HAR: A Machine Learning Essay on Human Activity Recognition.

Sergio H. S. de Quadros

2016-03-21

Remarks on reproducibility: R version 3.2.4 at *x86-64-pc-linux-gnu (64-bit)* platform running under *Ubuntu 14.04.4 LTS* and using the folowing libraries:

```
library(knitr);library(rmarkdown);library(ggplot2);library(magrittr);library(caret);library(gridExtra);
```

I achievied data set from http://groupware.les.inf.puc-rio.br/har on *the Weight Lifting Exercise Dataset* with:

```
# file names and url
URLtrain <- "http://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"
URLtest <- "http://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"
NAMEtrain <- "pml_training.csv"
NAMEtest <- "pml_testing.csv"
# Create directory
if (!file.exists("./figures")) {
        dir.create("./figures")
}
# Download
if (!file.exists(NAMEtrain)) {
        download.file(URLtrain, destfile=NAMEtrain)
}
if (!file.exists(NAMEtest)) {
        download.file(URLtest, destfile=NAMEtest)
}
# Load data and clean 'NA' & meaningless features
training <- read.csv(NAMEtrain, na.strings = c("NA", ""))
testing <- read.csv(NAMEtest, na.strings = c("NA", ""))
training1 <- training[, colSums(is.na(training)) == 0]
testing1 <- testing[, colSums(is.na(testing)) == 0]
training0 <- training1[, -c(1:7)]
testing0 <- testing1[, -c(1:7)]
```

The training data set has 19622 examples with 159 features in a supervised multiclass problem, the outcome has seven outcome classes: A, B, C, D, E corresponding to following activities: sitting, standing, standing up, sitting down and walking. Our test set has 20 examples.

I selected 52 features by exclusion of *NA* and meaningless ones, then I looked for high covariances between features (PCA?), skeewness(Box Cox transformations?) and distribution in the classes of outcome, but I didn't made those transformations because they wouldn't add accuracy for a non-linear multiclass models. Some figures was uploaded at *figures* file in GitHub site.

So I divided the training set into two subsets and used cross-validation in a 7-k fold:

- 65% for prediction and cross-validation;
- 35% to compute the out-of-sample errors.

```
fitControl <- trainControl(method = "cv", number=7)
set.seed(141593)
inTrain <- createDataPartition(training0$classe, p = 0.65)[[1]]
trainSub <- training0[ inTrain,]
testSub <- training0[-inTrain,]
```

I can approach a multiclass classification problem with logistic regression, SVM, random forest, decision trees, k-nearest neighbors and so on. My first choice was the fast *k-nearest neighbors*. Below we have out-of-sample optimistic assessment for k-nearest neighbors models' accuracies:

```
set.seed(141593)
mod_knn <- train(classe ~.,method="knn",trControl=fitControl,data=trainSub)
pred_knn <- predict(mod_knn,testSub)
predDF <- data.frame(pred_knn,testSub$classe)
uu <- confusionMatrix(pred_knn, testSub$classe)
uu
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 1861   67   15   18   22
##          B   20 1129   47    8   62
##          C   29   54 1092   78   45
##          D   30   35   30 1002   47
##          E   13   43   13   19 1086
##
## Overall Statistics
##
##                Accuracy : 0.8988
##                  95% CI : (0.8914, 0.9058)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8719
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.9529   0.8502   0.9123   0.8907   0.8605
## Specificity            0.9752   0.9753   0.9637   0.9753   0.9843
## Pos Pred Value         0.9385   0.8918   0.8413   0.8759   0.9250
## Neg Pred Value         0.9812   0.9645   0.9811   0.9785   0.9691
## Prevalence             0.2845   0.1934   0.1744   0.1639   0.1838
## Detection Rate         0.2711   0.1645   0.1591   0.1460   0.1582
## Detection Prevalence   0.2889   0.1844   0.1891   0.1666   0.1710
## Balanced Accuracy      0.9640   0.9127   0.9380   0.9330   0.9224
```

