [121]  0.859389087 -1.933822683  0.707328569  0.085670262  0.374478820
## [126]  0.501373933  1.074138453  0.818524584  0.790869759  0.392293507
## [131]  0.260564006  1.122318612  1.135750279 -1.601636788 -2.275402368
## [136] -0.830209609 -0.183093588  0.131612163 -0.011504713  0.498920357
## [141]  1.295615555  0.439108274  0.915606126  0.392904396 -0.172945448
## [146]  1.299140371  2.084413100  0.445124458  0.072084703  0.363039049
## [151] -0.783184934  0.920375750  0.855013917 -0.082174163  0.439526426
## [156] -2.706136686  1.039644251 -0.130715504 -0.710997130  2.565278772
## [161]  0.306257283  0.598982802  1.142512823 -1.046904481 -0.500876276
## [166]  0.303441126  1.112682797  1.523065802  2.040056062  0.154846414
## [171]  0.478211003  0.152863835  0.700883768  0.558847301  0.274722929
## [176]  0.083349618  0.094261322  0.196595344  0.870388186  0.249133314
## [181]  0.002047383  0.376146274  1.772552654  0.833903216  0.312774976
## [186]  0.398448625 -2.348595220  0.191623292  1.457502429  0.386339118
## [191]  0.111123394  0.644201387 -1.403049254  0.303293545  1.065223807
## [196]  0.946027095 -1.828954052  1.270052895  0.148013438  0.536713537
```

Model Accuracy

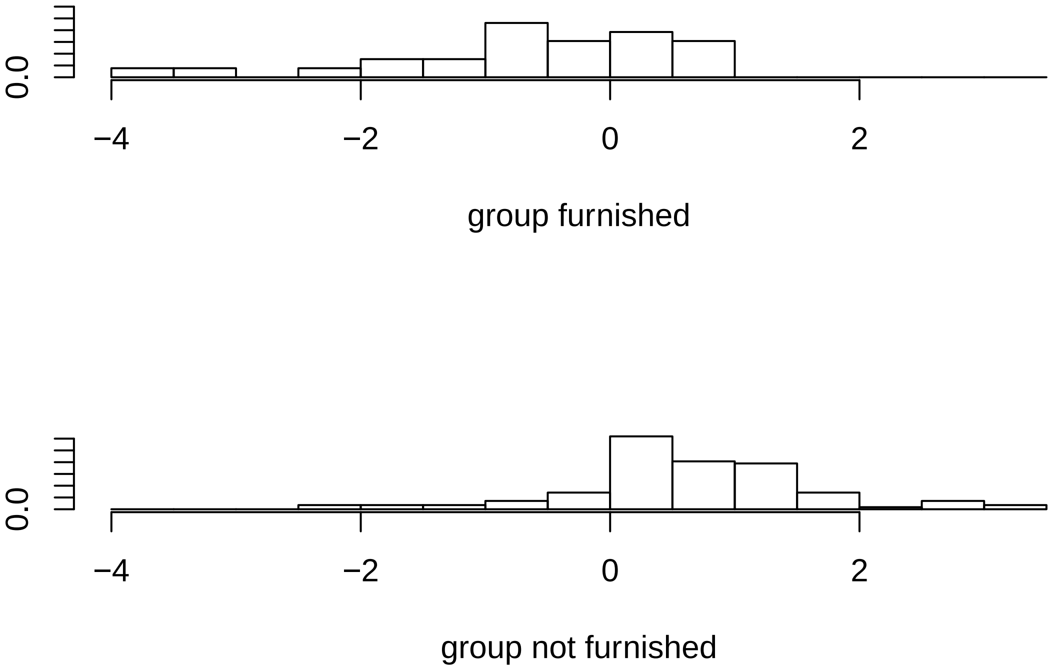
To observe the performance of the model, the test set is used to approximate accuracy.

