Stat 652 Project Guidelines

Brad McNeney

2017-11-05

Data

Pima Indians from the Southwest US have a high incidence of type 2 diabetes. In your project you will use information on eight covariates to predict diabetes status. The dataset is in the CSV file `pima-diabetes.csv` in this directory.

```
pima <- read.csv("pima-diabetes.csv")
dim(pima)
```

```
## [1] 768   9
```

```
head(pima,n=3)
```

```
##   Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
## 1           6     148            72            35       0 33.6
## 2           1      85            66            29       0 26.6
## 3           8     183            64             0       0 23.3
##   DiabetesPedigreeFunction Age Outcome
## 1                    0.627  50       1
## 2                    0.351  31       0
## 3                    0.672  32       1
```

The variables are described in more detail in the file `pima-diabetesDescr.txt` The variable `Outcome` is diabetes status.

As we discussed in the week 8 exercises, there are many zeros in the data that are implausible and probably represent missing values. You should use your own judgment on recoding missing values.

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. I have not held out a portion of the data to use for testing your prediction method. Instead, you should summarize the steps you took to estimate test error.

Grading Criteria

The criteria for the report are as follows. [Whether or not the project includes a review of another student's report, worth an additional 5 marks, is still pending.]

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 (17): 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 that were covered in class, and more in-depth descriptions of any methods that are related to, but not covered in class. 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.
- References (3): You must properly attribute the ideas and work of others.
- Grammar and spelling (2): Please proof-read your report.

Code (15 marks)

The code in your Appendix should be correct, readable and should knit without errors. A secondary consideration is efficiency. 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.