Anomaly detection in significant movements

By: Yousef Nami

Prepared for: Mr Juan Jimenez

Table of Contents:

1. Introduction

This document is a guide for reproducing the project in a python environment.

Most of the project research however has been completed in Jupyter notebooks. A jupyter notebooks environment is unfortunately required to access these.

On Atlassian, these are stored as .json files. An add-on can be used to convert the .json into a visual representation of the notebooks.

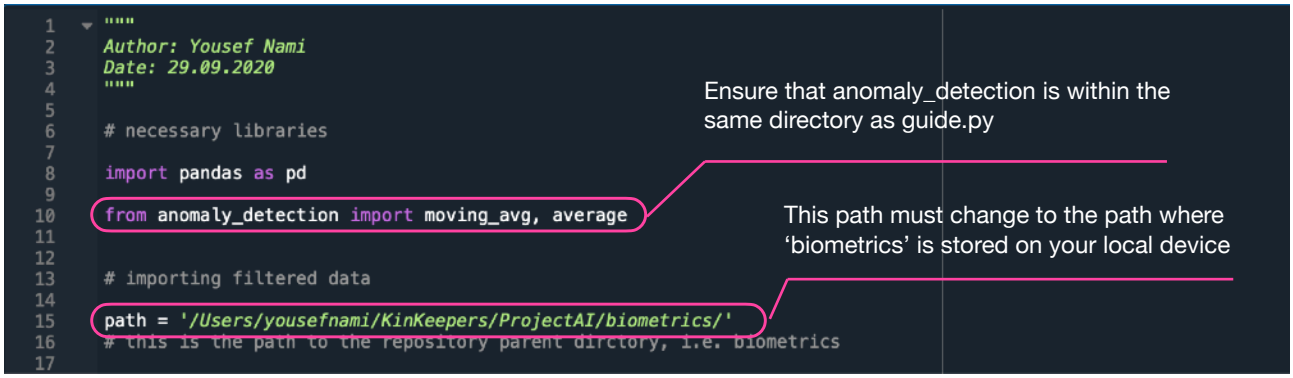
There are a number of guides here, all of which have their own python files. These are:

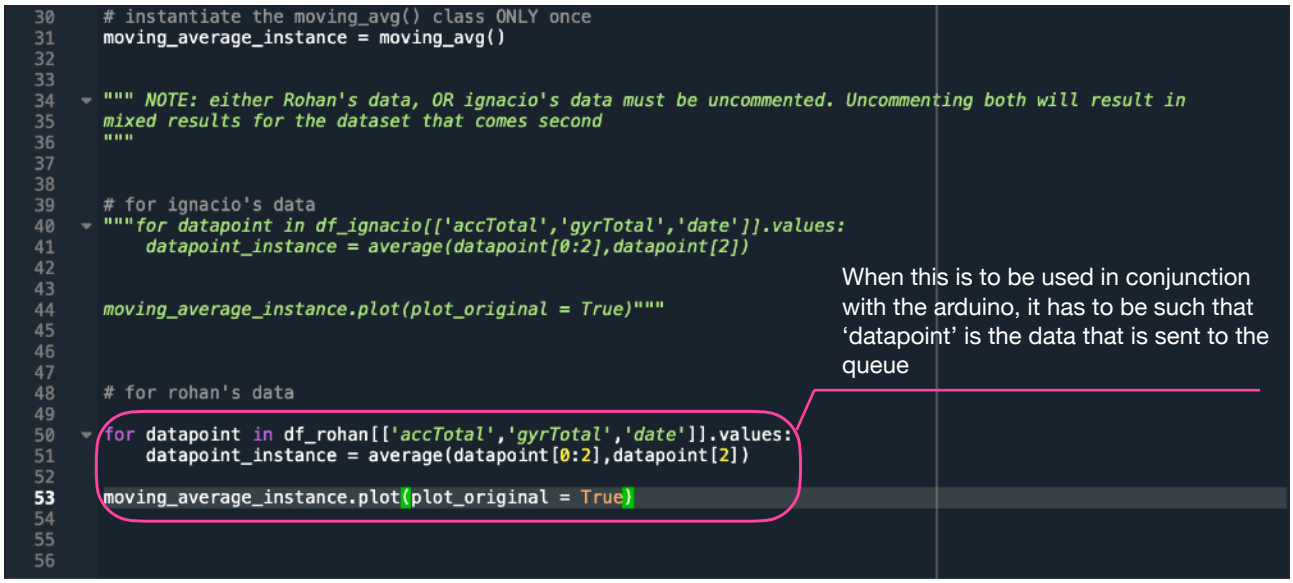
- **guide.py:** a guide to reproducing the moving average graphs. This is dependent on **anomaly_detection.py**. The data sources are: **ignacio_filtered.csv** and **rohan_filtered.csv**.
- **creating_filtered_data.py:** a guide to creating the filtered datasets: **ignacio_filtered.csv** and **rohan_filtered.csv.** This is dependent on file_reading.py. The data sources can be found under **movement_data_analysis/Data/** within the **Rohan** or **Ignacio** sub-directories
- **clustering_guide.py:** a guide to recreating the filter thresholds. This guide is longer, as it includes some preprocessing, evaluating the best model, then plotting the model's outcomes, and then converting then reducing the number of points int he desired outcome to allow curve fitting. It has been added here for completeness. This is dependent on **clustering.py** and **file_reading.py.** The data sources are the same as that for **creating_filtered_data.py,** as well as those under **movement_data_analysis/Data/ TestData**

