Samsung Data Report

A Human Activity Recognition

Linxiao Bai

Introduction and Data Property

- State of Objective:

  The objective of this project is to use existing features from cell phone gyroscope to recognize and classify human activity to six categories: **walk, walk up, walk down, running, lying and standing**. Although the training data is clearly labeled, and the problem is reduced to supervised classification problem, the primary challenge is to reduce dimensions of features as many as possible while preserving the quality of prediction. This report will illustrate two feature selection method our team experimented, as well as the performance of features on the test data.

- Data Overview:

  The data consists of activity behavior of 30 experiment subjects. Each subject contributes at least 200 observations with all six types of activity. While each record is clearly labeled, the activity behavior is recorded by the cellphone gyroscope forming 561 features. Features are normalized and some are transformation of others. So, high correlation between features are expected. Based on subset of subjects, data is split into training and testing. Training data consists of 21 subjects of a total of 7,352 observations. While testing data consists of 9 subjects of 2,947 observations.

  Figure1 shows count of each type of responses. The result shows that count of response is roughly uniformly distributed. No obvious skewness is observed. Stratified sampling method can be saved.

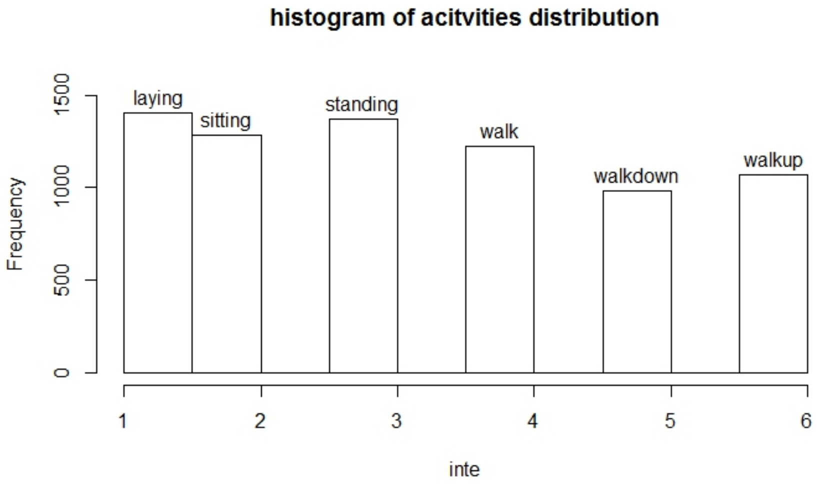


Figure 1. histogram of activity distribution

- Roadmap:

  The roadmap of the projects is to use two different variable selection strategy, Stepwise AIC (forward), and Lasso regression on the training data. After the features are determined, different classifiers are tested based on accuracy on the testing data to verify to performance of feature choices.

Methodology

- Forward AIC

  Stepwise AIC is performed based on the multinomial logistic regression. From the null model without any predictor, features are added one at a step. At each step, predictor yields largest decreases of AIC is added to the model. Fixing model with previous selection, forward selection continues adding the next best predictor to the model. For the sake of model complexity, Forward selection is forced to stop at 7th steps, yield a final model of maximum 7 predictors. Figure2 gives a visualization of AIC decision along each step. Where the feature added at step n is recorded as Table 1.

