Initial clean of Samsung data

Guy Maskall

05 Aug, 2020

Introduction

The name of the game. We've been presented with some activity accelerometer data. We wish to use summary statistics of the acceleration data, for example mean and standard deviation. However we lack confidence that those summary columns have been calculated correctly and this work seeks to validate this. Furthermore, successful validation will enable us to derive these features ourselves on future data.

Load the feature names.

```
data_path <- '../uci_har_dataset/'
feature_names <- read_delim(
    file.path(data_path, 'features.txt'), " ", col_names=c("index", "name")
)
```

```
## Parsed with column specification:
## cols(
##   index = col_double(),
##   name = col_character()
## )
```

```
length(feature_names)
```

```
## [1] 2
```

```
head(feature_names)
```

```
## # A tibble: 6 x 2
##   index name
##   <dbl> <chr>
## 1     1 tBodyAcc-mean()-X
## 2     2 tBodyAcc-mean()-Y
## 3     3 tBodyAcc-mean()-Z
## 4     4 tBodyAcc-std()-X
## 5     5 tBodyAcc-std()-Y
## 6     6 tBodyAcc-std()-Z
```

Some feature names are duplicated:

```
feature_names %>%
    count(name) %>%
    filter(n > 1)
```

```
## # A tibble: 42 x 2
##    name                             n
##    <chr>                        <int>
##  1 fBodyAcc-bandsEnergy()-1,16      3
##  2 fBodyAcc-bandsEnergy()-1,24      3
##  3 fBodyAcc-bandsEnergy()-1,8       3
##  4 fBodyAcc-bandsEnergy()-17,24     3
##  5 fBodyAcc-bandsEnergy()-17,32     3
##  6 fBodyAcc-bandsEnergy()-25,32     3
##  7 fBodyAcc-bandsEnergy()-25,48     3
##  8 fBodyAcc-bandsEnergy()-33,40     3
##  9 fBodyAcc-bandsEnergy()-33,48     3
## 10 fBodyAcc-bandsEnergy()-41,48     3
## # ... with 32 more rows
```

Three duplications. Is it always three? Three for X, Y, Z? Did someone forget to add the axis??

```
feature_names %>%
    count(name) %>%
    filter(!(n %in% c(1, 3)))
```

```
## # A tibble: 0 x 2
## # ... with 2 variables: name <chr>, n <int>
```

Good, feature names only appear either 1 or 3 times. They're either unique or, we suspect, for X, Y, and Z but unlabelled. Still cautious about this.

Load the training data.

X:

```
x_tr_file <- file.path(data_path, 'train', 'X_train.txt')
X_tr <- read_table(x_tr_file, col_names=F)
```

```
## Parsed with column specification:
## cols(
##   .default = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
dim(X_tr)
```

```
## [1] 7352  561
```

Consider the duplications. Let's look at some examples.

```
dups <- which(feature_names$name == "fBodyAcc-bandsEnergy()-1,16")
X_tr[1:10, dups]
```

```
## # A tibble: 10 x 3
##     X311   X325   X339
##    <dbl>  <dbl>  <dbl>
##  1 -1.00 -1.00  -0.995
##  2 -1.00 -0.999 -0.999
##  3 -1.00 -0.999 -0.999
##  4 -1.00 -1.00  -1.00
##  5 -1.00 -1.00  -1.00
##  6 -1.00 -1.00  -1.00
##  7 -1.00 -0.999 -1.00
##  8 -1.00 -0.999 -1.00
##  9 -1.00 -0.999 -1.00
## 10 -1.00 -0.999 -1.00
```

Y:

```r
y_tr_file <- file.path(data_path, 'train', 'y_train.txt')
y_tr <- read_lines(y_tr_file)
length(y_tr)
```

```
## [1] 7352
```