I tried another methods: *rpart* , *glm* and *gbm* that the accuracies was about 50%; then I used the *random forest method* - more accurate, but slower. Below we have out-of-sample optimistic assessment for random forest models' accuracies:

```
set.seed(141593)
mod_rforest <- train(classe ~.,method="rf", trControl=fitControl,data=trainSub)
pred_rforest <- predict(mod_rforest,testSub)
predDF2 <- data.frame(pred_rforest,testSub$classe)
vv <- confusionMatrix(pred_rforest, testSub$classe)
vv
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 1951   12    0    0    0
##          B    2 1314   12    0    0
##          C    0    2 1184   26    2
##          D    0    0    1 1099    0
##          E    0    0    0    0 1260
##
## Overall Statistics
##
##                Accuracy : 0.9917
##                  95% CI : (0.9893, 0.9937)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9895
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.9990   0.9895   0.9891   0.9769   0.9984
## Specificity            0.9976   0.9975   0.9947   0.9998   1.0000
## Pos Pred Value         0.9939   0.9895   0.9753   0.9991   1.0000
## Neg Pred Value         0.9996   0.9975   0.9977   0.9955   0.9996
## Prevalence             0.2845   0.1934   0.1744   0.1639   0.1838
## Detection Rate         0.2842   0.1914   0.1725   0.1601   0.1835
## Detection Prevalence   0.2859   0.1934   0.1768   0.1602   0.1835
## Balanced Accuracy      0.9983   0.9935   0.9919   0.9884   0.9992
```

This table summarizes *in* and *out-of-sample* errors by accuracies:

Methods	*In* Accuracy (%)	*Out-sample* Accuracy (%)
k-nearest neighbors	88.15	89.88
Random Forest	99.287	99.17

This figure presents the results in another way:

```
titulo <- "Cross-validation Accuracy"
trellis.par.set(caretTheme())
p3 <- plot(mod_knn)
trellis.par.set(caretTheme())
```

```
p4 <- plot(mod_rforest)
p5 <- ggplot(predDF,aes(x=pred_knn,y=testSub$classe))+geom_jitter(aes(colour=testSub$classe))+xlab("Cla
p6 <- ggplot(predDF2,aes(x=pred_rforest,y=testSub$classe))+
       geom_jitter(aes(colour=testSub$classe))+xlab("Class Predict")+ylab("Actual")+ggtitle("Random Fo
grid.arrange(p3,p4,p5,p6, ncol = 2, nrow = 2)
```