```
set.seed(101)
sample_n(house_data,10)
```

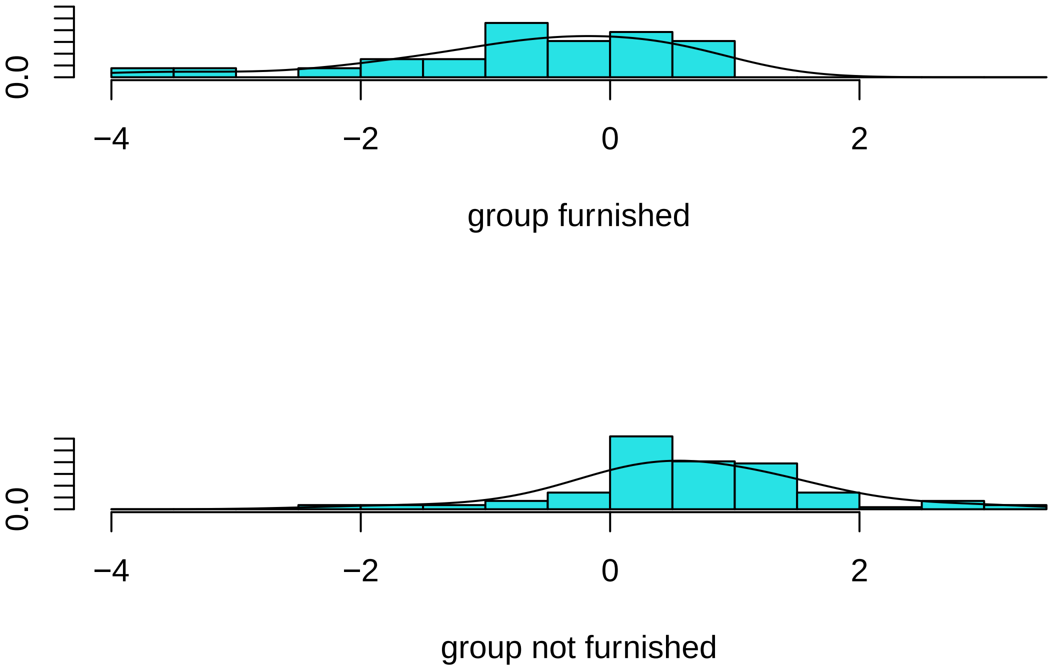
```
## # A tibble: 10 x 9
##     area rooms bathroom parking_spaces furniture    hoa rent_amount property_tax
##    <dbl> <dbl>    <dbl>          <dbl> <chr>      <dbl>       <dbl>        <dbl>
##  1   250     4        2              1 not furni~     0        2700          209
##  2    35     1        1              0 not furni~   270        1300            0
##  3    96     3        2              1 not furni~  1122        3050          231
##  4   137     3        3              1 furnished   1180        2900          214
##  5    40     1        1              1 not furni~     0        1200            0
##  6   301     4        5              4 furnished   4265       12500         1600
##  7    48     1        1              0 not furni~   309         700           28
##  8   140     2        3              2 furnished   1000        3000          113
##  9    70     2        1              1 not furni~   729         900          122
## 10    60     2        2              1 not furni~   440        1250           38
## # i 1 more variable: fire_insurance <dbl>
```

```
training_sample <- sample(c(TRUE, FALSE), nrow(house_data), replace = T, prob = c(0.75,0.25))
train <- house_data[training_sample, ]
test <- house_data[!training_sample, ]
lda.house <- lda(furniture ~ ., train)
plot(lda.house, col = as.integer(train$furniture))
```

```
## Warning in rect(breaks[-n], 0, breaks[-1L], est[[grp]], col = col, ...): NAs
## introduced by coercion
```



```
# Sometime bell curves are better
plot(lda.house, dimen = 1, type = "b")
```



```
lda.train <- predict(lda.house)
train$lda <- lda.train$class
table(train$lda,train$furniture)
```

```
##
##                  furnished not furnished
##   furnished              6             6
##   not furnished         20           107
```

```
# running accuracy on the training set shows how good the model is. It is not an indication of "true" a
lda.test <- predict(lda.house,test)
test$lda <- lda.test$class
table(test$lda,test$furniture)
```

```
##
##                  furnished not furnished
##   furnished              1             5
##   not furnished          8            47
```