Practical Machine Learning - Course Project

Hari Donthi

Sunday, December 21, 2014

Introduction

I pre-processed the data and applied the Random Forests method for my prediction model.

Pre-Processing

I applied 3 pre-processing steps.

Step 1. Used the nearZeroVar function to retain only non-zero values:

```
nzv <- nearZeroVar(training) training1 <- training[, -nzv]
```

Step 2. Removed variables that had all NAs in the test set. I did not want to do this, but I could not get the predict function to work when a variable in the test set had all NAs:

```
a <- nearZeroVar(testing, saveMetrics = TRUE) nullsInTest <- rownames(subset(a,nzv==TRUE))
training2 <- training1[,!(names(training1) %in% nullsInTest)]
```

Step 3. Removed 3 variables subjectively

```
drops <- c("X","user_name","num_window") training3 <- training2[,!names(training2) %in%
drops]
```

At the end of pre-processing, my training dataset had 55 predictors.

Training

I did a 2-fold cross-validation, 2-times using `method="repeatedcv"`.

“‘ > set.seed(825) > > #rf - removed nullsInTest, 2-fold,2-times, > fitControl <- trainControl(method = “repeatedcv”, number = 2, repeats = 2) > rfFit <- train(classe ~ ., data = training3, method = “rf”, trControl = fitControl) > rfFit Random Forest

19622 samples 55 predictor 5 classes: ‘A’, ‘B’, ‘C’, ‘D’, ‘E’

No pre-processing Resampling: Cross-Validated (2 fold, repeated 2 times)

Summary of sample sizes: 9810, 9812, 9811, 9811

Resampling results across tuning parameters:

```
mtry  Accuracy   Kappa      Accuracy SD   Kappa SD
 2    0.9871828  0.9837831  0.0010796399  0.001366566
37    0.9975028  0.9968414  0.0009434239  0.001193230
73    0.9967893  0.9959389  0.0009615179  0.001216269
```

Accuracy was used to select the optimal model using the largest value. The final value used for the model was mtry = 37. “‘

My model has .99 accuracy. The model was able to predict all the 20 tests correctly.

Acknowledgments

- http://groupware.les.inf.puc-rio.br/har
- Coursera discussion forums - huge help!