Modern Data Mining - HW 3

Anirudh Bajaj
Esther Shin
Matt LeBaron
Raman Chadha

Overview / Instructions

This is homework #3 of STAT 471/571/701. It will be **due on 28 October, 2018 by 11:59 PM** on Canvas. You can directly edit this file to add your answers. Submit the Rmd file, a PDF or word or HTML version with **only 1 submission** per HW team.

Note: To minimize your work and errors, we provide this Rmd file to guide you in the process of building your final report. To that end, we've included code to load the necessary data files. Make sure that the following files are in the same folder as this R Markdown file:

- `FRAMINGHAM.dat`
- `Bills.subset.csv`
- `Bills.subset.test.csv`

The data should load properly if you are working in Rstudio, *without needing to change your working directory.*

Solutions will be posted. Make sure to compare your answers to and understand the solutions.

R Markdown / Knitr tips

You should think of this R Markdown file as generating a polished report, one that you would be happy to show other people (or your boss). There shouldn't be any extraneous output; all graphs and code run should clearly have a reason to be run. That means that any output in the final file should have explanations.

A few tips:

- Keep each chunk to only output one thing! In R, if you're not doing an assignment (with the `<-` operator), it's probably going to print something.
- If you don't want to print the R code you wrote (but want to run it, and want to show the results), use a chunk declaration like this: `{r, echo=F}`
- If you don't want to show the results of the R code or the original code, use a chunk declaration like: `{r, include=F}`
- If you don't want to show the results, but show the original code, use a chunk declaration like: `{r, results='hide'}`.
- If you don't want to run the R code at all use `{r, eval = F}`.
- We show a few examples of these options in the below example code.
- For more details about these R Markdown options, see the documentation.
- Delete the instructions and this R Markdown section, since they're not part of your overall report.

Problem 0

Review the code and concepts covered during lecture, in particular, logistic regression and classification.

Problem 1

We will continue to use the Framingham Data (`Framingham.dat`) so that you are already familiar with the data and the variables. All the results are obtained through training data.

To keep our answers consistent, use a subset of the data, and exclude anyone with a missing entry. For your convenience, we've loaded it here together with a brief summary about the data.

We note that this dataset contains 311 people diagnosed with heart disease and 1095 without heart disease.

```
   0    1
1095  311
```

After a quick cleaning up here is a summary about the data:

```
# using the comment="     ", we get rid of the ## in the output.
summary(hd_data.f)
```

```
       HD             AGE            SEX           SBP              DBP
       0:1086   Min.   :45.00   FEMALE:730   Min.   : 90.0   Min.   : 50.00
       1: 307   1st Qu.:48.00   MALE  :663   1st Qu.:130.0   1st Qu.: 80.00
                Median :52.00                Median :142.0   Median : 90.00
                Mean   :52.43                Mean   :148.1   Mean   : 90.16
                3rd Qu.:56.00                3rd Qu.:160.0   3rd Qu.: 98.00
                Max.   :62.00                Max.   :300.0   Max.   :160.00
            CHOL            FRW             CIG
       Min.   : 96.0   Min.   : 52.0   Min.   : 0.000
       1st Qu.:200.0   1st Qu.: 94.0   1st Qu.: 0.000
       Median :230.0   Median :103.0   Median : 0.000
       Mean   :234.6   Mean   :105.4   Mean   : 8.035
       3rd Qu.:264.0   3rd Qu.:114.0   3rd Qu.:20.000
       Max.   :430.0   Max.   :222.0   Max.   :60.000
```

Part 1A

Conceptual questions to understand building blocks of logistic regression. All the codes in this part should be hidden.

i. Take a random subsample of size 5 from `hd_data_f` which only includes `HD` and `SBP`. Also set `set.seed(50)`. List the three observations neatly below. No code should be shown here.

ii. Write down the likelihood function using the five observations above.

iii. Find the MLE's based on this subset. Report the estimated logit function and the probability of `HD`=1. Briefly explain how the MLE's are obtained based on ii. above.

