Stat 652 Project Guidelines

Brad McNeney

2018-11-14

Data

The data are from the Canadian Community Health Survey (CCHS) – Healthy Aging module. Documentation for the study is available in the Documentation folder. Please read Sections 1 and 2 of the User Guide document (`CCHS_HA_User_Guide.pdf`). The project will have two parts.

Part I: Health Utility Indices and Cognitive Health Status

The file `HUI.csv` contains a training dataset comprised of 8 health-utility-index (HUI) variables. The 8 are a subset of the 13 variables described on pages 45-54 of the file `CCHS_HA_Derived_variables.pdf`. The first variable, `HUIDCOG` is a categorical variable with six levels of cognitive ability. See the Derived variables document for a description of the levels.

```
hui <- read.csv("hui.csv")
dim(hui)
```

```
## [1] 20000     9
```

```
summary(hui)
```

```
##            DHHGAGE       DHH_SEX                  HUIDCOG
##  55 TO 59 YEARS:3085   FEMALE:11385   COG. ATT. LEVE 1:13949
##  60 TO 64 YEARS:2982   MALE   : 8615   COG. ATT. LEVE 2:  496
##  85 AND OLDER  :2602                  COG. ATT. LEVE 3: 3764
##  65 TO 69 YEARS:2595                  COG. ATT. LEVE 4: 1268
##  70 TO 74 YEARS:1958                  COG. ATT. LEVE 5:  429
##  75 TO 79 YEARS:1928                  COG. ATT. LEVE 6:   71
##  (Other)       :4850                  NOT STATED      :   23
##             HUIGDEX                  HUIDEMO                  HUIGHER
##  LIM. HANDS/F   :  252   EMOT. ATT. LEV.1:14912   NO PROBLEMS    :17335
##  NOT STATED     :   10   EMOT. ATT. LEV.2: 4067   NOT STATED     :  296
##  USE OF HANDS/F.:19738   EMOT. ATT. LEV.3:  749   PROB./CORR.    : 1579
##                          EMOT. ATT. LEV.4:  183   PROB./NOT CORR.:  790
##                          EMOT. ATT. LEV.5:   39
##                          NOT STATED      :   50
##
##             HUIGMOB                  HUIGSPE                  HUIGVIS
##  NEED MECH. SUPP: 1580   NO PROBLEMS     :19837   NO PROBLEMS     : 4210
##  NO AID REQUIRED:  322   NOT STATED      :   11   NOT STATED      :  142
##  NO PROBLEMS    :17496   PARTIAL/NOT UND.:  152   VISUAL P. UNCOR.:  658
##  NOT STATED     :   16                            VISUAL PROB. COR:14990
##  REQUIRES HELP  :  586
##
##
```

Use the other HUIs to predict `HUIDGOC`. I have a hold-out test dataset on which you will eventually evaluate your predictions.

Part II: Predicting a HUI Score with Other Variables

In this part of the project you will use a larger (wider) dataset called `HStrain` that consists of 591 variables measured on 10000 subjects. You will find these data in the file `HStrain.csv.zip` in the Data directory; unzip it and use the R function `read.csv()` to read it into R. I have removed all of the subjects with `NOT STATED` entries in any of the 591 variables.

In the dataset, the categorical HUI variables have been removed and replaced by a single score called `HUIDHSI`. The description of `HUIDHSI` is as follows:

This derived variable is a Health Utilities Index which provides a description of an individual's overall functional health, based on eight attributes: vision, hearing, speech, ambulation (ability to get around), dexterity (use of hands and fingers), emotion (feelings), cognition (memory and thinking) and pain (in HUP module). The version of the index used in CCHS is adapted from the HUI Mark 3 (HUI3). The index is designed to produce an overall health utility score. This multi-attribute utility index produces a score ranging from 1.000 (perfect health), through 0.000 (health status equal to death) to -0.360 (health status worse than death).

Your task is to predict `HUIDHSI` with the other 590 variables in the `HStrain` dataset. You may work with `HUIDHIS` as-is (a quantitative variable), **or** break it into a binary variable that has value 0 if `HUIDHIS` is less than the median value of 0.905, and 1 otherwise. I have a hold-out test dataset on which you will eventually evaluate your predictions.

Project Length and Scope

Your report should be no more than 5 pages long, plus references. You must also include an Appendix of R code that can be used to reproduce the analyses refered to in the report. There is no page limit for the Appendix, but please use judgement about what to include. Too long and it is not likely to be read. You are encouraged to try several prediction methods, and can compare these methods, but your report should focus on one method in particular.

Grading Criteria

The criteria for the report are as follows.

Report (25 marks)

The report should have the following sections

1. Introduction (brief)
2. Data (brief)
3. Methods
4. Results
5. Conclusions and Discussion

The reports will be judged on the following criteria.

- Content (20): The content should be clear, accurate, complete and at the level of students in Stat 452. In the Methods you should provide a brief description of any statistical methods you use. Please restrict yourself to methods that were covered in class. You can mention methods not covered as areas of future work. Methods you considered but were not the focus of your report should be briefly mentioned here too. In Results you should summarize and interpret the fitted model. Though the primary goal is prediction, your insights into the data-generating process are important. Refer to the Appendix for the code that implements your prediction equation. In the Conclusions and Discussion present your conclusions, discuss short-comings of your approach, and, optionally, ideas for further work.

- Organization (3): Though the report is structured, you should present your ideas logically within each section.
- Grammar and spelling (2): Please proof-read your report.

Code (15 marks)

The code in your Appendix should look correct and be readable. Given the size of the dataset, you do not need to provide run-able code for all analyses. Use `{r, eval=FALSE}` in your computationally-intensive code chunks to prevent them from running. The Appendix will be judged on the following criteria.

- Software Details (2): List the version of R you are using and the names of all packages used in your analysis **at the beginning** of the Appendix. Please also provide an estimate of the time it will take to knit the code if more than about 2 minutes.
- Correctness (5): There should be no errors in data processing, function calls, etc.
- Readability (5): The steps of your analysis should be clearly layed out and it should be easy for the reader to find the final prediction equation/method.
- Efficiency (3): Please take steps to avoid computational inefficiencies, such as loops and excessive copying of large R objects.