Assignment 3

(Team 6)

**Goal**: Predicting whether a person makes over 50K a year.

**Sources for Data and its descriptions**: UCI Machine Learning Repository --https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data

https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names

**Key findings/Summary/Results**

**Methods and Analysis**

1. Data munging

; Data was imported from the url and was cleaned (removed the observations with “?” mark and checked outliers) before working on it.

About ? mark

Why we decide to delete ? instead of other procedures?

1. ? mark only appear only on 2nd, 7th 14th column and some row(observation) contains more than 2 missing value. If we delete those row, we still have 30162 observations which should be enough to start our analysis. Otherwise, our data set will be very messy.

2. Besides, we also noticed that the original data set has 48842 observations. And for preparation, 3620 observations with unknown data(NA) have already been deleted in our data-description file. Therefore we should continue the way we treat missing values.

2. Exploratory data analysis

; As requested, we used the first half of the data as a training data to explore data further.

\*\*\*Why we make these graph??

1. to have a intuitively look at our data. Find some pattern if there exist.

2. Find if we have a vector extremely correlated with our response vector (but answer is no)

3. \*Check the GAUSSIAN ASSUMPTION: workclass maybe Gaussian. But some of them obviously not: like Native country. Hoursperweek

and education, etc.

[insert Graphs]

3. Classification methods, models and algorithms that we had studied in class.

; We began with conducting linear classification models such as logistic regression. Logistic regression is a widely-used model when the response is categorical. If there are two possible outcomes, we use the binomial distribution, else we use the multinomial.

logistic model, as it is a simple linear regression model to start when having two classes for a response variable.

LDA:

Outlier detection: there’s no outlier in data for vector (fnlwgt and age). There’re outliers in the capital gain and capital loss. However, since we have no idea of what the data represents, we could not simply delete these predicors.

LDA result:

There are 2426 test samples being misclassified thus the test error is 0.1608647.

How about the training error?

There are 2486 train samples being misclassified thus the train error rate is 0.1648432.

Description (Analysis):

As for prediction on social science, predict error around 15% is already a really good prediction. However, since our goal is to predict “whether a person makes over 50K a year”. And when we looked in detail as the miss classification rate on test data set, we found 43.26% (1635 out of 3780)people with salary over 50K have been misclassified. Therefore LDA model may not be the model we’re finding.

QDA:

Just like LDA: 2982 of test samples were misclassified thus the test error rate is 0.1977322.

And 2970 of train samples were misclassified thus the train error is 0.1969365. It’s even worse than LDA. This is because QDA has a lot of coefficients need to be estimate in this data set. As usual, the tradeoff between LDA or QDA is one of bias and variance: LDA makes stronger assumptions and obtains estimates with lower variance

Besides, if take detailed look at QDA prediction, we could also find there’re 65% (2459 out of 3780) people with salary over 50K have been misclassified. It’s really bad prediction so we will not use QDA for our prediction.

RDA ( Regularised discriminant analysis)

We also tried RDA in r. However, since RDA is basically a combination of LDA and QDA. And neither of these models works well. Therefore the result didn’t show up well also. We got misclassification rate around 16.5% and still bad prediction on our goal.

4. Use the second half of the data for validation.

;

**Essential R scripts**

[insert R scripts that is important]

1. Data import/cleaning