hw5

Jenna

10/29/2020

Table of Contents

Q1

With the data set given here:

https://raw.githubusercontent.com/Vincent-Toups/bios611-project1/master/source_data/datasets_26073_33239_weight-height.csv

Repeat your GBM model. Contrast your results with the results for the previous exercise.

```
urlfile="https://raw.githubusercontent.com/Vincent-Toups/bios611-project1/master/source_data/datasets_2
q1=read_csv(url(urlfile))
```

```
## Parsed with column specification:
## cols(
##   Gender = col_character(),
##   Height = col_double(),
##   Weight = col_double()
## )
```

```
q1.analysis=q1%>%mutate(Male=(Gender=="Male"))

set.seed(392)
q1.train=sample(c(TRUE,FALSE),nrow(q1.analysis),rep=TRUE)
q1.test=(!q1.train)
q1.train = q1.analysis[q1.train,]
q1.test = q1.analysis[q1.test,]
gbm_model <- gbm(formula = Male ~ Weight+Height,
                    distribution = "bernoulli",
                    data = q1.train,
                    n.trees = 1000)
gbm_predict = (predict(gbm_model, newdata=q1.test, n.trees=1000,type="response")>0.5)
gbm_results=as.data.frame(cbind(Truth=q1.test$Male,Prediction=gbm_predict)) %>% mutate(correct=(Truth==
table(gbm_results$correct)
```

```
##
## FALSE  TRUE
##   408  4585
```

```
4585/(4585+408)
```

```
## [1] 0.9182856
```

This model with an accuracy of 91.8% is more accurate than my previous model with an accuracy of 49.8%.

Q2

Using the data set available here:

https://github.com/Vincent-Toups/bios611-project1/blob/master/source_data/datasets_38396_60978_charcters_stats.csv

1. Examine the dataset for any irregularities. Make the case for filtering out a subset of rows (or for not doing so).
2. Perform a principal component analysis on the numerical columns of this data. How many components do we need to get 85% of the variation in the data set?
3. Do we need to normalize these columns or not?
4. Is the "total" column really the total of the values in the other columns?
5. Should we have included in in the PCA? What do you expect about the largest principal components and the total column? Remember, a given principal component corresponds to a weighted combination of the original variables.
6. Make a plot of the two largest components. Any insights?

Q3

Use Python/sklearn to perform a TSNE dimensionality reduction (to two dimensions) on the numerical columns from the set above. You'll need lines like this in your Dockerfile:

```
RUN apt update -y && apt install -y python3-pip
RUN pip3 install jupyter jupyterlab
RUN pip3 install numpy pandas sklearn plotnine matplotlib pandasql bokeh
```

Once you've performed the analysis in Python (feel free to use a Python notebook) write the results to a csv file and load them into R. In R, plot the results.

Color each point by the alignment of the associated character. Any insights?

See the aliases file in Lecture 16 for how to launch your Jupyter Lab.

Q4

Reproduce your plot in Python with plotnine (or the library of your choice).

Q5

Using the Caret library, train a GBM model which attempts to predict character alignment. What are the final parameters that caret determines are best for the model.

Hints: you want to use the "train" method with the "gbm" method. Use "repeatedcv" for the characterization method. If this is confusing, don't forget to read the Caret docs.

Q6

A conceptual question: why do we need to characterize our models using strategies like k-fold cross validation? Why can't we just report a single number for the accuracy of our model?

Q7

Describe in words the process of recursive feature elimination.