Quality Predictions of Weight Lifting Exercises

Richard R.

January 30, 2016

Overview

This is an exploration of the Weight Lifting Exercise (WLE) data set to determine if accerlerometer data can help quantify how well partipants did an exercise. The variable `classe` will be used to build models that will attempt to predict if an exercise was done correctly.

WLE Dataset

The original WLE dataset[1] can be found here: Human Activity Recognition -HAR

For the Coursera Practical Machine Learning course the data was then divided into: pml-training and pml-testing

Basic Analysis

Used `str(tndata)` to determine structure of pml-training dataset. It is a data frame of 19622 observations of 160 variables.

```
library(caret)
tndata <- read.csv("./pml-training.csv")
tsdata <- read.csv("./pml-testing.csv")
str(tndata)
tndata <- tndata[,8:160] #Remove first seven columns
summary(tndata)          #Determined NA's summed to 19216
tndata <- tndata[,colSums(is.na(tndata)) != 19216]
tndata <- tndata[,colSums(tndata == "") != 19216]
tndata <- tndata[, -(grep(c("gyros"), names(tndata)))]
tndata <- tndata[, -(grep(c("magnet"), names(tndata)))]
```

Removed first seven columns since they are a "X" index, "user_name", "timestamp", and "window" describing details of the data collection. This was followed by removal of columns that had "NA" and empty strings numbering 19216.These were columns derived from the others such as "averages". Finally, columns that had "gyros" or "magents" were removed to focus on accelerometer data.

Cross Validation

The new 75% training and 25% test set were created to perform 1-fold cross validation. Also the seed was set to 2016.

```
set.seed(2016)
inTrain = createDataPartition(tndata$classe, p=3/4)[[1]]
training = tndata[inTrain,]
testing = tndata[-inTrain,]
```

Model Building

Exploring the correlation between `classe` and the remaining 29 variables.

Random Forrest (RF)

Using internal k-fold cross validation of 3.

```
modRF <- train(classe~., method="rf", data=training, trControl = trainControl(method="cv",number=3))
predRF <-predict(modRF,testing)
confusionMatrix(testing$classe, predRF)$overall['Accuracy']
```

Gradient Boosting Machine (GBM)

No traincontrol implemented. Data is already subsetted for cross validation.

```
modGBM <- train(classe~., method="gbm", data=training)
predGBM <-predict(modGBM,testing)
confusionMatrix(testing$classe, predGBM)$overall['Accuracy']
```

Linear Discriminant Analysis (LDA)

No traincontrol implemented. Data is already subsetted for cross validation.

```
modLDA <- train(classe~., method="lda", data=training)
predLDA <-predict(modLDA,testing)
confusionMatrix(testing$classe, predLDA)$overall['Accuracy']
```

Recursive Partitioning and Regression Trees (RPART)

No traincontrol implemented. Data is already subsetted for cross validation.

```
modRPART <- train(classe~., method="rpart", data=training)
predRPART <-predict(modRPART,testing)
confusionMatrix(testing$classe, predRPART)$overall['Accuracy']
```

Model Selection

Using 1-fold cross validation we find the model accuracies to be:

- RF = 0.990
- GBM = 0.952
- LDA = 0.594
- RPART = 0.546

The Random Forest model performed the best at 0.990 (99%). The expected out of sample error rate is 0.01 or **1%** for the final model.

Final Model

A final RF model is generated using the entire `pml-training` dataset.

```
modFinal <- train(classe~., method="rf", data=tndata, trControl = trainControl(method="cv",number=3))
```

Final Prediction

A final prediction using the `pml-testing` dataset.

```
predFinal <-predict(modFinal,tsdata)
predFinal
```

The final predictio is: [1] B A B A A E D B A A B C B A E E A B B B

Levels:

- A - correct exercise
- B - elbows forward
- C - lifting halfway
- D - lowering halfway
- E - hips forward

References

[1]Velloso, E.; Bulling, A.; Gellersen, H.; Ugulino, W.; Fuks, H. Qualitative Activity Recognition of Weight Lifting Exercises. Proceedings of 4th International Conference in Cooperation with SIGCHI (Augmented Human '13) . Stuttgart, Germany: ACM SIGCHI, 2013.