VU VRIJE UNIVERSITEIT AMSTERDAM

Assignment: Machine Learning for the Quantified Self Collecting own data and developing a model

Machine Learning for the Quantified Self 2024

Group: 16
Member names: (Cristian - Augustin Susanu, Filip-Mihai Muntean)
Emails: (c.susanu@student.vu.nl, f.muntean@student.vu.nl)
VUnetIDs: (csu100, mmi349)

June 9, 2024

Contents

1 Question 1

Define a clear research question and the data (what measurements do you want to use, what do you want to predict) you will use to answer the question.

Our aim is to predict the activity type (walking, running, sitting, standing) the user has been doing from the given dataset information by employing supervised learning. In terms of measurements to be used, data would be collected from the following sensors: Accelerometer, Liner Acceleration, Gyroscope, Location, Magnetic Field, Pressure and Proximity sensors.

The collected data would be used for determining the following:

- Accelerometer and Liner Acceleration: Predict activity type (sitting, standing, walking, biking, taking the train).
- Location (GPS) and gyroscope: Use location (GPS) movement to predict walking, running, standing.
- Light Sensor: Use light sensor information to detect if the activity was conducted indoors or outdoors.
- Gyroscope: Use the phone's gyroscope to collect raw data which is a rotation rate in rad/s.
- Linear Accelerometer: Measure the rate of change of velocity of an object in a specific direction.
- Magnetometer: Calculate and displaysthe FFT of magnetometer data.
- Proximity: Use the proximity sensor to measure times. You can determine the duration of close/far states or the time between two close/far states.

The data will be collected for approximately an hour and a half to two hours. Data collection would be done through the Phyphox mobile application, version 1.1.16 run on the IOS operating system. In terms of the frequency, 50/100 Hz would be employed for an accurate measurement of the data. For the accelerometer, linear accelerometer, magnetometer and gyroscope sensors around 108.000 data points would be collected and for the location sensor around 2000 data points would be collected. Based on the data quality, we will decide which features to keep, which features to drop and whether it is needed to do another round of data collection.

Supervised learning would be used with the collected data sets to train the algorithm and predict the type of activity undergone by the user.

2 Question 2

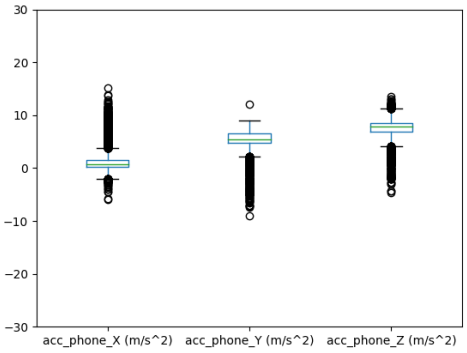
Collect the data and describe a summary of the data you have collected (an exploratory data analysis), similar to Chapter 2 of the book. Discuss choices you make (e.g. resolution with which you process your data).

In terms of issues we have encountered, the Phyphox application has encountered random crashes during recording data, reproduced on both the IOS and Android operating systems. Furthermore, the recording of data was paused when the screen turned off. These matters slowed down the process of data collection. For the later stages of the project, we are considering switching to an application such as Strava, would our data be deemed insufficient.

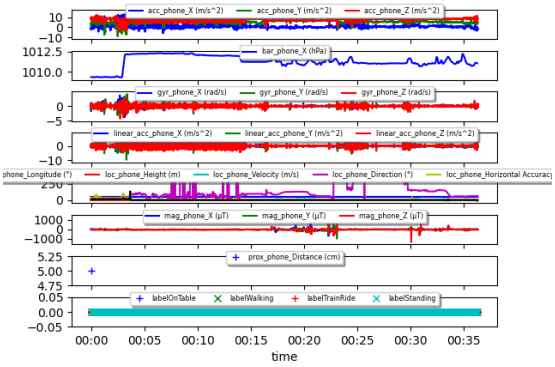
We have done an analysis based on a granularity of 60s and a granularity of 0.25s.

