Pratical Machine Learning Techniques for Classes classification

Damien Benveniste

20 novembre 2014

Introduction

The pupose of thus study is to design a machine learning model that could differentiate between good and bad forms of Biceps Curl execises based on-body sensing approach paramaters. The different classes follow the following format:

- Class A: exactly according to the specification
- Class B: throwing the elbows to the front
- Class C: lifting the dumbbell only halfwaw
- Class D: lowering the dumbbell only halfway
- Class E: throwing the hips to the front.

The input variables are a set of different accelerometer and gyrometer values recorded all over the body of the six participants.

```
training <- read.csv("pml-training.csv")
testing  <- read.csv("pml-testing.csv")
```

How to build the model

The algorythm

We are going to evaluate the accuracy on such a dataset of the Random Forrest algorythm. We choose this alrythm for few reasons:

- It is fast
- It is accurate for well distributed data
- It does not overfit
- There is no need for cross validation.

In random forests, there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error. Each tree is constructed using a different bootstrap sample from the original data. About one-third of the cases are left out of the bootstrap sample and not used in the construction of the kth tree.

Put each case left out in the construction of the kth tree down the kth tree to get a classification. In this way, a test set classification is obtained for each case in about one-third of the trees. At the end of the run, take j to be the class that got most of the votes every time case n was oob. The proportion of times that j is not equal to the true class of n averaged over all cases is the out-of-bag error estimate. This has proven to be unbiased in many tests.

Random forrest would not accurately classify the data if one or some of the classes were to be represented by a disproportionate number of samples (even with a small out-of-bag error estimate of the computed model). We check that the data is not inbalanced:

```r
summary(training$classe)
```

```
##    A    B    C    D    E
## 5580 3797 3422 3216 3607
```

Cleaning the data

Every learning algorythms speed will depend on the number of features that the algorythms are trying to learning from. In the case of the random forrest algothm the complexity is $(TkMlog(M))$