Prediction assignment

Practice Machine Learning

Fitness Data set

alt-data

26/01/2020

Assignment: The goal of your project is to predict the manner in which they did the exercise. This is the "classe" variable in the training set. You may use any of the other variables to predict with. You should create a report describing how you built your model, how you used cross validation, what you think the expected out of sample error is, and why you made the choices you did. You will also use your prediction model to predict 20 different test cases

Question:

Background
Using devices such as Jawbone Up, Nike FuelBand, and Fitbit it is now possible to collect a large amount of data about personal activity relatively inexpensively. These type of devices are part of the quantified self movement – a group of enthusiasts who take measurements about themselves regularly to improve their health, to find patterns in their behavior, or because they are tech geeks. One thing that people regularly do is quantify how much of a particular activity they do, but they rarely quantify how well they do it. In this project, your goal will be to use data from accelerometers on the belt, forearm, arm, and dumbell of 6 participants. They were asked to perform barbell lifts correctly and incorrectly in 5 different ways. More information is available from the website here: http://web.archive.org/web/20161224072740/http:/groupware.les.inf.puc-rio.br/har (see the section on the Weight Lifting Exercise Dataset).

Data
The training data for this project are available here:
https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv
The test data are available here:
https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv

The data for this project come from this source: http://web.archive.org/web/20161224072740/http:/groupware.les.inf.puc-rio.br/har. If you use the document you create for this class for any purpose please cite them as they have been very generous in allowing their data to be used for this kind of assignment.

Introduction

Using the HAR dataset on Fitness movements the aim of this analysis is to predict the manner in which the subjects did the exercise. * built a model * cross validation *

START: with import

```
#IMPORT DATA READ IT IN//

training<-read.csv('https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv'
)

testing<-read.csv('https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv')

# str(training)

# summary(training)

# both used to explore the data but the output is huge so not included in the html submis
sion.
```

Exploratory analysis

The data is in a number of small time series for each user.

Many of the variables are non alvaible. All the variables whose names contain skew or kurtosis are only present when the New Window variable=yes.

```
#import library
library(parallel)
library(doParallel)
```

```
## Loading required package: foreach
## Loading required package: iterators
```

```
cluster <- makeCluster(detectCores() - 1) # convention to leave 1 core for OS
registerDoParallel(cluster)
fitControl <- trainControl(method = "cv",
                           number = 5,
                           allowParallel = TRUE)
```

Model building and selection

Skew, and kurtosis measures captured the essential features of the time series

```
trainNEW<-subset(training, new_window=="yes")
```

trained a model on the fitness data set gave accuracy of 0.9966667 with mtry=6939 it's important to note that the testing data is not like this (it is new_window=n) so can't be used to predict.

Final selection of data

Since most of the data set was empty, it not worth using. Hence removing is best choice. Choice 3 factors to have no correlation: user name, cvtd_timestamp and new_window - "new_window","X","raw_timestamp_part_1","raw_timestamp_part_2", "num_window".

```
not_emptyCol <- function(x) all(class(x)!="logical")
```

```
testNEW<-select_if(testing, not_emptyCol)

a<-names(testNEW)

a<-a[1:59]

names(testNEW)[60]<-"classe"

trainNEW<-select(training,c(a,"classe"))

trainNEW2<-select(trainNEW,-c("user_name","cvtd_timestamp","new_window","X","raw_timestamp_part_1","raw_timestamp_part_2", "num_window"))

testNEW2<-select(testNEW,-c("user_name","cvtd_timestamp","new_window","X","raw_timestamp_part_1","raw_timestamp_part_2", "num_window"))


modSmall <- caret::train(classe ~ ., method="rf",data=trainNEW2,verbose=FALSE,trControl = fitControl)

modSmall
```

```
## Random Forest

##

## 19622 samples

##    52 predictor

##     5 classes: 'A', 'B', 'C', 'D', 'E'

##

## No pre-processing

## Resampling: Cross-Validated (5 fold)

## Summary of sample sizes: 15699, 15697, 15698, 15697, 15697

## Resampling results across tuning parameters:

##

##   mtry  Accuracy   Kappa

##    2    0.9942410  0.9927148

##   27    0.9944447  0.9929726

##   52    0.9889405  0.9860082

##

## Accuracy was used to select the optimal model using the largest value.

## The final value used for the model was mtry = 27.
```

```
stopCluster(cluster)

registerDoSEQ()
```

OOB estimate of error rate: 0.4%

optimal model of:

mtry Accuracy Kappa
2 0.99 0.99