Attribute	% Missing Values	Mean	Std Dev	Min	Max
acc_phone_X (m/s$\hat{2}$)	0.00 / 0.00	1.10 / 1.10	1.54 / 1.78	-1.84 / -5.87	9.65 / 15.22
acc_phone_Y (m/s$\hat{2}$)	0.00 / 0.00	5.41 / 5.41	1.55 / 1.71	-4.72 / -9.02	7.83 / 12.03
acc_phone_Z (m/s$\hat{2}$)	0.00 / 0.00	7.67 / 7.67	1.13 / 1.66	-2.49 / -4.58	9.41 / 13.47
bar_phone_X (hPa)	0.10 / 77.78	1011.30 / 1011.31	0.77 / 0.79	1009.27 / 1009.27	1012.31 / 1012.31
gyr_phone_X (rad/s)	0.00 / 0.00	-0.00 / -0.00	0.08 / 0.27	-1.09 / -3.85	0.67 / 3.64
gyr_phone_Y (rad/s)	0.00 / 0.00	-0.00 / -0.00	0.10 / 0.25	-0.62 / -3.17	0.76 / 4.04
gyr_phone_Z (rad/s)	0.00 / 0.00	-0.01 / -0.01	0.13 / 0.34	-1.14 / -4.29	0.58 / 3.43
linear_acc_phone_X (m/s$\hat{2}$)	0.00 / 0.00	0.02 / 0.02	0.17 / 0.66	-0.55 / -4.91	1.63 / 7.48
linear_acc_phone_Y (m/s$\hat{2}$)	0.00 / 0.00	-0.07 / -0.07	0.19 / 0.70	-2.21 / -9.74	0.73 / 4.68
linear_acc_phone_Z (m/s$\hat{2}$)	0.00 / 0.00	0.02 / 0.02	0.16 / 1.14	-1.02 / -5.05	1.68 / 6.96
loc_phone_Latitude (°)	5.40 / 76.83	52.34 / 52.34	0.00 / 0.00	52.32 / 52.32	52.34 / 52.34
loc_phone_Longitude (°)	5.40 / 76.83	4.90 / 4.90	0.03 / 0.03	4.86 / 4.86	4.96 / 4.96
loc_phone_Height (m)	5.40 / 76.83	3.61 / 3.62	3.11 / 3.13	-8.03 / -8.25	12.42 / 12.48
loc_phone_Velocity (m/s)	10.20 / 77.55	4.61 / 4.61	6.30 / 6.30	0.00 / 0.00	19.46 / 20.12
loc_phone_Direction (°)	10.90 / 77.69	145.19 / 145.57	108.81 / 111.16	0.73 / 0.18	359.41 / 359.95
loc_phone_Horizontal Accuracy (m)	5.40 / 76.83	7.11 / 6.28	9.60 / 7.21	4.71 / 4.71	99.00 / 99.00
loc_phone_Vertical Accuracy (°)	5.40 / 76.83	4.17 / 3.87	4.74 / 4.28	1.14 / 1.13	106.92 / 134.87
mag_phone_X (µT)	0.00 / 0.00	-11.91 / -11.90	56.40 / 63.17	-344.62 / -921.81	445.16 / 787.30
mag_phone_Y (µT)	0.00 / 0.00	-21.25 / -21.24	47.13 / 54.17	-453.75 / -992.12	321.26 / 657.25
mag_phone_Z (µT)	0.00 / 0.00	-31.02 / -31.02	51.48 / 61.13	-467.32 / -1351.68	436.60 / 1271.68
prox_phone_Distance (cm)	99.90 / 99.99	5.00 / 5.00	nan / nan	5.00 / 5.00	5.00 / 5.00
label_OnTable	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00
label_Walking	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00
label_TrainRide	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00
label_Standing	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00

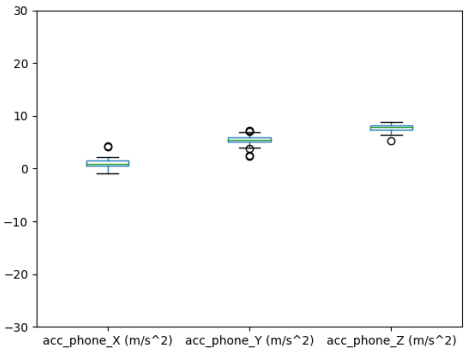
Table 1: Statistics of our Dataset



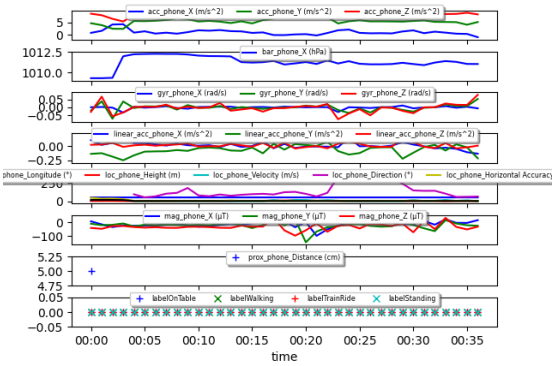
(a) Accelerometer Features with $\Delta t = 0.25$s



