Human-Activity Recognition Classification

Mini-project

Principles of Machine Learning (CS-4801)

Submitted By :

Kuldeep Singh Bhandari

Third Year Undergraduate

Department of Computer Science

Indian Institute of Technology, Palakkad

1. Problem Statement

→ The experiments have been carried out with a group of 30 volunteers within an age bracket of 19-48 years. Each person performed six activities (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING) wearing a smartphone (Samsung Galaxy S II) on the waist. Using its embedded accelerometer and gyroscope, it has captured 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz. The experiments have been video-recorded to label the data manually. The obtained dataset has been randomly partitioned into two sets, where 70% of the volunteers was selected for generating the training data and 30% the test data. Propose a good multi class classification method.

2. Using Cross-validation set

→ Learning the parameters of a prediction function and testing it on the same data is a methodological mistake: a model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data. This situation is called overfitting. To avoid it, it is common practice when performing a (supervised) machine learning experiment to hold out part of the available data as a Validation Set : X_val and y_val.

3. Model Selection

Here, i am using k-fold technique to split the given Dataset into k-different folds using one fold as Validation Set and rest as Training Set. We have to do this with every fold (all k-folds), that is, keep one fold as Validation Set and rest as Training Set. In each split we will train our model and estimate the correct model parameters by analyzing the accuracy in Validation Set for each parameter in different splits of Dataset. The parameters which performs the best will be chosen for the model predicting the Test Data.

4. Finding best parameters

To find best parameters for a model, we will be doing k-split to the Dataset and for each parameter, we will be calculating its performance on each split and we will take the mean of validation scores obtained in each split and we will measure that score with the best score we have observed so far and if this parameter is better than the best parameter, then mark this parameter as the best parameter. Then, after finding the best parameter, we will find the split which gives best Validation Score for the model with this best parameter. And we will use this model to predict the test dataset.

5. Should PCA be used?

Since, the dataset contains 10299 instances and 561 attributes. The number of attributes is quite high so it might take time to train the model. Also, large number of attributes can lead to Overfitting. If the model overfits because of large number of attributes, we can use Principal Component Analysis (PCA) to reduce the number of attributes so that the model will contain new useful features. On applying PCA in the given dataset with variance 0.9, we observe that around only 63 principal components are required and with variance 0.85, around only 40 principal components are required. But when i trained my model on Logistic, SVM and Random Forest, i observed that the Training Score and Validation Score is not improving, rather it is decreasing. This may be happening because when we are applying PCA with variance 0.85, we may be losing some useful feature and because of that our model is not getting properly trained with PCA. So, in this case, it is better to train the model without PCA so that our model will be properly trained with all the useful features.

We can verify that PCA will not increase the Test Score. Before applying PCA, we need to perform Data Centralization. Data centralization is a major step before applying PCA to a Data Set. Below is a sample output we observed after applying Logistic Regression to the model with PCA and without PCA. The performance of the model without PCA is way better than that with PCA.

Output of Logistic Regression without PCA

Training score by Logistic Regression (liblinear) : 0.9954435931849224

Validation score by Logistic Regression (liblinear) : 0.9392274925705115

Test score by Logistic Regression (liblinear) : 0.9657278588394977

Output of Logistic Regression with PCA

Training score by Logistic Regression (liblinear) : 0.9333514689880305

Validation score by Logistic Regression (liblinear) : 0.9292182182181

Test score by Logistic Regression (liblinear) : 0.9182219205972175

6. Approach

I have used three classification algorithms to perform multi-classification namely

- Logistic Regression
- Support Vector Classifier
- Random Forest Classifier

I have chosen Logistic Regression because on analyzing features, i observed that the data may be linearly separable and Linear Regression works really well when data is linearly separable.

I have chosen Support Vector Classifier with 'rbf' kernel and 'linear' kernel and Validation and Test score were quite similar. It strengthened my belief that the data is linearly separable

because 'rbf' kernel is a non-linear classifier but can also classify linear separable data as well and linear classifier specifically classify linearly separable data.

I have chosen Random Forest Classifier because of its bagging and boosting features. But random forest could not perform well as compared to SVM and Logistic Regression. The potential reason for this could be "Overfitting" as we can see that difference between training score and validation score is quite significant.

7. Performance Measure of Models

Below is the performance of three models on given dataset :

Note : The training score and validation score is the mean score obtained by taking mean of scores in each split during k-fold cross-validation.

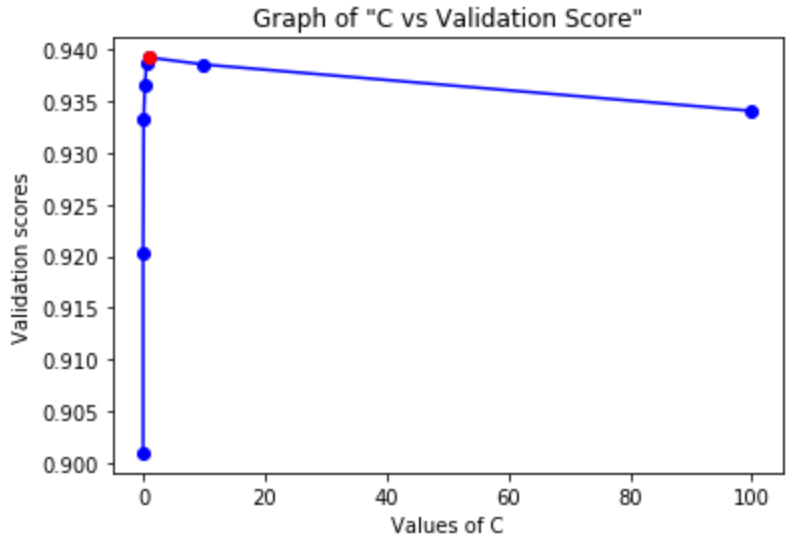
Table of Performance measure of different models

