Prediction Assignment Writeup for JHU/Coursera Practical Machine Learning Class

dandybits

Synopsis: This project involves analyzing of the Weight Lifting Exercises (WLE) Dataset. This document describes the approach for building classification model that allows to distinguish properly conducted weight lifting excercise movements from those conducted with common mistakes.

This research was conducted as a test assignment for the Data Science Certification on Coursera.. The code for this assignment is available on Github

For more information about the collection and the original analyis of the WLE dataset see research article Qualitative Activity Recognition of Weight Lifting Exercises by Velloso, E.; Bulling, A.; Gellersen, H.; Ugulino, W.; Fuks, H.

Exploring the WLE dataset The first step in data analysis is loading the data and converting it into a data.table format that makes available a richer data processing capabilities than default dataframe format.

```r
## load libraries
library(lattice)
library(ggplot2)
library(caret)
library(rpart)

## load data
wle.data <- read.csv("data\\pml-training.csv")
wle.grade <- read.csv("data\\pml-testing.csv")
```

Observation notes on WLE dataset As described in the paper referenced above as well as revealed in exploring the data, the dataset contains records with various levels of granularity. There are ‘timestamp’-level records that contain a set of sensor measurements as well as ‘summary’ records that contained averaged measurements for time windows of several sizes from 0.5 sec to 2.5 sec. This makes the task somewhat ambiguous. We are trying to predict if a record belongs to a properly exectuted movement while any meaningful classification only applies at the level of the entire set of records for a particular movement.

Moreover, since the surrogate identifier for the movement, num_window attribute, is present in the test data set, it is possible to predict based on the num_window attribute alone.

While this may seem trivial, similar ‘over_inclusive’ datasets occasionally caused unintended results even in high-profile ML competitions.

```r
## splitting data for model validation
set.seed(130265)
inTrain <- createDataPartition(wle.data$classe, p = 0.7, list = FALSE)
wle.train <- wle.data[inTrain,]
wle.test <- wle.data[-inTrain,]
fit.winonly.rpart <- rpart(classe ~ num_window, data=wle.train, method = "class", cp=0.0025)
```

Predicting based on window_num only

The aboove approach gave 100% accurate results on the prediction quiz.

```r
predict(fit.winonly.rpart, wle.grade, type = "class")
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 
##  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
```

Predicting based on meaningful predictors

Conclusions