Prediction of Activity Levels

S. Schwarzbek

20 March 2015

With so many data collection points
>...it is now possible to collect a large amount of data about personal activity relatively inexpensively. ... 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.

```
library(caret)
```

```
## Loading required package: lattice
## Loading required package: ggplot2
```

```
PMLfull <- read.table(
  "C:\\Users\\steve\\Desktop\\Coursera\\MachLearn\\pml-training.csv",
  header = TRUE, sep = ",", na.strings = "#DIV/0!" )
PMLuseful <- PMLfull[, c(2, 6:11, 37:49, 60:68, 84:86, 102, 113:124, 140, 151:160 )]
PMLuseful$classe <- as.factor(PMLuseful$classe)
PMLuseful$user_name <- as.factor(PMLuseful$user_name)
```

Data

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

___ The data for this project come from this source: 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. ___

Variable Inspection

We've reduced the variable kept in the fits for modeling. Easiest to jettison are those with many NA's.

Analysis

RandomForest is thorough, but takes too long for writing up reports. We reload a saved version for display.

```
#two.mdl <-train(classe ~ ., PMLuseful[, -c(2,3)], method = "rf")
two.mdl <- readRDS("RandomForestResults")
varImp(two.mdl)
```

```
## Loading required package: randomForest
## randomForest 4.6-10
## Type rfNews() to see new features/changes/bug fixes.
```

```
## rf variable importance
##
##   only 20 most important variables shown (out of 57)
##
##                       Overall
## roll_belt              100.00
## pitch_forearm           60.01
## yaw_belt                55.07
## pitch_belt              45.85
## magnet_dumbbell_y       43.95
## magnet_dumbbell_z       43.46
## roll_forearm            42.46
## accel_dumbbell_y        23.96
## accel_forearm_x         18.23
## magnet_dumbbell_x       17.95
## roll_dumbbell           17.46
## magnet_belt_z           16.14
## accel_dumbbell_z        15.32
## accel_belt_z            14.82
## magnet_forearm_z        14.78
## total_accel_dumbbell    14.37
## gyros_belt_z            13.25
## magnet_belt_y           13.10
## yaw_arm                 12.21
## magnet_belt_x           10.81
```

```
predict(two.mdl, PMLtest[, -c(2,3)])
```

```
##  [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
```

Our expected out of bag error is

```
knn.mdl <- knn3(classe ~ ., data = PML, k=5)
knn.pred <- predict(knn.mdl, PMLtest[, -c(2,3)])
knn.pred
```

```
##         A   B C   D   E
##  [1,] 0.0 1.0 0 0.0 0.0
##  [2,] 1.0 0.0 0 0.0 0.0
##  [3,] 0.2 0.8 0 0.0 0.0
##  [4,] 1.0 0.0 0 0.0 0.0
##  [5,] 1.0 0.0 0 0.0 0.0
##  [6,] 0.0 0.0 0 0.0 1.0
##  [7,] 0.0 0.0 0 1.0 0.0
##  [8,] 0.0 1.0 0 0.0 0.0
##  [9,] 1.0 0.0 0 0.0 0.0
## [10,] 1.0 0.0 0 0.0 0.0
## [11,] 0.0 1.0 0 0.0 0.0
## [12,] 0.0 0.0 1 0.0 0.0
## [13,] 0.0 0.6 0 0.4 0.0
## [14,] 1.0 0.0 0 0.0 0.0
```

```
## [15,] 0.0 0.2 0 0.0 0.8
## [16,] 0.0 0.0 0 0.0 1.0
## [17,] 1.0 0.0 0 0.0 0.0
## [18,] 0.0 1.0 0 0.0 0.0
## [19,] 0.0 1.0 0 0.0 0.0
## [20,] 0.0 1.0 0 0.0 0.0
```

Appendix Class Notes

What you should submit

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.

1. Your submission should consist of a link to a Github repo with your R markdown and compiled HTML file describing your analysis. Please constrain the text of the writeup to $<$ 2000 words and the number of figures to be less than 5. It will make it easier for the graders if you submit a repo with a gh-pages branch so the HTML page can be viewed online (and you always want to make it easy on graders :-).
2. You should also apply your machine learning algorithm to the 20 test cases available in the test data above. Please submit your predictions in appropriate format to the programming assignment for automated grading. See the programming assignment for additional details.

Reproducibility

Due to security concerns with the exchange of R code, your code will not be run during the evaluation by your classmates. Please be sure that if they download the repo, they will be able to view the compiled HTML version of your analysis.

Please upload a link to the github repository containing your .Rmd or .md file and your compiled HTML file performing your analysis. BI LinkMathPreviewEdit: Rich

Evaluation/feedback on the above work

Note: this section can only be filled out during the evaluation phase. Has the student submitted a github repo?

Does the submission build a machine learning algorithm to predict activity quality from activity monitors?

To evaluate the HTML file you may have to download the repo and open the compiled HTML document.

Alternatively if they have submitted a repo with a gh-pages branch, you may be able to view the HTML page on the web. If the repo is:

https://github.com/DataScienceSpecialization/courses/tree/master/08_PracticalMachineLearning/001predictionMotivation

then you can view the HTML page here:

http://datasciencespecialization.github.io/courses/08_PracticalMachineLearning/001predictionMotivation/

Do the authors describe what they expect the out of sample error to be and estimate the error appropriately with cross-validation?

Weight Lifting Exercises Dataset

On-body sensing schema

This human activity recognition research has traditionally focused on discriminating between different activities, i.e. to predict "which" activity was performed at a specific point in time (like with the Daily Living Activities dataset above). The approach we propose for the Weight Lifting Exercises dataset is to investigate "how (well)" an activity was performed by the wearer. The "how (well)" investigation has only received little attention so far, even though it potentially provides useful information for a large variety of applications,such as sports training.

In this work (see the paper) we first define quality of execution and investigate three aspects that pertain to qualitative activity recognition: the problem of specifying correct execution, the automatic and robust detection of execution mistakes, and how to provide feedback on the quality of execution to the user. We tried out an on-body sensing approach (dataset here), but also an "ambient sensing approach" (by using Microsoft Kinect - dataset still unavailable)

Six young health participants were asked to perform one set of 10 repetitions of the Unilateral Dumbbell Biceps Curl in five different fashions: exactly according to the specification (Class A), throwing the elbows to the front (Class B), lifting the dumbbell only halfway (Class C), lowering the dumbbell only halfway (Class D) and throwing the hips to the front (Class E).

Class A corresponds to the specified execution of the exercise, while the other 4 classes correspond to common mistakes. Participants were supervised by an experienced weight lifter to make sure the execution complied to the manner they were supposed to simulate. The exercises were performed by six male participants aged between 20-28 years, with little weight lifting experience. We made sure that all participants could easily simulate the mistakes in a safe and controlled manner by using a relatively light dumbbell (1.25kg).

Read more: http://groupware.les.inf.puc-rio.br/har#ixzz3PJ8LvOf1