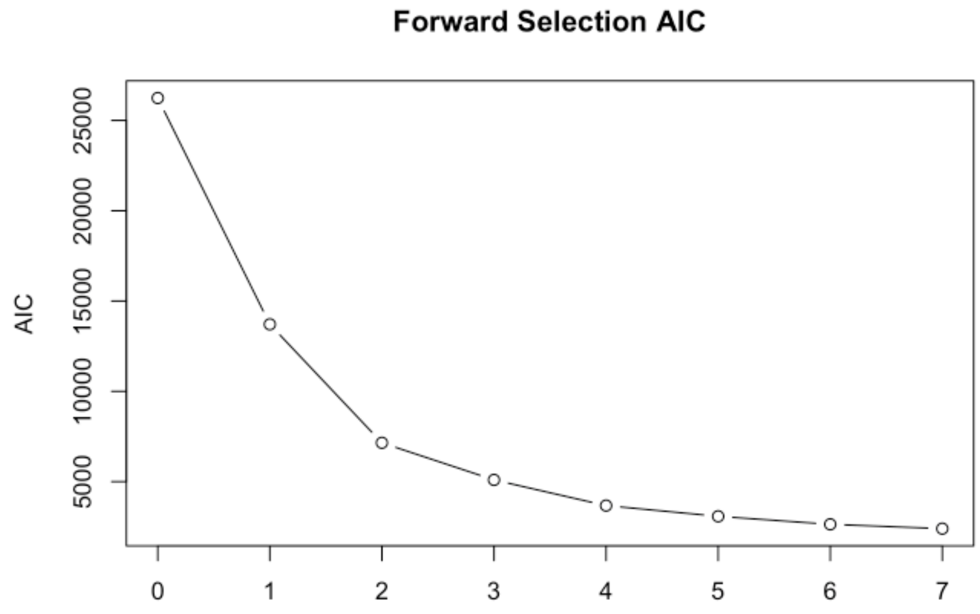


Figure 2. AIC Changes at Each Step

Step	Added Feature
0	Null Model
1	tBodyAcc.max...X
2	tGravityAcc.min...X
3	angle.Y.gravityMean.
4	tGravityAcc.arCoeff...X.1
5	tBodyAcc.correlation...X.Y
6	tGravityAcc.arCoeff...Y.2
7	fBodyGyro.meanFreq...X

Table 1. Features Added at Each Step

The Curve from figure 2 shows great drops of AIC at first 3 steps. This shows the adding of first three variables are most beneficial to the model. Also, slope of the curve flattens as step continues. After step 4, the curve is very flat and adding features are no longer as beneficial as previous steps. Based on this observation, final features are arbitrarily selected as top four features: ***tGravityAcc.min...X, tBodyAcc.max...X, angle.Y.gravityMean., and tGravityAcc.arCoeff...X.1.***

This set of features is then tested on the testing data, using random forest classifier because it is the classifier gives the best performance according to a separate experiment. The experiment result is irrelevant to feature selection process hence will not be shown in this report. Table 2 shows the final confusion table on test data.

Predicted	Ground Truth					
	Laying	sitting	Standing	Walk	Walkdown	walkup
Laying	537	0	0	0	0	0
Sitting	0	356	105	0	0	0
standing	0	135	426	0	0	0
Walk	0	0	0	440	53	98
Walkdown	0	0	0	14	302	35
Walkup	0	0	1	42	65	338

Table 2. Confusion Table for the StepAIC Result.

The overall accuracy is reported to be 0.814. However, looking at the right bottom corner of the confusion table, walking group records, walk, walkdown, and walkup are easily misclassified with each other. This is because this feature set does not provide information to distinguish walking group well.

Further tests of performance of forward selection is performed. Figure 3 shows the trend of testing accuracy against the number of features added into the model by forward selection. Notice that accuracy at 4-predictors is reported to be 0.814 as shown before.

