Weight-lifting Exercise Qualitative Prediction

Emmet Murphy

November 23, 2016

Overview

A random forest classification model is fitted to data from Velloso et al. (2013). The model predicts how well a person performs a weight-lifting exercise, with 95% accuracy.

Data preparation

The size of the data and computational limitations drove the design of data partitions between train and validation. We originally split 70/30, but a random forest model required hours to train on available hardware. Therefore we set aside 90% of the data for validation and used the remaining 1,964 samples for training.

```
trainMaster <- read.csv('pml-training.csv')
set.seed(199)
library(caret)
inTrain <- createDataPartition(y=trainMaster$classe,
                               p=0.1, list=FALSE)
train <- trainMaster[inTrain,]
validation <- trainMaster[-inTrain,]
test <- read.csv('pml-testing.csv')
```

Model selection

Data exploration techniques such as pair-wise plots proved not very helpful with this data, given the large number of variables and difficulty in interpreting their influence. Therefore for our model selection we relied heavily on the original paper, which used random forests with bagging. With only random forest using default parameters, the model achieves 95% accuracy (good enough for our purposes).

Data preparation

The data has variables that are not useful as predictors, including:
- The first 7 columns that describe the data, such as name and timestamps
- Variables that have NA whenever `new_window` is "no" (98% of the records)
- Variables for kurtosis, skewness, amplitude, max and min. They are factor variables with many levels, including blanks and "#DIV/0!"

Variables meeting the above criteria are removed, leaving 52 predictors (down from 159).

```
new_window_count <- length(train[train$new_window=='no',1])
trainFeatures <- subset(train, select='classe')
nonPredictors <- c('X', 'user_name', 'raw_timestamp_part_1', 'raw_timestamp_part_2', 'cv
td_timestamp', 'new_window', 'num_window')
removedFeatures <- subset(train, select=nonPredictors)
for (col in colnames(train)) {
    if (col %in% nonPredictors) next
    if (regexpr('^kurtosis_|^skewness_|^amplitude_|^max_|^min_', col)[1] == -1 && sum(i
s.na(train[[col]])) != new_window_count) {
        trainFeatures[[col]] = train[[col]]
    }
    else {
        removedFeatures[[col]] = train[[col]]
    }
}
validationFeatures <- subset(validation, select=colnames(trainFeatures))
```

Fitting the model

After training the random forest model we print the variable importance, for reference.

```
fitRf <- train(classe ~ .,data=trainFeatures,method="rf",prox=TRUE)
#print(fitRf)
varImp(fitRf)
```

```
## rf variable importance
##
##   only 20 most important variables shown (out of 52)
##
##                      Overall
## roll_belt             100.00
## pitch_forearm          61.13
## roll_forearm           48.42
## magnet_dumbbell_z      48.20
## magnet_dumbbell_y      44.76
## yaw_belt               42.83
## pitch_belt             32.18
## roll_dumbbell          23.76
## magnet_dumbbell_x      22.85
## accel_dumbbell_y       22.77
## accel_forearm_x        19.35
## magnet_belt_z          16.64
## magnet_forearm_z       15.48
## magnet_belt_y          14.64
## roll_arm               14.53
## accel_belt_z           13.61
## accel_dumbbell_z       13.15
## total_accel_dumbbell   12.78
## gyros_dumbbell_y       12.49
## yaw_dumbbell           12.06
```

Finally, the prediction and confusion matrix for the validation data suggest we can expect about 95% accuracy for out-of-sample data.

```
#trainPred <- predict.train(fitRf, trainFeatures)
#confusionMatrix(trainPred, trainFeatures$classe)
validationPred <- predict.train(fitRf, validationFeatures)
confusionMatrix(validationPred, validationFeatures$classe)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 4931  174   12    2    0
##          B   27 3129  142    8   43
##          C   29   95 2856  131   18
##          D   12   16   67 2747   23
##          E   23    3    2    6 3162
##
## Overall Statistics
##
##                Accuracy : 0.9528
##                  95% CI : (0.9496, 0.9559)
##     No Information Rate : 0.2844
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9403
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.9819   0.9157   0.9276   0.9492   0.9741
## Specificity            0.9851   0.9846   0.9813   0.9920   0.9976
## Pos Pred Value         0.9633   0.9343   0.9128   0.9588   0.9894
## Neg Pred Value         0.9927   0.9799   0.9847   0.9901   0.9942
## Prevalence             0.2844   0.1935   0.1744   0.1639   0.1838
## Detection Rate         0.2793   0.1772   0.1617   0.1556   0.1791
## Detection Prevalence   0.2899   0.1897   0.1772   0.1622   0.1810
## Balanced Accuracy      0.9835   0.9501   0.9544   0.9706   0.9859
```

References

Velloso, E., A. Bulling, H. Gellersen, W. Ugulino, and H. Fuks. 2013. “Qualitative Activity Recognition of Weight Lifting Exercises.” *Proceedings of 4th International Conference in Cooperation with SIGCHI (Augmented Human ’13)*. ACM SIGCHI. http://groupware.les.inf.puc-rio.br/har#weight_lifting_exercises#ixzz4Qu8t3jbV (http://groupware.les.inf.puc-rio.br/har#weight_lifting_exercises#ixzz4Qu8t3jbV).