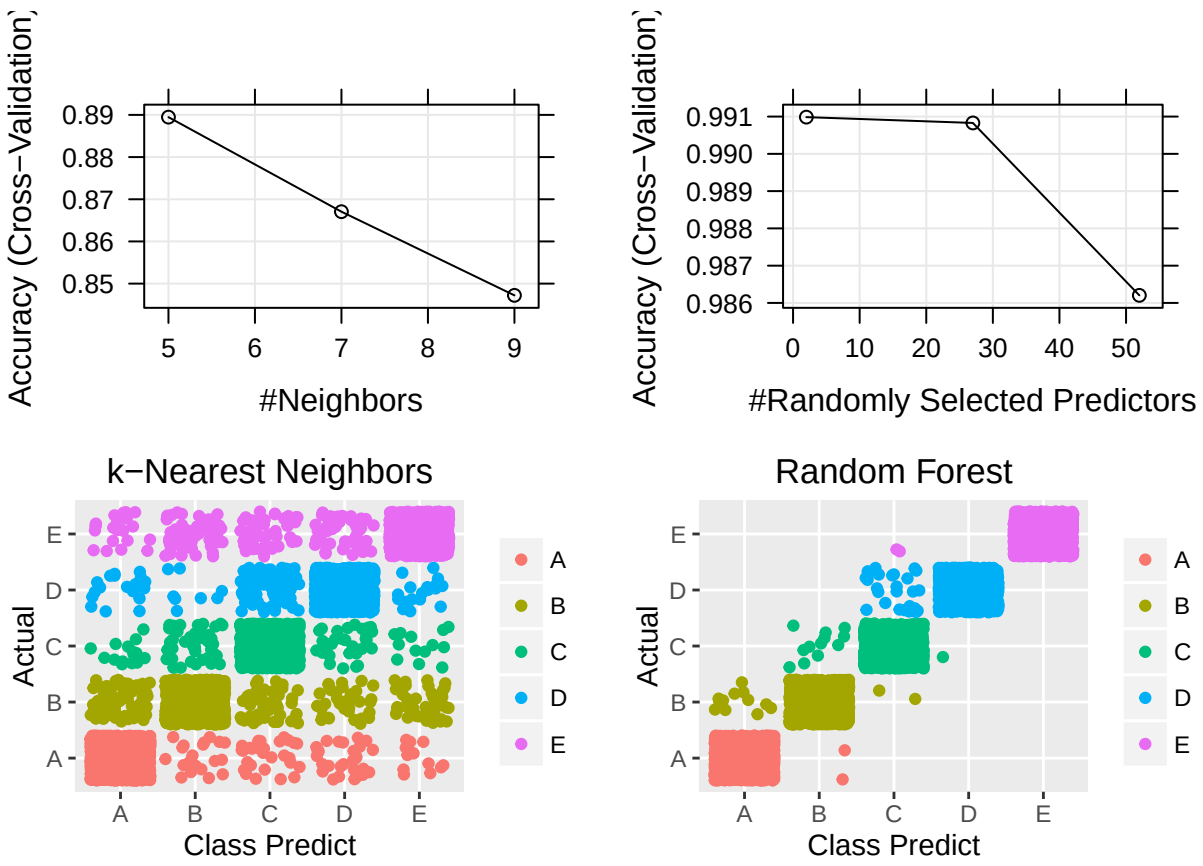

Figure 1 *Top* Both with accuracies by cross-validated k=7 preprocess: *left* k-Nearest Neighbors models; *rigth* random forest. *Bottom* Out-of-sample accuracies: *left* k-Nearest Neighbors models; *rigth* random forest. Random Forest model have better performance than k-Nearest Neighbors ones.

We must choose the *Random Forest model* because it had the best out-of-sample accuracy to made new predictions on testing set and submit the answers at end:

```
# predictions <- predict(mod_rforest, testing0)
# predictAnswers <-  function(x) {
#         n <- length(x)
#         for(i in 1:n) {
#                 filename <- paste0("problem_id_", i, ".txt")
#                 write.table(x[i], file="answer_shsq.txt",
#                     quote=FALSE, row.names=FALSE, col.names=FALSE)
#         }
# }
# predictAnswers(predictions)
```

Bibliography

- 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.

- Ugulino, W.; Cardador, D.; Vega, K.; Velloso, E.; Milidiu, R.; Fuks, H. Wearable Computing: Accelerometers' Data Classification of Body Postures and Movements. Proceedings of 21st Brazilian Symposium on Artificial Intelligence. Advances in Artificial Intelligence - SBIA 2012. In: Lecture Notes in Computer Science. , pp. 52-61. Curitiba, PR: Springer Berlin / Heidelberg, 2012. ISBN 978-3-642-34458-9. DOI: 10.1007/978-3-642-34459-6_6.

- 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.

- Ugulino, W.; Ferreira, M.; Velloso, E.; Fuks, H. Virtual Caregiver: Colaboração de Parentes no Acompanhamento de Idosos. Anais do SBSC 2012, IX Simpósio Brasileiro de Sistemas Colaborativos , pp. 43-48. São Paulo, SP: IEEE, 2012. ISBN 978-0-7695-4890-6.