Homework 9

your name here

Due Wed, May 4, 2022 at 11:00 PM in D2L

Instructions

Use this .Rmd file as a template for your homework. Please use D2L to turn in both the Knitted PDF output and your R Markdown file. Your .Rmd file should compile on its own if it is downloaded by your instructor.

Data

For this homework, you will use the `NHANES` data set in the `NHANES` package. You will need to install the `NHANES` package, then type `?NHANES` to view a description of the data set.

Load packages and data

```
library(tidyverse)
library(NHANES) # You may need to install this package
library(skimr)
```

Exercises

1. Examine Data Set

Before any modeling occurs, it is a good idea to examine the data set to get a sense of what the variables in the data set look like. We've seen the `summary` and `str` functions in base R, but another nice function to view a quick summary of a data frame is the `skim` function in the `skimr` library.

Use the `skim` function to print a summary of the data set. Write a few sentences commenting on a few interesting aspects of this data set. (For example, only 88 observations have a value for `HeadCirc` – the rest are missing. Thus, this variable will be of little value when fitting models.)

2. Create Training / Testing Data

Run the code below to separate the `NHANES` data set at random into 70% training and 30% testing. The `set.seed` function ensures that everyone is using the same training and testing data set.

```
set.seed(05042022)
num.obs <- nrow(NHANES)
test.ids <- base::sample(1:num.obs, size=round(num.obs*.3))
test.set <- NHANES[test.ids,]
train.set <- NHANES[(1:num.obs)[!(1:num.obs) %in% test.ids],]
```

3. Supervised Learning

The `NHANES` data set contains a binary variable `SleepTrouble` that indicates whether each person has trouble sleeping. For each of the following models below,

- Build a classifier for `SleepTrouble` using your training data set
- Report its effectiveness on your testing data set
- Make an appropriate visualization of the model
- Interpret the results - what have you learned about people's sleeping habits?

a. Logistic regression
b. Decision tree
c. Random forest

You may use whatever variables you like, except for `SleepHrsNight` and `ID`.

4. Unsupervised Learning

a. Compare and contrast hierarchical clustering with k-means clustering.

b. Describe a situation where clustering could be useful.

Extra Credit

The `nasaweather` package contains data about tropical storms from 1995–2005 in the data set `storms`.

a. Create a scatterplot of these storms with wind speed on the y-axis, pressure on the x-axis, and colored by type of storm.

b. Use a k-means clustering algorithm to cluster these data based on the two variables `wind` and `pressure` with four clusters. Describe and visualize how well your clusters match up to the types of storms given in the data set.

Cite Sources

Remove this text and cite the sources used on your homework here.