Machine Learning Engineer Nanodegree

Capstone Proposal

Mostafa Hamed AbdElmasoud Ali

October 4th, 2019

CLASSIFICATION OF HUMAN MOVEMENTS VIA SMARTPHONE SENSOR DATA

Domain Background

It's no surprise that wearable devices have become so popular in recent years. The ability to track movements and activities throughout the day personalizes the technology and provides individuals with the tools to take control of and manage their own habits. However, because these devices are designed to be small and unobtrusive, the onboard hardware often consists of rudimentary sensors that only capture three dimensional accelerations and tilt at a set frequency. As a result machine learning is a common technique to identify patterns to draw further intelligence from this data and correctly classify resulting activities from these otherwise meaningless signals.

With a few exceptions, up to this point most wearables really only target activity as a single cluster mimicking an advanced pedometer to record differences between if the individual is standing, walking or running. However, activities and fitness have progressed with constantly evolving standards and methods related to the way we keep our bodies healthy. To truly track accurate and categorical progression with these devices it's important to correctly classify a multitude of specific activities as each one may impact the overall fitness results for each individual differently

- https://arxiv.org/abs/1806.05226v3

Problem Statement

Utilizing machine learning, statistical techniques and a data set of labeled signal attributes this report will explore the applicability of using a classification model to accurately predict and differentiate between a set of unique yet similar activities. If successful this information can be applied to a wide variety of wearable devices to bucket and independently track different activities over the course of a period of time

A solution to this problem can result in tremendous impacts and applications in both health and fitness related tracking for a wide range of individuals

Datasets and Inputs

The data set for this particular problem was obtained from the UCI Machine Learning Repository. It was collected via gyroscope and accelerometer sensors onboard a waist mounted Smartphone (Samsung Galaxy II) from a group of 30 volunteers between the ages of 19-48. Each individual performed multiple iterations of 3 static postures (sitting, standing, lying) and 3 dynamic activities (walking, walking upstairs, walking downstairs). Additionally, data was also collected on transitional movements between the static postures: sit-to-lie, lie-to-sit, sit-to-stand, stand-to-sit, lie-to-stand and stand-to-lie. Included with this data set are the original raw tri-axial signals from the accelerometer and gyroscope for each participant as well as a preprocessed feature set created through multi-stage filtering and derivation with respect to time

The preprocessed dataset was already split randomly into training and testing sets with 7761 and 3162 examples, respectively. These values result in approximately 29% of the data reserved for the testing set. Each example is composed of feature vector containing 561 attributes. The 17 main signal variables created from the raw signals, 8 of which have tri-axial .values (denoted by –XYZ), can be seen in Table 1

Variable Name	Variable Description
tBodyAcc-XYZ	Time Domain Body Acceleration
tGravityAcc-XYZ	Time Domain Gravitational Acceleration
tBodyAccJerk-XYZ	Linear Acceleration Derived in Time
tBodyGyro-XYZ	Time Domain Body Angular Velocity
tBodyGyroJerk-XYZ	Angular Velocity Derived in Time
tBodyAccMag	Body Acceleration Magnitude using Euclidean Norm
tGravityAccMag	Gravitational Acceleration Magnitude using Euclidean Norm
tBodyAccJerkMag	Linear Acceleration Magnitude using Euclidean Norm
tBodyGyroMag	Body Angular Velocity Magnitude using Euclidean Norm
tBodyGyroJerkMag	Angular Velocity Magnitude using Euclidean Norm
fBodyAcc-XYZ	Frequency Domain FFT of Body Acceleration
fBodyAccJerk-XYZ	Frequency Domain FFT of Linear Acceleration
fBodyGyro-XYZ	Frequency Domain FFT of Body Angular Velocity

fBodyAccMag	Frequency Domain FFT Euclidean Norm of Body Acc.
fBodyAccJerkMag	Frequency Domain FTT of Euclidean Norm of Linear Acc.
fBodyGyroMag	Frequency Domain FFT of Euclidean Norm of Body Ang. Vel.
fBodyGyroJerkMag	Frequency Domain of FFT of Ang. Velocity Derived in Time

there are significantly less examples for the transition classes. Nothing in particular was done at this point to the dataset to compensate for this class imbalance as it would reduce the size of the dataset too significantly, but it's definitely something to be aware of before proceeding just in case any unexpected results creep up during an analysis phase

Solution Statement

For dimensionality reduction, principle component analysis (PCA) was used

For the machine learning algorithms, four different types were chosen in the early stages of exploration to see how they responded primarily to the PCA reductions. These four algorithms are Linear Support ,Vector Classifier, Gaussian Naïve Bayes Classifier, Support Vector Classifier with an 'rbf' kernel, and finally a Random Forest Classifier

Benchmark Model

As previously mentioned, the algorithms and solutions will be benchmarked against the following metrics: accuracy, training and prediction times, as well as a visual of the confusion matrix to verify if the algorithm is in fact making mistakes they are not serious (precision vs recall). The seriousness of the mistakes will be determined by if the algorithm classifies a mistake as a similar activity or a non-related activity. For example, even if it is only a few mistakes, it would be undesirable to settle on an algorithm that incorrectly classifies "Standing" as "Walking Upstairs." A mistake of this nature could seriously compromise the output data if it were being used for fitness or health tracking. It would appear as if the individual had been exercising from time to time throughout the day when maybe they were just standing still for most of it. On the other hand, if an alternate algorithm only made mistakes by classifying various examples of "Standing" as "Sitting," this might be the better choice even if accuracy is not as .high as the others tested

I will use the following models and compare between F1 score and Accuracy score (choose the best one as my bench model with different no_components of PCA) :

Support Vector Classifiers are a set of machine learning algorithms that attempt to draw linear decision boundaries that maximize the margin between the set of classes.

