Computer Assignment 7 - Logistic Regression and LDA

Machine Learning, Spring 2020

Rui Li

1994 Census Data

Let's walk through an example done during lecture. First, load the `adults.csv` data (downloaded originally from here).

```
adult = read.csv("adults.csv")
adult$X = NULL
```

Next, we will run the logistic regression model to predict the classifier `income`, which marks whether a given adult makes $\leq \$50k$ (coded as a 0) or $\geq \$50k$ (coded as a 1). To assess the performance of our model, we will only build the model on 75% of our data so that we can later use the remaining 25% as a testing data set.

```
set.seed(13)
training_size <- round(.75 * nrow(adult))  # training set size
indices = sample(1:nrow(adult), training_size)
training_set <- adult[indices,]
testing_set <- adult[-(indices),]
m1 <- glm(income ~ ., data = training_set, family = binomial('logit'))
summary(m1)

##
## Call:
## glm(formula = income ~ ., family = binomial("logit"), data = training_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7181  -0.5927  -0.2568  -0.0662   3.2070
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -9.495730   0.280207 -33.888  < 2e-16 ***
## age                      0.030356   0.001707  17.778  < 2e-16 ***
## workclassOther/Unknown  -1.611046   0.749696  -2.149   0.0316 *
## workclassPrivate         0.092437   0.055316   1.671   0.0947 .
## workclassSelf-Employed  -0.174826   0.071272  -2.453   0.0142 *
## education_num            0.322409   0.009698  33.244  < 2e-16 ***
## marital_statusMarried    1.984032   0.068998  28.755  < 2e-16 ***
## marital_statusSeparated -0.093530   0.166238  -0.563   0.5737
## marital_statusSingle    -0.510317   0.085823  -5.946 2.75e-09 ***
## marital_statusWidowed   -0.016882   0.153900  -0.110   0.9126
## occupationOther/Unknown  1.348355   0.749483   1.799   0.0720 .
## occupationProfessional   0.753131   0.068996  10.915  < 2e-16 ***
## occupationSales          0.490633   0.065412   7.501 6.35e-14 ***
## occupationService        0.090910   0.069874   1.301   0.1932
## occupationWhite-Collar   0.788020   0.054021  14.587  < 2e-16 ***
## raceAsian-Pac-Islander   0.191524   0.246788   0.776   0.4377
## raceBlack                0.341624   0.234546   1.457   0.1452
## raceOther               -0.223151   0.362520  -0.616   0.5382
## raceWhite                0.535386   0.223601   2.394   0.0166 *
## sexMale                  0.406079   0.053430   7.600 2.96e-14 ***
## hours_per_week           0.030613   0.001670  18.326  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27024  on 24420  degrees of freedom
## Residual deviance: 17967  on 24400  degrees of freedom
## AIC: 18009
##
## Number of Fisher Scoring iterations: 6
```

Examine the summary of our logistic regression model. Comment on the significance of each of our predictors. Do any of the significant predictors surprise you? Provide an interpretation of what the `Estimate` value is for the predictor age. Your answer should say something about this value's relation to the log-odds.

```
According to the summary above, we can notice that `age`, `education years`,
`married status`, `male`, `working hours`, and `occupation types` have the mo
st significant level to the income.

The significance of `age` and `working hours` surprise me. It is not resonabl
e that one is more likely to have income > $50k when he is older and working
longer.

Also, for every one unit change in `age`, the log odds of income>=$50k increa
ses by 0.03.
```

Now that you have created a model on the `training_data`, use the `predict` function in **R** to use your model to classify the data in the `testing_data`. What proportion of values were classified correctly?

```
pre_test = as.data.frame(predict(m1, newdata = testing_set, type = "response
"))
pre_class = pre_test$`predict(m1, newdata = testing_set, type = "response")`>
=1/2
true_class = testing_set$income=='>50K'
glm_correct = (sum((pre_class+true_class)==2)+sum((pre_class+true_class)==0))
/length(testing_set$income)
```

```
The proportion of correctly classified values is 0.8267.
```

Build a LDA model on the `training_data`, and see how well it performs classifying the observations in the `testing_data`. Compare the proportion of values classified correctly to this same metric we just calculated for the logistic regression model.

```
library(MASS)
library(dplyr)

##
## Attaching package: 'dplyr'

## 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

model_lda = lda(income~., data = training_set)
lda_predictions <- model_lda %>% predict(testing_set)
lda_correct = sum(lda_predictions$class==testing_set$income)/length(testing_s
et$income)

The lda_corrrect is 0.8267, and glm_correct is 0.08267. These two values are
quite similar.
```

Magazine Reseller Data

Now, let us apply these same methods on some new data, the given data `kids.csv`. Each observation of this data set records the demographics of a person as well as whether or not they bought a magazine.