	Logistic Regression	SVM	Random Forest Classifier
Training Score	0.995	0.996	0.981
Cross-Validation Score	0.939	0.932	0.920
Test Score	0.962	0.952	0.911

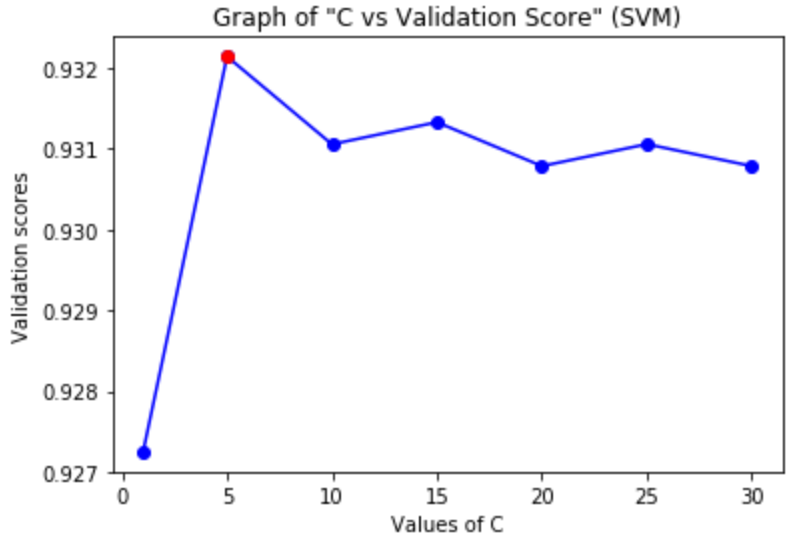
Here , we can observe that Logistic Regression is giving the best Cross-Validation and best Test Score.

Graph for Optimal C

Optimal C for Logistic Regression : 1



Optimal C for SVM : 5



8. Confusion Matrices

Logistic Regression :

We can see that recall and precision both are really good for Logistic Regression. Its precision is not as good as that of SVM but overall F1-score is more. We can see that Logistic Regression finds difficult to predict WALKING DOWNSTAIRS precisely. For WALKING DOWNSTAIRS, it has comparatively bad recall and bad precision.

F1-score : 0.969 (average parameter = 'micro')

Actual / Predicted	WALKING	WALKING UPSTAIRS	WALKING DOWNSTAIRS	SITTING	STANDING	LAYING
WALKING	236	0	8	0	1	0
WALKING UPSTAIRS	0	209	5	0	0	0
WALKING DOWNSTAIRS	6	8	180	3	0	0
SITTING	0	0	0	257	0	0
STANDING	0	0	0	14	261	0
LAYING	0	0	0	0	0	281

SVM :

Its precision is really good but recall for STANDING and SITTING is not that good. Because of that, it is performing slightly less better than Logistic Regression.

F1-score : 0.958 (average parameter = 'micro')

Actual / Predicted	WALKING	WALKING UPSTAIRS	WALKING DOWNSTAIRS	SITTING	STANDING	LAYING
WALKING	245	0	0	0	0	0
WALKING UPSTAIRS	0	214	0	0	0	0
WALKING DOWNSTAIRS	1	0	196	0	0	0
SITTING	0	0	0	223	34	0
STANDING	0	0	0	24	250	0
LAYING	0	0	0	0	2	279

Random Forest Classifier :

We can see that RFC is not able to give good precision or good recall. As we can see from the table below, the precision for WALKING DOWNSTAIRS is not good as it is getting confused with WALKING UPSTAIRS and WALKING DOWNSTAIRS. Recall for WALKING UPSTAIRS is also not good because of the confusion between WALKING UPSTAIRS and WALKING DOWNSTAIRS.

F1-score : 0.938 (average parameter = 'micro')

Actual / Predicted	WALKING	WALKING UPSTAIRS	WALKING DOWNSTAIRS	SITTING	STANDING	LAYING
WALKING	238	3	4	0	0	0
WALKING UPSTAIRS	2	202	8	0	2	0
WALKING DOWNSTAIRS	5	40	152	0	0	0
SITTING	0	0	0	246	11	0
STANDING	0	0	0	15	260	0
LAYING	0	0	0	0	0	281

8. Challenges faced

- **Model Selection** - How to find the optimal parameters for the model ?
- **Using PCA** - Should PCA be used or not?
- **Training Time** - Since, the dataset contains 561 attributes and 10299 instances, training the model takes the time.
- **How to split Dataset** - Which technique to use for splitting the data set ?

9. Conclusion

I have trained three models for Human-Activity Classification Dataset and observed their performances. As we can see that out of three, Logistic Regression and SVM are performing really well and Logistic Regression is performing overall better as it has better Validation Score and also better F1-score. Also, from the experience Logistic Regression gives the similar score in Test Dataset as Validation Dataset. So considering Reliability and Performance of Logistic Regression, i will stick with Logistic Regression as my model for predicting the Test Dataset.