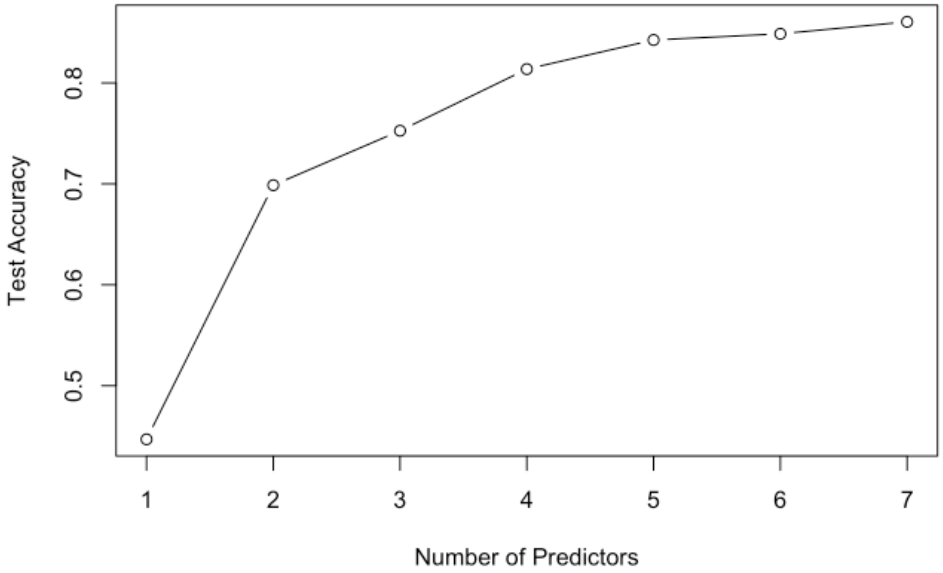


Figure 3. Test Accuracy Against Number of Predictors

Figure 3 also shows that at step 2, 3, 4, 5 the test accuracy is roughly linearily increasing. After step 5, the trend shows tendency to flatten. Accuracy stops the trend of linear increasement. Having observed the testing data, 5-feature-model maybe as good as 4-feature-model.

- Lasso Approach

  The Lasso adds penalized term to non-zero parameters, forcing some features to be “dropped-out” along the way, where severity of panalty is measured as lambda. To proceed with Lasso, selection of lambda is carried out with 10-fold-cross-validation on training data. Figure 4 demonstrates the choice of lambda and its impact on the goodness of fit.

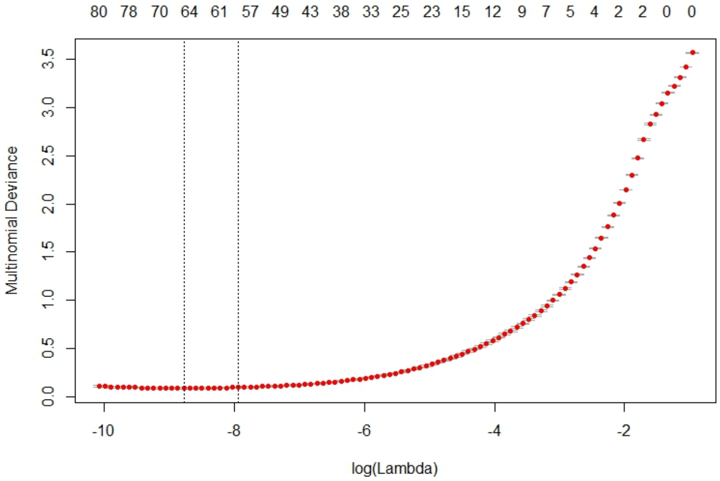
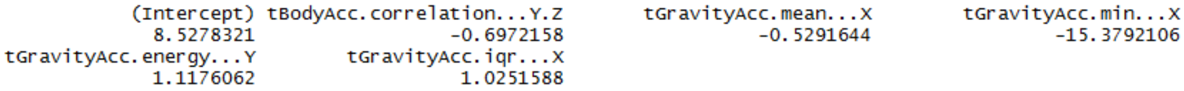


Figure 4. Choice of Lambda Influencing Model

Where the goodness of fit is meausured as Multinomial Deviance, the lower the better. The result of 10-fold-cross-validation shows an optimum selection of log-lambda around -8. This choice of lambda will be used for the final fit of the Lasso model.

