CWRU DSCI351-451: Project 4 - SamSensor Supervised Machine Learning

Roger H. French, JiQi Liu

02 December, 2018

Contents

4.1.0.1 Semester Project 4: SamSensor Supervised Machine Learning (ML)

Machine learning

- is a powerful tool for data analysis,
- allowing for modeling and prediction
  - of complex data structures.

In this lab we will be using

- the rotational and acceleration data
  - from Samsung Galaxy S3 cell phones
  - from the UCI Machine Learning Repository
  - a YouTube about the data is here
- to predict what activity people holding the phones are doing.

This data contains all of the variables from the phone

- and the "activity" column that shows
  - what activity was being done at each measurement.

The features and features_info files

- can help you understand the variables.

We will be focusing on 3 machine learning methods,

- Regression (or Decision) Tree,
  - using the rpart package
- Random Forest,
  - using the randomForest package
- Support Vector Machine,
  - using the e1071 package.

4.1.0.1.1 Question 1: Supervised and Unsupervised Machine Learning

Explain supervised and unsupervised machine learning in your own words.

ANSWER:

Read in the SamsungData.rda data for this lab,

- how many predictors and responses are there,
- how many different activities are there
- and what are they?

You will need to split this data into training and testing datasets.

- use the caret package to do this splitting
- its a swiss army knife for accessing > 100 ML methods

```r
library(caret)
```

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

```
## Loading required package: ggplot2
```

```r
 ??caret::caret
```

ANSWER:

In your opinion,

- is variable selection necessary for this data set
- and if so can it be reasonably done by a human inferring
    - or assuming which variables are important?

ANSWER:

- Explain why this lab is supervised machine learning and not unsupervised

ANSWER:

4.1.0.1.2 Question 2: Data Cleaning and “-” in Column Names

There have been some changes to R recently

- and “-” can no longer be used in column names
- without causing problems so we have to do some data cleaning

Such is life for data scientists

- Edit the column names to format them better
- Replace all parentheses and commas “(” “)” and “,” with periods “.”
- Replace all dashes “-” with underscores "_"
- Regular expression (regex) can help you do this

Lets build a simple Regression Tree

- to try to predict the activity using the training data

Use the rpart() function from the rpart package

- to build a basic regression tree,
- be sure and keep this model,
- we’ll use it later

```r
library(rpart)
?rpart
```

Plot the model with plot() and text()

- and plot the error per split with the rsq.rpart() function
  - the splits in the trees show the level of branches in the tree

How many splits were there in your regresion tree and

- does this make sense based on the number of activities?

ANSWER:

Explain how someone could estimate the activity using this tree

- if they had unlabled data

ANSWER:

4.1.0.1.3 Question 3: Regression Trees and Random Forest ML

In your own words,

- explain what a random forest is and
- what is its relationship with regression trees
  - like the one you just fitted in Question 2?

ANSWER:

Use the randomForest function

- in the randomForest package
- to fit a model to the training data

```r
library(randomForest)
```

```
## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##
##     margin
```

```r
?randomForest
```

Show your results with the plot() and print() functions on the model,

- explain the meaning of result.

Use the importance() function

- to find the most important variable as determined by the random forest
- what was this variable measuring?

Using ggplot and geom_density,

- plot the distribution of this variable
  - for each activity in one plot,

- can you explain why this variable is important in classification?

Were there any activities

- that the model had more trouble distinguishing,
- if so does is make sense that they were hard to distinguish?

ANSWER:

What was the error of the random forest

- and how did it compare to the regression tree?

ANSWER:

4.1.0.1.4 Question 4: Support Vector Machine ML Model

In your own words

- explain what a support vector machine is.

ANSWER:

Use the e1071 package

- [I don't know why they named it that either][1]
- to build an SVM model on the data set

```
library(e1071)
?e1071
```

```
## No documentation for 'e1071' in specified packages and libraries:
## you could try '??e1071'
```

4.1.0.1.5 Question 5: Training and Testing Your 3 ML Models

Read in the testing data and

- use the predict() function
  - to predict the activity of the test data with
    - your regression tree,
    - your random forest,
    - and your SVM model

Compare the predictions for each activity and

- compare them to the labeled activity in the testing data
  - What is the error of each model?

Which model do you think

- does the best job of predicting the activity?
- explain your decision

ANSWER:

4.1.0.2 Links

[1]: The package authors belonged to vienna university of technology [at that time]. Institut für Statistik und Wahrscheinlichkeitstheorie. Their statistic department has code :e107. And they were in the computational intelligence section of that department, called e1071.

http://www.r-project.org

http://rmarkdown.rstudio.com/

<!– # Keep a complete change log history at bottom of file. # Complete Change Log History # v0.00.00 - 1405-07 - Nick Wheeler made the blank script ##########