Coursera Data Analysis: Assignment 2

Introduction

The data provided were collected from 21 subjects who held smartphones while engaged in each of six different physical activities (laying down, sitting, standing, walking forward, walking up stairs, or walking down stairs). As they performed these activities, the phones' accelerometers recorded movement in three dimensions over time, and summaries of these measurements were calculated (e.g., mean x-axis acceleration, maximum y-axis acceleration, and so forth). In all, 561 different summary measurements were provided per observation.

Each subject was measured approximately 56 times doing each activity (range 36-95), for a total N of 7,352 observations.

The object of this assignment is to devise a classificatory model that predicts which of the six activities the subject was performing from the various measures, and then assess the validity of that model using held-out data.

Methods

Data from the first four subjects (N = 1,315 total observations) were taken as the subsample from which we devised the predictive model, and the last four subjects (N = 1,485 total observations) were held out to validate the model.

The impractically large number of potential predictor variables led me to make my first task data reduction. If I had a better grasp of the subject matter of the data I might have been able to pare down the list of potential predictors on the basis of theory or prior knowledge—alas, that was not the case.

In order to come up with a reasonably short list of predictors to consider, I wrote a function that automated the process of:

1. Computing a linear model like lm(::potential predictor:: ~ activity, data = testSample).
2. Taking the absolute value of the resulting coefficients for the different factor levels for activity.
3. Storing these coefficients, along w/the name of the ::potential predictor:: in a data frame.

The result was a 561 x 6 data frame with one row per potential predictor, and one column per type of activity like so:

	(Intercept)	activitysitting	activitystanding	activitywalk	activitywalkdown	activitywalkup
tBodyAcc_mean_X	0.25705323	0.010208385	0.023421192	0.021356777	0.030997	0.005923601
tBodyAcc_mean_Y	0.02143523	0.013951408	0.007535796	0.004222005	0.007074	0.007611915
tBodyAcc_mean_Z	0.11376189	0.008403373	0.00725857	0.003179067	0.005872	0.004580008
tBodyAcc_std_X	0.953866	0.023371078	0.023878167	0.648496681	1.10932	0.753962614
tBodyAcc_std_Y	0.92577857	0.024261075	0.012592549	1.002759452	1.020932	1.015836767
tBodyAcc_std_Z	0.92377594	0.007951225	0.003196748	0.542151028	0.614306	0.65997809

Once I had this, I took the list of top 8 predictor variables for each activity by sorting and subsetting, and then took the union of all these sets.[1] This left me with a set of only 23 variables to inspect (because many of the same predictors were related to each activity).

Analysis

Training Data

With the data frame thus reduced to the presumably best individual predictors of activity, I was able to do a k-means cluster analysis on the predictors and see how well the resulting clusters predicted activity in the training set. Because there were six activities to predict I chose 6 clusters. The call I used was:

```
cl <- kmeans(testSample[, vars_to_use], centers = 6, nstart = 100)
```

This resulted in the following clusters (with my interpretations)

cluster	laying	sitting	standing	walk	walkdown	walkup	Interpretation
1	192	0	0	0	0	0	Laying
2	0	1	0	242	34	198	Walking (ambiguous)
3	29	0	0	0	0	0	Laying
4	0	145	151	0	0	0	Standing (ambiguous)
5	0	52	76	0	0	0	
6	0	0	0	24	159	12	Walk down

Creating a function to predict a specific activity on the basis of the cluster score is a simple matter, and results in the following crosstabulation:

predicted	actual					
	laying	sitting	standing	walk	walkdown	walkup
laying	221	0	0	0	0	0
standing	0	197	227	0	0	0
walk	0	1	0	242	34	198
walkdown	0	0	0	24	159	12

The chi-squared statistic (which tests a null hypothesis of no association between the two variables) for this table is 3357, with 15 degrees of freedom, resulting in a p-value of 0. The implication is that the two variables are indeed highly dependent on one another, which is obvious from the frequencies in the table.

This data is visualized in Figure 1.

In terms of overall accuracy in the training sample, the predictions were correct 65% of the time (849 times out of 1315).

[1] This was not my own idea—I read a class forum post where someone proposed doing this and thought it was a good idea. The programming was all my work however.

Validation Sample

To evaluate how this model works in an entirely new sample, we take the cluster centers from our training data model and use them to predict cluster scores for the validation data. This is easily done with the cl_predict() function in the CLUE library.[i] Again translating the cluster numbers into predictions yields the following crosstabulation:

Predicted	Actual					
	laying	sitting	standing	walk	walkdown	walkup
laying	293	0	0	0	0	0
standing	0	264	283	0	0	0
walk	0	0	0	219	51	216
walkdown	0	0	0	10	149	0

Here again the chi-squared test of independence yields a p-value of 0, indicating that the predictions are indeed related to the actual activities. These predictions were correct 64% of the time—944 times out of 1,485 observations.

Conclusions

The model has particular difficulty discriminating between sitting and standing, and between walking forward and walking up stairs. Other than those, it was quite good at predicting laying down, and walking down stairs. While this model is probably useful for a number of purposes, more sophisticated methods (such as CART) may be able to wring more accuracy out of these data.

References

[i] Kurt Hornik (2005). A CLUE for CLUster Ensembles. Journal of Statistical Software 14/12. URL http://www.jstatsoft.org/v14/i12/.