Computer Assignment 7 - Logistic Regression and LDA

Machine Learning, Spring 2020

YOUR NAME

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 $\leqslant \$50k$ (coded as a 0) or $\geqslant \$50k$ (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)
```

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.

YOUR ANALYSIS HERE

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?

```
YOUR CODE, AND ANALYSIS, HERE
```

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.

```
YOUR CODE, AND ANALYSIS, HERE
```

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.

```
set.seed(13)
YOUR CODE, AND ANALYSIS, HERE
```

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