Final

This is for the final project of practical machine learning coursera.

The training data:

https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv (https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv)

The test data:

https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv (https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv)

Reference:

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.

The paper claimed to use random forest, 10-cross-validation with bagging.

First load libraries:

```
library(ggplot2)
library(reshape2)
library(AppliedPredictiveModeling)
library(caret)
```

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

```
setwd("/Users/xiem1/Desktop/class/practical_machine_learning/")
```

Read in training data:

```
trainingData<-read.csv("pml-training.csv",header = TRUE)
## get the traning data with label
trainingWLE<-trainingData[,setdiff(c(grep('accel',names(trainingData)),ncol(trainingD
ata)),grep('var_',names(trainingData)))]

### OK, testing here
testingData<-read.csv("pml-testing.csv",header = TRUE)
testingWLE<-testingData[,setdiff(c(grep('accel',names(testingData)),ncol(testingData)
),grep('var_',names(testingData)))]
```

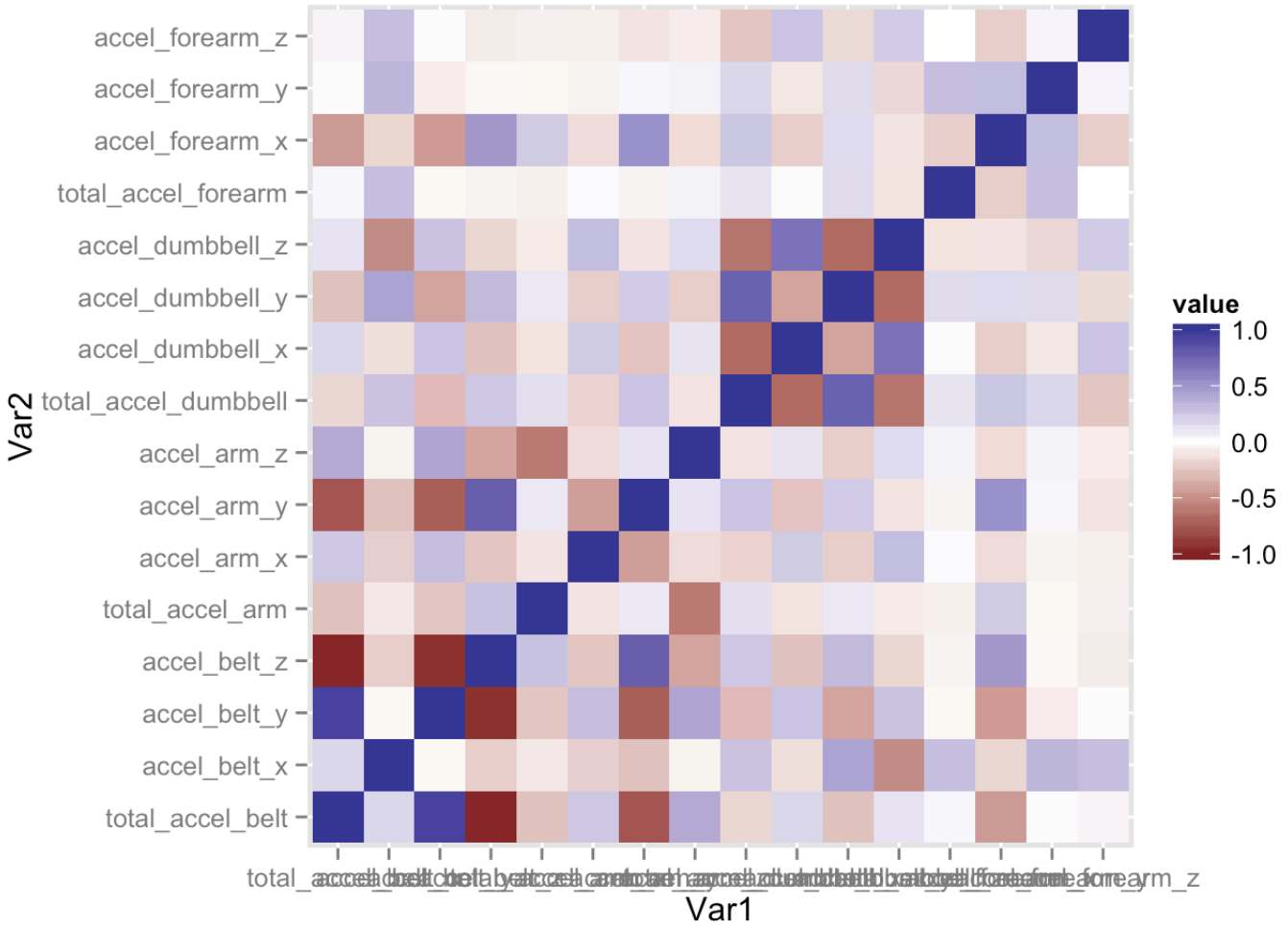
Let's see how data distributed

```
trainingWLE_data<-subset(trainingWLE,select = -classe)
summary(trainingWLE_data)
```

```
##  total_accel_belt  accel_belt_x       accel_belt_y     accel_belt_z
##  Min.   : 0.00    Min.   :-120.000   Min.   :-69.00   Min.   :-275.00
##  1st Qu.: 3.00    1st Qu.: -21.000   1st Qu.:  3.00   1st Qu.:-162.00
##  Median :17.00    Median : -15.000   Median : 35.00   Median :-152.00
##  Mean   :11.31    Mean   :  -5.595   Mean   : 30.15   Mean   : -72.59
##  3rd Qu.:18.00    3rd Qu.:  -5.000   3rd Qu.: 61.00   3rd Qu.:  27.00
##  Max.   :29.00    Max.   :  85.000   Max.   :164.00   Max.   : 105.00
##  total_accel_arm  accel_arm_x       accel_arm_y      accel_arm_z
##  Min.   : 1.00   Min.   :-404.00   Min.   :-318.0   Min.   :-636.00
##  1st Qu.:17.00   1st Qu.:-242.00   1st Qu.: -54.0   1st Qu.:-143.00
##  Median :27.00   Median : -44.00   Median :  14.0   Median : -47.00
##  Mean   :25.51   Mean   : -60.24   Mean   :  32.6   Mean   : -71.25
##  3rd Qu.:33.00   3rd Qu.:  84.00   3rd Qu.: 139.0   3rd Qu.:  23.00
##  Max.   :66.00   Max.   : 437.00   Max.   : 308.0   Max.   : 292.00
##  total_accel_dumbbell accel_dumbbell_x  accel_dumbbell_y
##  Min.   : 0.00        Min.   :-419.00   Min.   :-189.00
##  1st Qu.: 4.00        1st Qu.: -50.00   1st Qu.:  -8.00
##  Median :10.00        Median :  -8.00   Median :  41.50
##  Mean   :13.72        Mean   : -28.62   Mean   :  52.63
##  3rd Qu.:19.00        3rd Qu.:  11.00   3rd Qu.: 111.00
##  Max.   :58.00        Max.   : 235.00   Max.   : 315.00
##  accel_dumbbell_z  total_accel_forearm accel_forearm_x   accel_forearm_y
##  Min.   :-334.00   Min.   :  0.00      Min.   :-498.00   Min.   :-632.0
##  1st Qu.:-142.00   1st Qu.: 29.00      1st Qu.:-178.00   1st Qu.:  57.0
##  Median :  -1.00   Median : 36.00      Median : -57.00   Median : 201.0
##  Mean   : -38.32   Mean   : 34.72      Mean   : -61.65   Mean   : 163.7
##  3rd Qu.:  38.00   3rd Qu.: 41.00      3rd Qu.:  76.00   3rd Qu.: 312.0
##  Max.   : 318.00   Max.   :108.00      Max.   : 477.00   Max.   : 923.0
##  accel_forearm_z
##  Min.   :-446.00
##  1st Qu.:-182.00
##  Median : -39.00
##  Mean   : -55.29
##  3rd Qu.:  26.00
##  Max.   : 291.00
```