Part 1B

Goal: Identify important risk factors for `Heart.Disease.` through logistic regression. Start a fit with just one factor, `SBP`, and call it `fit1`. Let us add one variable to this at a time from among the rest of the variables.

```
fit1 <- glm(HD~SBP, hd_data.f, family=binomial)
summary(fit1)
fit1.1 <- glm(HD~SBP + AGE, hd_data.f, family=binomial)
summary(fit1.1)
fit1.2 <- glm(HD~SBP + SEX, hd_data.f, family=binomial)
summary(fit1.2)
fit1.3 <- glm(HD~SBP + DBP, hd_data.f, family=binomial)
summary(fit1.3)
fit1.4 <- glm(HD~SBP + CHOL, hd_data.f, family=binomial)
summary(fit1.4)
fit1.5 <- glm(HD~SBP + DBP, hd_data.f, family=binomial)
summary(fit1.5)
```

```
fit1.6 <- glm(HD~SBP + FRW, hd_data.f, family=binomial)
summary(fit1.6)
fit1.7 <- glm(HD~SBP + CIG, hd_data.f, family=binomial)
summary(fit1.7)
```

i. Which single variable would be the most important to add? Add it to your model, and call the new fit `fit2`.

We will pick up the variable either with highest $|z|$ value, or smallest p value. From all the two variable models we see that `SEX` will be the most important addition on top of the SBP. And here is the summary report.

```
## How to control the summary(fit2) output to cut some junk?
## We could use packages: xtable or broom.
library(xtable)
options(xtable.comment = FALSE)
fit2 <- glm(HD~SBP + SEX, hd_data.f, family=binomial)
xtable(fit2)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.5703	0.3897	-11.73	0.0000
SBP	0.0187	0.0023	8.05	0.0000
SEXMALE	0.9034	0.1398	6.46	0.0000

ii. Is the residual deviance of `fit2` always smaller than that of `fit1`? Why or why not?

iii. Perform both the Wald test and the Likelihood ratio tests (Chi-Squared) to see if the added variable is significant at the .01 level. What are the p-values from each test? Are they the same?

Part 1C - Model building

Start with all variables. Our goal is to fit a well-fitting model, that is still small and easy to interpret (parsimonious).

i. Use backward selection method. Only keep variables whose coefficients are significantly different from 0 at .05 level. Kick out the variable with the largest p-value first, and then re-fit the model to see if there are other variables you want to kick out.

ii. Use AIC as the criterion for model selection. Find a model with small AIC through exhaustive search. Does exhaustive search guarantee that the p-values for all the remaining variables are less than .05? Is our final model here the same as the model from backwards elimination?

iii. Use the model chosen from part ii. as the final model. Write a brief summary to describe important factors relating to Heart Diseases (i.e. the relationships between those variables in the model and heart disease). Give a definition of "important factors".

Part 1D - Prediction

Liz is a patient with the following readings: `AGE=50, GENDER=FEMALE, SBP=110, DBP=80, CHOL=180, FRW=105, CIG=0`. What is the probability that she will have heart disease, according to our final model?

Part 2 - Classification analysis

a. Display the ROC curve using `fit1`. Explain what ROC reports and how to use the graph. Specify the classifier such that the False Positive rate is less than .1 and the True Positive rate is as high as possible.

b. Overlay two ROC curves: one from `fit1`, the other from `fit2`. Does one curve always contain the other curve? Is the AUC of one curve always larger than the AUC of the other one? Why or why not?

c. Estimate the Positive Prediction Values and Negative Prediction Values for `fit1` and `fit2` using .5 as a threshold. Which model is more desirable if we prioritize the Positive Prediction values?

d. (Optional/extra credit) For `fit1`: overlay two curves, but put the threshold over the probability function as the x-axis and positive prediction values and the negative prediction values as the y-axis. Overlay the same plot for `fit2`. Which model would you choose if the set of positive and negative prediction values are the concerns? If you can find an R package to do so, you may use it directly.

Part 3 - Bayes Rule

Bayes rules with risk ratio $\frac{a_{10}}{a_{01}} = 10$ or $\frac{a_{10}}{a_{01}} = 1$. Use your final model obtained from 1 B) to build a class of linear classifiers.

a. Write down the linear boundary for the Bayes classifier if the risk ratio of $a_{10}/a_{01} = 10$.
b. What is your estimated weighted misclassification error for this given risk ratio?
c. Recall Liz, our patient from part 1. How would you classify her under this classifier?

Now, draw two estimated curves where x = posterior threshold, and y = misclassification errors, corresponding to the thresholding rule given in x-axis.

d. Use weighted misclassification error, and set $a_{10}/a_{01} = 10$. How well does the Bayes rule classifier perform?
