Machine Learning Project: Predict Exercise Manner

Executive Summary

In this project, data from accelerometers on the belt, forearm, arm, and dumbell of 6 participants were analyzed to predict how well an exercise was performed. Machine Learning techniques to classify exercise categorgy will be trained on a subset of data and tested on the remaining data in order to minimize both bias and variance.The following algorithms were used:
* k-Nearest Neighbors
* Decision Tree
* Random Forest

The random forest model has proved to be the most robust with an accuracy of aproximately 97% when predicting class of excerise.

Getting Data and Cleaning Data

The following scripts were used to download the data. Additionally, feature selection will be performed, to reduce the scope of the problem, due to the many variables that are available.

1. Load libraries

```
library(tidyverse)
library(caret)
library(randomForest)
library(rpart)
library(skimr)
library(knitr)
```

2. Get Data

```
training_url <-"https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"
test_url <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"

if(!dir.exists("data")){dir.create("data")}
if(!file.exists("data/training.csv")){download.file(training_url, "data/training.csv")}
if(!file.exists("data/test.csv")){download.file(test_url, "data/test.csv")}

raw_training = read.csv("data/training.csv")
raw_test = read.csv("data/test.csv")
```

3. Feature Selection

```
#keep only columns associated with gyros, accel, magnet
gyros <- grep("^gyro",names(raw_training) )
accel <- grep("^accel",names(raw_training) )
magnet <- grep("^magnet",names(raw_training) )
training <- select(raw_training, gyros, accel, magnet, classe)
```

4. Test and Train split

```
#split data into test and training subsets
set.seed(123)
training_sample <- createDataPartition(training$classe, p=.5, list=F)
```

```
testing <- training[-training_sample,]
training <- training[training_sample,]
```

EDA

```
skimmed <- skim_to_wide(training)
#skimmed[, c(1:5, 9:11, 13, 15:16)]
kable(skimmed)
```

Machine Learning

The dependent variable as a 5 level factor making the problem a classification therefore three popular machine
learning algorithms will be used to predict Excercise class

```
#define cross validation for all models
#set number = 5 to save time
trnControl <- trainControl(method = "cv", number=5, returnData=F)

set.seed(123)

mod.knn <- train(classe~ ., data=training, method="knn", trControl=trnControl )
#train model not working correctly
#mod.tree <- train(classe~., data=training, method="rpart", type="class")
mod.tree <- rpart(classe~., data=training, method="class")
mod.rf <- train(classe ~ ., data=training, method="rf", trControl = trnControl, ntree=100, importance=F
```

Model Selection

Choose the model that performs well on both the training and testing data set

```
models <- c("knn", "tree", "random forest")
test_acc <- c(sum(predict(mod.knn, testing) == testing$classe)/nrow(testing),
              sum(predict(mod.tree, testing, type="class") == testing$classe)/nrow(testing),
              sum(predict(mod.rf, testing) == testing$classe)/nrow(testing)
           ) %>% round(4)

train_acc <- c(sum(predict(mod.knn, training) == training$classe)/nrow(training),
               sum(predict(mod.tree, training, type="class") == training$classe)/nrow(training),
               sum(predict(mod.rf, training) == training$classe)/nrow(training)
            ) %>% round(4)
df.models <- data.frame(Model = models, Train = train_acc, Test = test_acc)
kable(df.models %>% arrange(desc(test_acc)))
```

Model	Train	Test
random forest	1.0000	0.9792
knn	0.9213	0.8509
tree	0.5952	0.5785

Conclusion

Based on the model selection parameters it appears the random forest model performs the best at excericse pattern detection.

Predictions

Given the random forest model below table represents the prediction of the unseen data

```r
kable(data.frame(Row = 1:20,User = raw_test$user_name, Prediction = predict(mod.rf, raw_test)))
```

Row	User	Prediction
1	pedro	B
2	jeremy	A
3	jeremy	B
4	adelmo	A
5	eurico	A
6	jeremy	E
7	jeremy	D
8	jeremy	B
9	carlitos	A
10	charles	A
11	carlitos	B
12	jeremy	C
13	eurico	B
14	jeremy	A
15	jeremy	E
16	eurico	E
17	pedro	A
18	carlitos	B
19	pedro	B
20	eurico	B