Prediction Assignment Writeup

Chris Little

06/17/2015

Introduction

This is a writeup illustrating the application of machine learning (specifically 5-fold cross-validated random forests) to data from [Velloso, et al. 2013] in order to predict activity type from sensor data.

Data Clean-up

First, we load the training & testing data from disk.

```
pml_train <- read.csv("./data/pml-training.csv", na.strings = c("", "NA", "#DIV/0!", '""'))
pml_test <- read.csv("./data/pml-testing.csv", na.strings = c("", "NA", "#DIV/0!", '""'))
```

Then, we clean the data frames up by removing irrelevent columns. Columns 1-7 contain non-sensor data.

```
pml_train <- pml_train[, 8:ncol(pml_train)]
pml_test <- pml_test[, 8:ncol(pml_test)]
```

Columns with near zero variance won't be useful as discriminating predictors, so they are removed.

```
nzv_cols <- nzv(pml_train)
pml_train <- pml_train[, -nzv_cols]
pml_test <- pml_test[, -nzv_cols]
```

Columns that are entirely NAs in the test set obviously can't be used for prediction, so they are removed.

```
no_na_cols <- !as.logical(colSums(is.na(pml_test)))
pml_train <- pml_train[,  no_na_cols]
pml_test <- pml_test[, no_na_cols]
```

Model Training and Cross-Validation

Next, using random forests with PCA-preprocessing, a predictor is trained from the training set. All of the sensor data is used to build the model that predicts classe. 5-fold cross-validation is performed by the trainer. (This model takes a long time to train, so it is serialized to disk and only trained from the data if it is absent from disk.)

```
if (!file.exists("rf_fit.rds")) {
  tc <- trainControl(method = "cv", number = 5)
  rf_fit <- train(classe ~ ., data = pml_train,  preProcess = c("pca"),  method = "rf",
                  prox = TRUE, trControl = tc)
  saveRDS(rf_fit, file = "rf_fit.rds")
  rf_fit
} else {
  readRDS(rf_fit, file = "rf_fit.rds")
}
```

```
## Random Forest
##
## 19622 samples
##    52 predictors
##     5 classes: 'A', 'B', 'C', 'D', 'E'
##
## Pre-processing: principal component signal extraction, scaled, centered
## Resampling: Cross-Validated (5 fold)
##
## Summary of sample sizes: 15697, 15697, 15698, 15698, 15698
##
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa      Accuracy SD  Kappa SD
##    2    0.9808889  0.9758226  0.002449964  0.003099801
##   27    0.9712059  0.9635746  0.004327133  0.005475870
##   52    0.9707473  0.9629965  0.003517072  0.004449230
##
## Accuracy was used to select the optimal model using  the largest value.
## The final value used for the model was mtry = 2.
```

With the selected model, accuracy is 0.981 (so we expect an error rate of 0.019) and κ is 0.976.

A cross-validated confusion matrix of the model's predictions illustrates its good fit:

```
confusionMatrix(rf_fit)
```

```
## Cross-Validated (5 fold) Confusion Matrix
##
## (entries are percentages of table totals)
##
##            Reference
## Prediction    A    B    C    D    E
##          A 28.3  0.3  0.0  0.0  0.0
##          B  0.0 18.9  0.2  0.0  0.1
##          C  0.1  0.2 17.1  0.6  0.1
##          D  0.0  0.0  0.1 15.7  0.1
##          E  0.0  0.0  0.0  0.0 18.1
```

References

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. http://groupware.les.inf.puc-rio.br/har