The variables given are as follows:

1. Household Income (Income; rounded to the nearest $1,000.00)
2. Gender (IsFemale = 1 if the person is female, 0 otherwise)
3. Marital Status (IsMarried = 1 if married, 0 otherwise)
4. College Educated (HasCollege = 1 if has one or more years of college education, 0 otherwise)
5. Employed in a Profession (IsProfessional = 1 if employed in a profession, 0 otherwise)
6. Retired (IsRetired = 1 if retired, 0 otherwise)
7. Not employed (Unemployed = 1 if not employed, 0 otherwise)
8. Length of Residency in Current City (ResLength; in years)

9. Dual Income if Married (Dual = 1 if dual income, 0 otherwise)
10. Children (Minors = 1 if children under 18 are in the household, 0 otherwise)
11. Home ownership (Own = 1 if own residence, 0 otherwise)
12. Resident type (House = 1 if the residence is a single-family house, 0 otherwise)
13. Race (White = 1 if the race is white, 0 otherwise)
14. Language (English = 1 is the primary language in the household is English, 0 otherwise)

as well as a binary classifier Buy that marks whether or not a given person bought a magazine.

Randomly assign 75% of your data as the training data and the other 25% as your testing data. Build a logistic regression model on the training, discuss the model summary and significance of the parameter values, and assess the model's performance in the testing data. Then build a LDA model on the training data and see how well it performs classifying the observations in the testing data. Compare the proportion of values classified correctly to this same metric we just calculated for the logistic regression model.

This last exercise leaves much of the process up to you! Go through previous code, google issues, and read manual pages if you get lost.

```
#insert the data sheet and create the training and test data
kid = read.csv("kids.csv")
kid$Obs.No.=NULL

set.seed(13)
training_size <- round(.75 * nrow(kid)) # training set size
indices = sample(1:nrow(kid), training_size)
training_kid <- kid[indices,]
testing_kid <- kid[-(indices),]

#set up the glm model
glm_kid <- glm(Buy ~ ., data = training_kid, family = binomial('logit'))
summary(glm_kid)

##
## Call:
## glm(formula = Buy ~ ., family = binomial("logit"), data = training_kid)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.43823  -0.09806  -0.01121  -0.00178   2.55715
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.835e+01  2.635e+00  -6.964 3.31e-12 ***
## Income            1.969e-04  2.629e-05   7.490 6.91e-14 ***
## Is.Female         1.385e+00  5.313e-01   2.607  0.00913 **
## Is.Married        6.666e-01  7.006e-01   0.952  0.34131
## Has.College      -2.293e-02  5.162e-01  -0.044  0.96457
## Is.Professional   3.094e-01  5.427e-01   0.570  0.56857
## Is.Retired       -2.559e+00  1.190e+00  -2.151  0.03147 *
## Unemployed        1.185e+00  5.632e+00   0.210  0.83331
## Residence.Length  3.292e-02  1.698e-02   1.939  0.05250 .
## Dual.Income      -6.396e-02  6.089e-01  -0.105  0.91633
## Minors            9.535e-01  5.305e-01   1.797  0.07227 .
## Own               9.445e-01  6.449e-01   1.465  0.14303
## House            -5.403e-01  6.812e-01  -0.793  0.42764
## White             2.003e+00  6.513e-01   3.075  0.00211 **
## English           1.970e+00  1.058e+00   1.862  0.06257 .
## Prev.Child.Mag    2.086e+00  9.015e-01   2.314  0.02068 *
## Prev.Parent.Mag   8.173e-01  7.728e-01   1.057  0.29028
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 473.37  on 504  degrees of freedom
## Residual deviance: 135.17  on 488  degrees of freedom
## AIC: 169.17
##
## Number of Fisher Scoring iterations: 9
```

According to the summary above, we can notice that `income`, `female`, and `white` have the greatest significance on the `buy` logistic model.

```
#glm correct for kids
kid_test = as.data.frame(predict(glm_kid, newdata = testing_kid, type = "response"))
pre_kid = kid_test$`predict(glm_kid, newdata = testing_kid, type = "response")`>=1/2
true_kid = testing_kid$Buy=='1'
kid_glm_correct = (sum((pre_kid+true_kid)==2)+sum((pre_kid+true_kid)==0))/length(testing_kid$Buy)
```

```
#lda for kids
library(MASS)
library(dplyr)
kids_lda = lda(Buy~., data = training_kid)
kid_lda_predictions <- kids_lda %>% predict(testing_kid)
kid_lda_correct = sum(kid_lda_predictions$class==testing_kid$Buy)/length(testing_kid$Buy)
```

The glm corrects for kids is 0.9345, and the lda corrects for kids is 0.9226, which are quite similar.

This data, and the information about it, was gotten from here.