**5104 Project**

Alex, LAM Siu Hung (1006201460)

Janice, CHUNG Wai Tung (1155118014)

Twiggy, WONG Tsz Wing (1004666311)

Kit, LAU Chiu Kit (1155120306)

Frankie, CHAN Yiu Fung (1155010561)

WONG Yiu Chung (1155017920)

**Typo and Question**

129 Principal Component Analysis is a **dimension**

131, 140 ... to reduce the number of **dimension**

170 **REgression**

186 ...control the computational nuances?

189...**valodation**, bias-variance trade-off?

191...**than** that from K-fold

**AUC**

Receiver Operating Characteristic (ROC) curve is the plot of true positive rate (tpr) vs false positive rate (fpr). We can assess the performance of the model by the area under the ROC curve. As a rule of thumb, 0.9-1 = excellent; 0.8-0.9=good; 0.7-0.8=fair; 0.6-0.7=poor; 0.5-0.6=fail.

**Accuracy**

When measuring accuracy of a classification model, we can use F1 score. F1 score is harmonic mean of precision and recall. By using it, we can avoid some misleading situations when using error rate for comparing the performance.

F1=2 x Precision x Recall / (Precision+Recall)

where

Precision = True Positive / (False Postive + True Postive)

Recall = True Positive / (True Positive + False Negative)

**Logarithmic loss**

Logarithmic loss measures the performance of a classification model where the prediction input is a probability value between 0 and 1. The value of Log loss increases when the predicted probability diverges from the actual label.

When the predicted probability approaches 1, log loss slowly decreases. But if the predicted probability decreases, the log loss value would increase rapidly.

<http://wiki.fast.ai/index.php/Log_Loss>

**Overview**

This report explores the Weight Lifting Exercises Dataset and attempt to predict the type of performance based on data from various sensors on the body. The target variable here refers to five different type of performances namely sitting-down, standing-up, standing, walking, and sitting.

**Twiggy:**

**Is the dataset we use is Weight Lifting Exercises Dataset? Not sitting down, standing-up, standing, walking, and sitting. Should be:**

* Exactly according to the specification (Class A),
* throwing the elbows to the front (Class B),
* lifting the dumbbell only halfway (Class C),
* lowering the dumbbell only halfway (Class D)
* and throwing the hips to the front (Class E).

In the dataset, Six young health participants were asked to perform one set of 10 repetitions of the Unilateral Dumbbell Biceps Curl in five different fashions: exactly according to the specification (Class A), throwing the elbows to the front (Class B), lifting the dumbbell only halfway (Class C), lowering the dumbbell only halfway (Class D) and throwing the hips to the front (Class E).

Read more: <http://groupware.les.inf.puc-rio.br/har#ixzz5nRfC1iBV>

**Introduction to the problem**

The awareness of Human Activity Recognition - HAR is increasing among the pervasive computing research community. Some traditional human activity recognition research focused on discriminating between different activities, i.e. to predict "which" activity was performed at a specific point in time. However, the “how (well)’ investigation is yet to broadly covered, while the quality of wearer detection can help for to apply on other real-life area, such as sports training.

In this study, the objective is to investigate "how" an activity was performed by the wearer, which means we want to assess whether the activity recognition techniques could detect mistakes in weight-lifting exercises. In the dataset, six male participants aged between 20-28 years were asked to perform one set of 10 repetitions of the Unilateral Dumbbell Biceps Curl in five different fashions: exactly according to the specification (Class A), throwing the elbows to the front (Class B), lifting the dumbbell only halfway (Class C), lowering the dumbbell only halfway (Class D) and throwing the hips to the front (Class E). Class A corresponds to the specified execution of the exercise, while the other 4 classes correspond to common mistakes.

The data has recorded users performing the same activity correctly and with a set of common mistakes with wearable sensors. And in the following parts, we would use the different data mining tools to test the accuracy of result.

*Ref Bulling, A.; Gellersen, H.; Ugulino, W.; Fuks, H. Qualitative Activity Recognition of Weight Lifting Exercises. Proceedings of 4th Augmented Human (AH) International Conference in cooperation with ACM SIGCHI (Augmented Human'13) . Stuttgart, Germany: ACM SIGCHI, 2013.*

Read more: <http://groupware.les.inf.puc-rio.br/work.jsf?p1=11201#ixzz5nRkO3Lrh>

**Describe the dataset**

**Twiggy: Shall we prepare a table to describe the variables like ah Sir notes (For final 53 variables)? Seems it can explain the dataset more clearly? Any way we can understand each variables refer to…?**

|  |  |  |
| --- | --- | --- |
| **Column** | **Name** | **Description** |
| *1* | *xxx* | *continuous: xxxx* |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Data preparation**

We first combine original training and testing dataset. The combined dataset is then randomized and has 160 variables in total. We then perform data cleaning to the original dataset and now remains 53 variables. In our next step, we sliced 80% of the dataset as training and the remaining 20% as testing.

We then further reduce the dimension of the dataset using Principal Component Analysis (PCA). PCA is a dimension reduction technique. A reduced dataset allows faster processing and smaller storage. In the context of data mining, PCA reduce the number of variables to be used in a model by focusing only on the components accounting for the majority of the variance. Highly correlated variables are also removed as a result of PCA. In this report, PCA is performed on the original dataset to reduce the number of dimension while retaining 99% of the information. Now, 37 variables remain.

**Methods to use**

**Findings, Compare the results**

**Conclusion**

The KNN model is chosen due to its best performance in Log Loss while its performance in Mean F1, Accuracy and AUC are similar among the top three models. The accuracies of KNN as shown in the confusion matrices using training and testing dataset are of high performance with accuracy rates of 0.9883 and 0.975 respectively.

**NB**

Seed(5104) ID=80% Accuracy=50.57%

pr A B C D E

A 320 34 10 0 11

B 82 486 52 20 140

C 544 160 500 221 103

D 145 45 69 305 88

E 43 65 42 66 374

Seed(123) ID=70% Accuracy=72.56%

pr A B C D E

A 450 33 22 0 13

B 137 668 122 37 252

C 798 278 717 334 151

D 238 71 116 464 143

E 41 64 49 140 549