Assignment 4

edelsonc

September 7, 2016

Using data to answer questions first requires an understanding of the quality of the data. This can be assessed in a number of forms, but can be broadly broken down into qualitative and quantitative measurements.

We will assess the quality of the data used in last weeks exercise, and from that determine to what extent we can predict discharge status.

The data can be accessed at the following link: https://archive.ics.uci.edu/ml/machine-learning-databases/00296/dataset_diabetes.zip

Additionally, the homepage for the data is located here: https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008

Qualitative Measures

A quick look around the homepage immediately leads us to believe that this data comes from a reliable source. An NIH grant financed it, it was distributed by CERNER, is currently hosted by University of California Irvine, and was used in a peer reviewed journal article. Additionally, the information was initially collected from 130 US hospitals.

All of these institutes are fairly trust worthy, and the sequence of data transfer means that this data has been reviewed multiple times at different levels for different things. For instance, just on its homepage, we can see a description of many of the basic attributes of the data. These include Set Characteristics, Missing Values, Instances, Attributes, Attribute Types, and a few others. This lets us know that, at the bare minimum, someone as looked through the data and cataloged these point.

Furthermore, if we go and look at the article produced using the data (which is linked on the homepage) we can learn a lot more about the entries of the data, as well as the intended purpose of the data. This helps us learn if the information is biased as well as what each attribute means for the question at hand.

These points combined lead us to believe that the data is reasonably trustworthy, and that it can be used for analysis.

Quantitative Measures

From UCI's site we already know that there are missing values in the data set. If we then look at the paper we see that they documented the percentage of missing values in the each attribute. We can confirm this with a quick test

```r
f_path <-"~/Desktop/Data_Science/EDA/assignment_4/dataset_diabetes/diabetic_data.csv"

diabetic_data <- read.csv(f_path)

diabetic_data[diabetic_data == "?"] <- NA

100 * sum(complete.cases(diabetic_data))/nrow(diabetic_data) # ratio of complete to total rows
```

```
## [1] 1.0249
```

So Just over 1 % of our data is complete…a most dismal statistic. However, if we looked at the paper, we say that 97% of weight was missing, as well as large chuncks of a few other catagories. If we don't think these values are important for discharge status, which is a desicion we'd have to carefully consider, we can simply remove them

```
diabetic_reduced <- diabetic_data[c(7,8)]  # reduces to discharge, admin type and admin id

str(diabetic_reduced)
```

```
## 'data.frame':    101766 obs. of  2 variables:
##  $ admission_type_id       : int  6 1 1 1 1 2 3 1 2 3 ...
##  $ discharge_disposition_id: int  25 1 1 1 1 1 1 1 1 3 ...
```

```
100 * sum(complete.cases(diabetic_reduced))/nrow(diabetic_reduced)  # complete ratio
```

```
## [1] 100
```

So now with this particular subset we have ALL of the data value. This means if we wish to draw conclusions about discharge status based on subsets of the data, we can confidently do that. However, all of these catagories, and many of the others, have attribute values that mean the information is missing. If we look at the data key, we see that 5,6 and 8 all mean missing values for admission, while 18, 25, and 26 means missing in discharge. So if we then look at the number of complete cases for these values

```
# find where data is actually missing
complete_col <- !(diabetic_reduced$admission_type_id %in% c(5,6,8)) & !(diabetic_reduced$discharge_dispo
# select only rows with complete data
diabetic_nonna <- diabetic_reduced[complete_col,]
# ratio of complete to incomplete
ratio_comp <- 100*nrow(diabetic_nonna)/nrow(diabetic_reduced)
```

This then shows us that 86.15% of the values from these two catagories are complete, enough certainly to perform some form of analysis.

Conclusion

Based on the source of the data, the fact that it has been reviewed multiple times by large trustworthy institutions, that there is a wealth of information online regarding it, and that we can select subsets of the data that are mostly complete, I believe this data can be used to perdict discharge status.