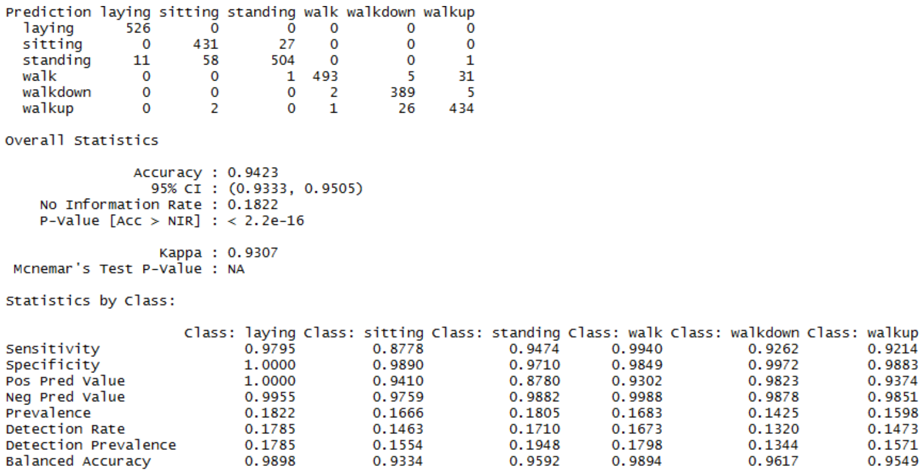
The final Lasso model returns a linear model with 6 non-zero paramerters, that is 5 features are selected. Figure 5 shows each feature and its corresponding parameter.



```
          (Intercept) tBodyAcc.correlation...Y.Z       tGravityAcc.mean...X        tGravityAcc.min...X
            8.5278321                 -0.6972158                 -0.5291644                -15.3792106
 tGravityAcc.energy...Y          tGravityAcc.iqr...X
            1.1176062                  1.0251588
```

Figure 5. Features and Their Parameters

The performace of the model is again tested on the test data. A total of 0.9494 of accuracy is observed. Table 3 shows the detailed confusion table and other goodness of fit measurement of the Lasso approach.



```
Prediction laying sitting standing walk walkdown walkup
  laying      526       0        0    0        0      0
  sitting       0     431       27    0        0      0
  standing     11      58      504    0        0      1
  walk          0       0        1  493        5     31
  walkdown      0       0        0    2      389      5
  walkup        0       2        0    1       26    434

Overall Statistics

               Accuracy : 0.9423
                 95% CI : (0.9333, 0.9505)
    No Information Rate : 0.1822
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9307
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: laying Class: sitting Class: standing Class: walk Class: walkdown Class: walkup
Sensitivity                 0.9795         0.8778          0.9474      0.9940          0.9262        0.9214
Specificity                 1.0000         0.9890          0.9710      0.9849          0.9972        0.9883
Pos Pred Value              1.0000         0.9410          0.8780      0.9302          0.9823        0.9374
Neg Pred Value              0.9955         0.9759          0.9882      0.9988          0.9878        0.9851
Prevalence                  0.1822         0.1666          0.1805      0.1683          0.1425        0.1598
Detection Rate              0.1785         0.1463          0.1710      0.1673          0.1320        0.1473
Detection Prevalence        0.1785         0.1554          0.1948      0.1798          0.1344        0.1571
Balanced Accuracy           0.9898         0.9334          0.9592      0.9894          0.9617        0.9549
```

Table 3. Summary of Goodness of Fit of Lasso Approach.

Conclusion

The two methods this project selected as feature selection yield very different final feature set as well as different performance on test data.

Forward AIC approach yields four features: ***tGravityAcc.min...X, tBodyAcc.max...X, angle.Y.gravityMean., and tGravityAcc.arCoeff...X.1***. Where test accuracy is reported to be 0.814 on a random forest classifier.

On the other hand, the Lasso approach yields 5 features, ***tBodyAcc-correlation()-Y,Z, tGravityAcc-mean()-X, tGravityAcc-min…X, tGravityAcc-energy()-Y, and tGravityAcc-iqr()-X***. The overall accuracy on the test data using lasso model is reported to be 0.9494. This shows a better performance than the previous approach.

Comparing the performance of two approaches, the Lasso-generated 5 features are selected as the result: ***tBodyAcc-correlation()-Y,Z, tGravityAcc-mean()-X, tGravityAcc-min…X, tGravityAcc-energy()-Y, and tGravityAcc-iqr()-X***. With an overall accuracy of 0.9494 on the test.

Reference

1. "An Introduction to Feature Selection." Machine Learning Mastery. October 30, 2016. Accessed February 14, 2017. http://machinelearningmastery.com/an-introduction-to-feature-selection/.
2. Kuhn, Max. "The caret Package." Site not found · GitHub Pages. Accessed February 14, 2017. http://topepo.github.io/caret/pre-processing.html#identifying-correlated-predictors.
3. CRAN - Package glmnet. (n.d.). Retrieved February 14, 2017, from https://cran.r-project.org/web/packages/glmnet/index.html