To see correlation between features:

```
## Warning: Non Lab interpolation is deprecated
```



Conclude: there is no outstanding correlation between features.

I will use 10 cross validation which is popular for this kind of research.

```
set.seed(222)
fitControl <- trainControl(method = "cv",number = 10)
mod<-train(classe~. , method="rf",data=trainingWLE,trControl = fitControl)
```

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

```
mod
```

```
## Random Forest
##
## 19622 samples
##    16 predictor
##     5 classes: 'A', 'B', 'C', 'D', 'E'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 17661, 17660, 17659, 17661, 17659, 17658, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa      Accuracy SD  Kappa SD
##    2    0.9551523  0.9432585  0.003339126  0.004231470
##    9    0.9502093  0.9370004  0.004653575  0.005893891
##   16    0.9369587  0.9202200  0.006516783  0.008244873
##
## Accuracy was used to select the optimal model using  the largest value.
## The final value used for the model was mtry = 2.
```

```
pred<-predict(mod,testingWLE)
```

My estimation is that my model accuracy will be 95.5% using mtry=2 with SD 4.59e-3.

now here is my prediction

```
pred<-predict(mod,testingWLE)
finalResult<-data.frame(pred,testingWLE[,"problem_id"])
finalResult
```

```
##    pred testingWLE....problem_id..
## 1     B                          1
## 2     A                          2
## 3     C                          3
## 4     A                          4
## 5     A                          5
## 6     E                          6
## 7     D                          7
## 8     B                          8
## 9     A                          9
## 10    A                         10
## 11    B                         11
## 12    C                         12
## 13    B                         13
## 14    A                         14
## 15    E                         15
## 16    E                         16
## 17    A                         17
## 18    B                         18
## 19    B                         19
## 20    B                         20
```