G4058 Assignment 4

Due: Wednesday, April 6, 2016 by 6PM

April 1, 2016

Dowload the dataset.RData file from CourseWorks to your working directory and load it into R via

```
load("dataset.RData")
str(dataset)

## 'data.frame': 10000 obs. of  6 variables:
##  $ Amount.Requested    : num  1500 6000 6000 3000 35000 ...
##  $ Debt.To.Income.Ratio: num  17.32 23.6 13.88 1.95 22.08 ...
##  $ Zip.Code            : chr  "797" "986" "434" "891" ...
##  $ State               : chr  "TX" "WA" "OH" "NV" ...
##  $ Employment.Length   : num  0 10 0 0 0 0 0 5 10 0 ...
##  $ y                   : int  0 1 0 0 0 0 0 1 0 0 ...
```

In these data, the outcome of interest is whether a person loan was approved by the bank. The variables are

- `Amount.Requested`: The proposed amount for the loan
- `Debt.To.Income.Ratio`: The ratio of the applicant's debt (excluding mortgages and the proposed loan) payments each month to the applicant's stated monthly income
- `Zip.Code`: The 3-digit zip code of the applicant
- `State`: The state where the applicant lives
- `Employment.Length`: The number of years that the applicant has worked at the same job. `10` indicates at least ten years, `0` indicates less than one year, and `-1` indicates unemployed.
- `y`: A binary variable indicating whether the loan was approved

Split the data into a training set containing the first half of the observations and a testing set containing the second half.

1 Plot

Using the training dataset, make a scatterplot of `Amount.Requested` and `Debt.To.Income.Ratio`. Use the `pch` argument to plot different symbols depending on whether **y** is 1 or 0.

2 Initial Model

Use the `glm()` function in R to estimate a logit model for this outcome in the training data as a function `Amount.Requested`, `Debt.To.Income.` and their interaction. How well does this model perform when predicting the outcome in the testing data where the criterion is the proportion of times the model classifies a testing observation correctly, according to the conventional rule that observations with a probability greater than 0.5 are classified as 1 and observations with a probability less than or equal to 0.5 are classified as 0?

3 Expanded Model

Does including any additional predictor(s) improve the performance of the model when predicting the outcome in the testing data?

4 Using optim

Re-estimate the model in question 3 by defining a log-likelihood function and maximizing it with the `optim` function. Verify that your parameter estimates are very similar to those in question 3.

5 LDA and QDA

Using the same predictors as in question 3, use Linear Discriminant Analysis and Quadratic Discriminant Analysis to estimate the parameters of the model with the training data and predict the outcome in the testing data. How do these two techniques perform relative to the logit models you estimated previously?

6 LASSO Penalization

Use the `glmnet` function in the glmnet package to estimate this logit model with L1 penalization. Does penalization of the fit in the training data improve the classification accuracy in the testing data?