Are you doing this in right manner?

Anton Mishchuk

9/3/2017

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.

Summary

The goal of your project is to predict the manner in which people do the exercises.

Two prediction model where used: * regression trees ("rpart") * random forest ("rf")

"rpart" shows bad performance with accuracy ~50%. But "rf" shows outstanding accuracy ~99%.

Then the fitted model used to predict results for test data.

Libraries

"dplyr" library was used for preparing data. "caret" package used for traing. I've also used "doMC"" library to train in parallel because "rf" model takes too much time.

```
library(caret)
library(dplyr)
library(doMC)
registerDoMC(cores = 8)
```

Data

I've removed first 7 colunms because they have some descriptive and time data. Also I removed all the column that has NA data:

```
data <- read.csv("pml-training.csv", na.strings=c("","NA","#DIV/0!"))
test_data <- read.csv("pml-testing.csv")

data <- data[, -c(1:7)]
clean_data <- data[, sapply(data, function(x) !any(is.na(x)))]
```

Data were splitted into trainig and validating sets (50% for each)

```
inTrain = createDataPartition(clean_data$classe, p = 0.5)[[1]]
training = clean_data[ inTrain,]
validating = clean_data[-inTrain,]
```

Prediction models

Desicion tree ("rpart") and Random Forests ("rf") methods were used. Seed '123456' was used

```
set.seed(123456)
```

Regression trees

```
modFit <- train(classe ~ ., data = training, method = 'rpart')
pred.valid <- predict(modFit, validating)
rpart.conf <- confusionMatrix(pred.valid, validating$classe)
```

So the accurracy is: 0.5647299

Random forests

```
modFit <- train(classe ~ ., data = training, method = 'rf')
pred.valid <- predict(modFit, validating)
rf.conf <- confusionMatrix(pred.valid, validating$classe)
```

So the accurracy is: 0.988685

Using rf model to predict on test data

```
predict(modFit, test_data)
```

```
##  [1] B A B A A E D B A A B C B A E E A B B B
## Levels: A B C D E
```