GuassianNB for its simplicity

Naïve Bayes Classifiers are a family of classifiers built on the principles and foundation of the Bayes Theorem.

Finally, a Random Forest Classifier was used which is an ensemble method of decision trees. Decision trees are an algorithm that attempts to build a set of decision paths (branches) that split the data at multiple levels eventually creating buckets (leaves) that minimize classification error.

Metrics

precision = # of true positives / (# of true positives + # of false positives)

recall = # of true positives / (# of true positives +# of false negatives)

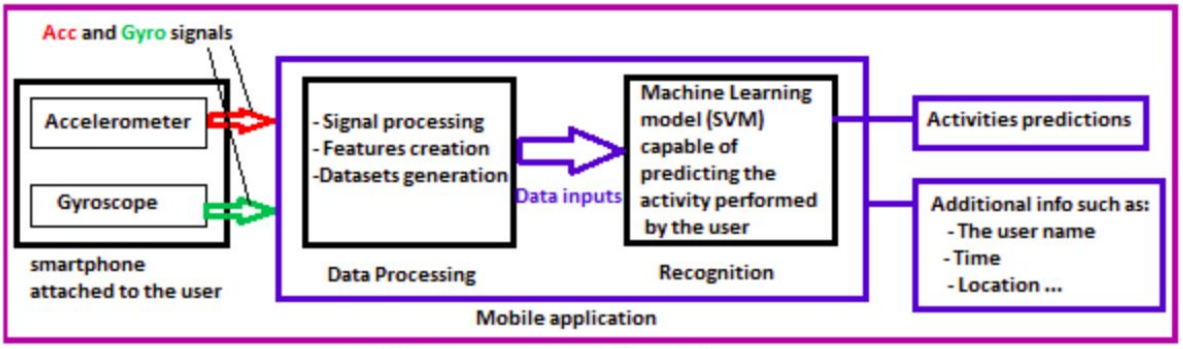
These two pieces of information are critical in classification because the cost of making a mistake may be different depending on what type of mistake it actually is. For example, related to this application, if someone is using a device for fitness tracking while doing a set of pushups the cost of a classification mistake may not be as large if the algorithm incorrectly assigns the activity to bicep curls (which still primarily use the arms) as opposed to lunges (which is more related to the legs). In machine learning the model can never be 100% perfect, so it often makes sense to choose the model that makes the “best mistakes” according to the application.

In addition to the metrics mentioned above, training and prediction times were also recorded for the various models. This can be extremely important when deploying the model to a device with hardware limitations or looking to scale it to larger datasets. Depending on the application, if two models perform similarly in terms of accuracy, training and prediction times can become a good differentiator between which one will be the final solution.

Project Design

To fulfil remote monitoring systems’ requirements Jorge Luis Reyes Ortiz has developed a complete Human Activity Recognition System able to detect and recognize 12 different activities performed by humans in their daily living in online mode using smartphones. The recognition part of his system is based on an SVM model already trained, capable of predicting activities performed by users. Necessary datasets of users’ movements will be collected from smartphone sensors (accelerometer and

gyroscope), processed and then fed to the prediction model to recognize performed .activities



The HAR system working in Online Mode

To build the final model embedded in this HAR system. An offline version of this system needs to be created for two reasons

.To construct the `signal processing pipeline` and process the original data collected .1

.To build the train-test pipelines and test different state of art ML algorithms to choose the optimal one .2

steps:
Exploratory Visualization
PCA Analysis for dimensionality Reduction and Visualization
Some Test Algorithms on a 3-Component PCA Reduction
Algorithm performance comparison with plots
Accuracies vs Times Plot for final mold on PCA vs Entire Featureset

if there will be more small steps and plots than these steps, i will config it for sure in project.

References

.Ensemble Methods." 1.11. Ensemble Methods — Scikit-learn 0.17.1 Documentation .1.11" N.p., n.d. Web. 15 July 2016

.Support Vector Machines." 1.4. Support Vector Machines — Scikit-learn 0.17.1 Documentation .N.p., n.d. Web. 10 July 2016

.Naive Bayes." 1.9. Naive Bayes — Scikit-learn 0.17.1 Documentation .1.9" .1.4" .N.p., n.d. Web. 8 July 2016

Grid Search: Searching for Estimator Parameters." 3.2. Grid Search: Searching for .3.2" Estimator
.Parameters — Scikit-learn 0.17.1 Documentation. N.p., n.d. Web. 18 July 2016

Jorge-L. Reyes-Ortiz, Luca Oneto, Albert Sama, Xavier Parra, Davide Anguita. Transition-Aware Human
Activity Recognition Using Smartphones. Neurocomputing. Springer 2015