Machine Learning Capstone Project

Definition

Project Overview

The aim of Human Activity Recognition (HAR) is to establish the actions performed by someone given a set of observations about the person and the surrounding environment. This has many uses in healthcare applications, for example monitoring the daily activity of elderly people or the monitoring of activity levels in the overweight and obese. Recognition of activity can be accomplished by retrieving information from various sources such as sensors worn on the body. Dedicated motion sensors on different body parts such as the waist, wrist, chest and thighs achieve good classification performance, however these sensors are usually uncomfortable and do not provide a long-term solution for activity monitoring[5]. Smartphones (with embedded built-in sensors such as accelerometers and gyroscopes) are an alternative solution for HAR. These devices provide a flexible, affordable and self-contained solution to automatically and unobtrusively monitor Activities of Daily Living (ADL).

The database to be used in this project, the Human Activity Recognition database, was created from recording the activities of daily living (ADL) of 30 subjects whose activities were recorded by a waist-mounted smartphone with embedded inertial sensors. The objective of this project is to classify activities recorded in the database into one of six activities (standing, sitting, laying down, walking, walking downstairs and walking upstairs).
I have an undergraduate degree in Exercise and Health Sciences and a Masters in Sports and Exercise Science and therefore have a personal interest in this problem and an understanding of the benefit of accurate activity classification.

Problem Statement

The rapid rise in obesity levels has led to increased clinical and public health interest in effective weight loss programming. The monitoring of physical activity levels using a device has been shown to improve the efficacy and results when following a weight loss programme. The objective of this project is to record the activity levels of subjects using a Samsung Galaxy SII phone and classify activities into one of the six activities performed by processing the recorded data through a Machine Learning (ML) algorithm. This will be done using a Support Vector Machine (SVM) approach which has been used in previous projects to analyse the HAR database.

Metrics

Log loss was used as an evaluation metric which is defined below:

Log loss = $-\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$

Log Loss measures the accuracy of a classification model by giving a probability value of between 1 and 0 for a predicted value. The better the machine model is, the lower the probability value will be with the perfect model having a log loss of 0.

Analysis

Data Exploration

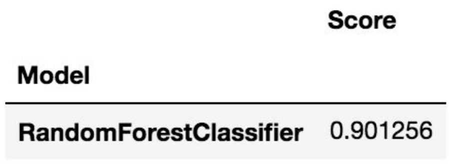
The dataset, train.csv, was used to train the data. In the dataset there are 10299 instances (WALKING = 1226, WALKING_UPSTAIRS = 1073, WALKING_DOWNSTAIRS = 986, SITTING = 1286, STANDING = 1374, LYING = 1407).

Algorithms and Techniques

It was proposed that a support vector machine (SVM) approach was used to analyse the data. This approach has been shown to perform well in similar studies, such as a study by Karantonis et al [7] . where a system provided an accuracy of 90.8% using data collected from 6 volunteers for the classification of 12 ADL using a waist-mounted triaxial accelerometer. A similar study [8] obtained a recognition performance of 93.9% used a chest-mounted accelerometer to classify 5 ADL.

Benchmark

The data was trained and tested on a random forest classifier using the sklearn.ensemble.RandomForestClassifier package from scikit-learn. When the final solution is applied to the data, the results can then be compared to the random forest model (the benchmark model) so an objective comparison can be made to see if the benchmark model is outperformed. Running the random forest classifier model shows an accuracy of 90%.



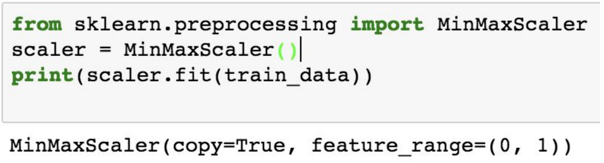
	Score
Model	
RandomForestClassifier	0.901256

There are a number of studies that have used the ADL dataset and have shown an accuracy of 90-96%. The benchmark to be compared against in this project will be taken from the the paper by et Anguits et al 9 . This study showed an overall accuracy of 96% for the test data composed of 2497 patterns. This may be used as a secondary benchmark model.

Methodology

Data Preprocessing

As an SVM is being used, the data was preprocessed and scaled using the min-max scaler sklearn.preprocessing package from scikit-learn. This will translate each feature so that it is within a given range.



```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
print(scaler.fit(train_data))
```

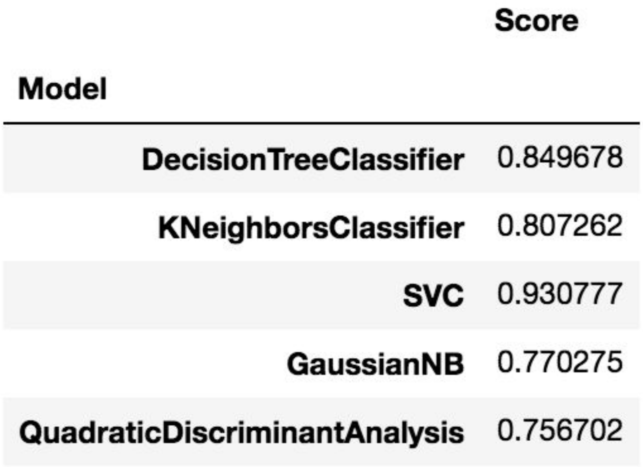
```
MinMaxScaler(copy=True, feature_range=(0, 1))
```

Implementation

A number of algorithms were tested, namely:

- DecisionTreeClassifier
- KNeighborsClassifier
- SVC
- GaussianNB
- QuadraticDiscriminantAnalysis

The following results are as follows and show the correct choice of algorithm to use was the support vector machine as it had the highest accuracy (93%).



	Score
Model	
DecisionTreeClassifier	0.849678
KNeighborsClassifier	0.807262
SVC	0.930777
GaussianNB	0.770275
QuadraticDiscriminantAnalysis	0.756702

Conclusion

Reflection

This was the most interesting, but also the most challenging, of all the projects on the Nanodegree. Interesting in that it was on a subject of my choosing which is of personal interest and relevance to me but challenging in that the project was open and there was a wide number of projects and subject matter which could be chosen which provides its own set of problems.

Improvement

There are a number of algorithms which were not tested which may provide better results. The same applies to the preprocessing of the data. Given more time and experimentation there may well be an improvement to the results and accuracy of predicting physical activity.