2. Guide to reproducing work

2.1. guide.py

The guide.py file should be available under 'Python files'. This section highlights parts of the guide.py file that may be unclear or problematic, as well as it's companion files (and data sources).





Data sources:

- **rohan_filtered.csv:** contains all the data generated by Rohan, i.e. under *biometrics/movement_data_analysis/Data/Rohan* after the thresholds have been applied. To be found *under biometrics/movement_data_analysis/Data/Filtered_data*. This can be reproduced by following steps from **creating_filtered_data.csv**
- **ignacio_filtered.csv:** similar to above, except that it's generated by Ignacio's data which is found under *biometrics/movement_data_analysis/Data/Ignacio*

2.2. creating_filtered_data.py

The **creating_filtered_data.py** file should be available under *biometrics/movement_data_analysis/Data/Python\ files*.

The same thing with the paths described under Section 2.1 applies here.

Note that **clustering.py** and **file_reading.py** must also be in the same directory as **creating_filtered_data.py**

```
13  # important libraries
14
15  from file_reading import directory_to_df, special_treatment
16  import numpy as np
17  from clustering import polynomial
18
19  path = ('/Users/yousefnami/KinKeepers/ProjectAI/biometrics/'
20          'movement_data_analysis/Data/') #path to the data folder
21
22  subdir_names = ['Ignacio/','Rohan/'] # note, this must be a list!
23
24  for i in range(len(subdir_names)):
25      subdir_names[i] = path + subdir_names[i]
26
27  # sanity check
28  print(subdir_names)
29
30  df, filenames = directory_to_df(subdir_names)
31  print(filenames) # names of the files that were read
```

Note, to recreate 'rohan_filtered.csv' and 'ignacio_filtered.csv', you must perform the entire operation separately for each item in subdir_names.

In essence, directory_to_df converts all of the files within the list of directories it receives into the same data frame. Note that subdir_names MUST be a list, so even if only Rohan's data is desired, subdir_names = ['Rohan/'] NOT subdir_names = 'Rohan/'

2.3. clustering_guide.py

The **clustering_guide.py** file should be available under *biometrics/movement_data_analysis/Data/Python\ files.*

Once again, be careful of the paths specified in the python file.

```
45  significance = [1,-1,1,-1,1,-1,-1,-1,1] # -1 == sig, 1 == non sig
46  scaler = MinMaxScaler()
47  X_test = []
48  Y_test = []
49
```

Significance is based on knowing the order of the files that within the TestData directory. Whether they are significant or insignificant was determined by the team using votes. A description of this can be found as a comment to ticket IOT-117.

3. Further considerations

Considerations can be found under the summary Jupyter notebooks, or their PDF files under the relevant folders. Paths:

- Anomaly detection summary:

biometrics/movement_data_analysis/Anomaly\ Detection\ in\ Significant\ Movements/

- Clustering summary:

biometrics/movement_data_analysis/Clustering\ movement\ data/