e. Use weighted misclassification error, and set $a_{10}/a_{01} = 1$. How well does the Bayes rule classifier perform?

Problem 2

How well can we predict whether a bill will be passed by the legislature?

Hundreds to thousands of bills are written each year in Pennsylvania. Some are long, others are short. Most of the bills do not even get to be voted on ("sent to the floor"). The chamber meets for 2-year sessions. Bills that are not voted on before the end of the session (or which are voted on but lose the vote) are declared dead. Most bills die. In this study we examine about 8000 bills proposed since 2009, with the goal of building a classifier which has decent power to forecast which bills are likely to be passed.

We have available some information about 8011 bills pertaining to legislation introduced into the Pennsylvania House of Representatives. The goal is to predict which proposals will pass the House. Here is some information about the data:

The response is the variable called `status.` `Bill:passed` means that the bill passed the House; `governor:signed` means that the bill passed both chambers (including the House) and was enacted into law; `governor:received` means that the bill has passed both chambers and was placed before the governor for consideration. All three of these statuses signify a success or a PASS (Meaning that the legislature passed the bill. This does not require it becoming law). All other outcomes are failures.

Here are the rest of the columns:

- `Session` : in which legislative session was the bill introduced
- `Sponsor_party` : the party of the legislator who sponsored the bill (every bill has a sponsor)
- `Bill_id` : of the form HB-[bill number]-[session], e.g., `HB-2661-2013-2014` for the 2661st House Bill introduced in the 2013-2014 session.
- `Num_cosponsors` : how many legislators cosponsored the bill
- `Num_d_cosponsors` : how many Democrats cosponsored the bill
- `Num_r_cosponsors` : how many Republicans cosponsored the bill
- `Title_word_count` : how many words are in the bill's title
- `Originating_committee` : most bills are sent ("referred") to a committee of jurisdiction (like the transportation committee, banking & insurance committee, agriculture & rural affairs committee) where they are discussed and amended. The originating committee is the committee to which a bill is referred.
- `Day_of_week_introduced` : on what day the bill was introduced in the House (1 is Monday)
- `Num_amendments` : how many amendments the bill has

- `Is_sponsor_in_leadership` : does the sponsor of the bill hold a position inside the House (such as speaker, majority leader, etc.)
- `num_originating_committee_cosponsors` : how many cosponsors sit on the committee to which the bill is referred
- `num_originating_committee_cosponsors_r` : how many Republican cosponsors sit on the committee to which the bill is referred
- `num_originating_committee_cosponsors_d` - how many Democratic cosponsors sit on the committee to which the bill is referred

The data you can use to build the classifier is called `Bills.subset`. It contains 7011 records from the full data set. I took a random sample of 1000 bills from the 2013-2014 session as testing data set in order to test the quality of your classifier, it is called `Bills.subset.test`.

Your job is to choose a best set of classifiers such that

- The testing ROC curve pushes to the upper left corner the most, and has a competitive AUC value.
- Propose a reasonable loss function, and report the Bayes rule together with its weighted MIC.
- You may also create some sensible variables based on the predictors or make other transformations to improve the performance of your classifier.

Here is what you need to report:

1. Write a summary about the goal of the project. Give some background information. If desired, you may go online to find out more information.
2. Give a preliminary summary of the data.
3. Based on the data available to you, you need to build a classifier. Provide the following information:
    - The process of building your classifier
    - Methods explored, and why you chose your final model
    - Did you use a training and test set to build your classifier using the training data? If so, describe the process including information about the size of your training and test sets.
    - What is the criterion being used to build your classifier?
    - How do you estimate the quality of your classifier?
4. Suggestions you may have: what important features should have been collected which would have helped us to improve the quality of the classifiers.

Final notes: The data is graciously lent from a friend. It is only meant for you to use in this class. All other uses are prohibited without permission.