(b) TotalFeatures with $\Delta t = 0.25$s



(c) Accelerometer Features with $\Delta t = 0.25$s



(d) TotalFeatures with $\Delta t = 60$s

Figure 1: A figure with four subfigures

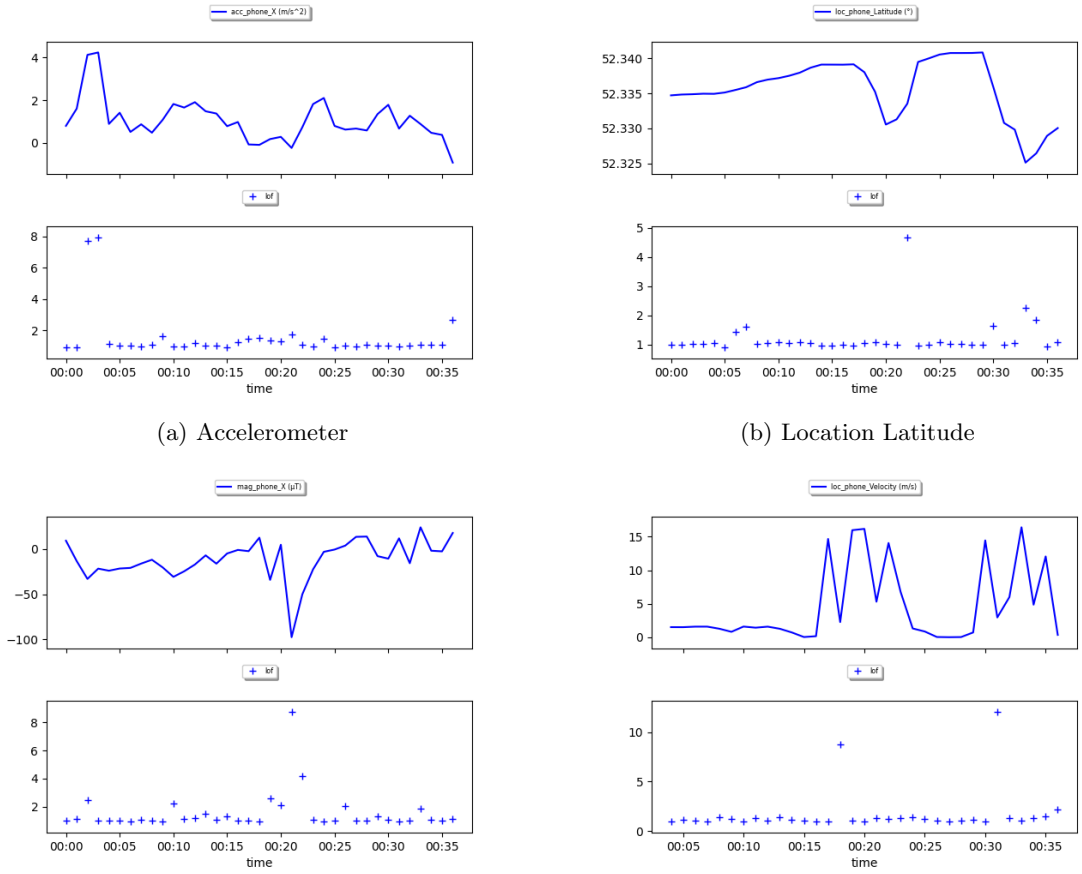
Details of the whole dataset are present in figures 2d 2b. Data was gathered in different movement states, which can be seen by the accelerometer, showing high variability, while the gyroscope reports quite low variability. On the contrary, the magnetometer reports some high variability, which translates to the fact that the individual has changed environments, usually due to going on a walk or running. There was a sensor malfunction which can be seen through the proximity data, almost entirely missing. We have opted for showcasing these statistics, since such a table was presented in the

3 Question 3

Remove noise and handle missing values using an appropriate technique (i.e. those discussed in Chapter 3 of the book). Again, discuss the choices you make and provide a good rationale.

Based on the statistics table most of our data does not contain missing values. However, there are several missing values present in the barometer feature (`bar_phone_X`, as well as in the columns

designating the Location variable, namely `loc_phone_*`, while 99.99% of our values are missing in the `prox_phone_Distance` variable. Since most of the data within the Proximity feature is missing, we have decided to exclude it from the rest of the preprocessing. For now, we have experimented with locating outliers based on the LOF method, with $K = 5$.



(a) Accelerometer

(b) Location Latitude

(c) Magnetometer

(d) Location Longitude

Figure 2: Outlier detection of several features with $\Delta t = 60$s

Imputing of missing values was done using the PCA method. After preprocessing, our data looks as such:

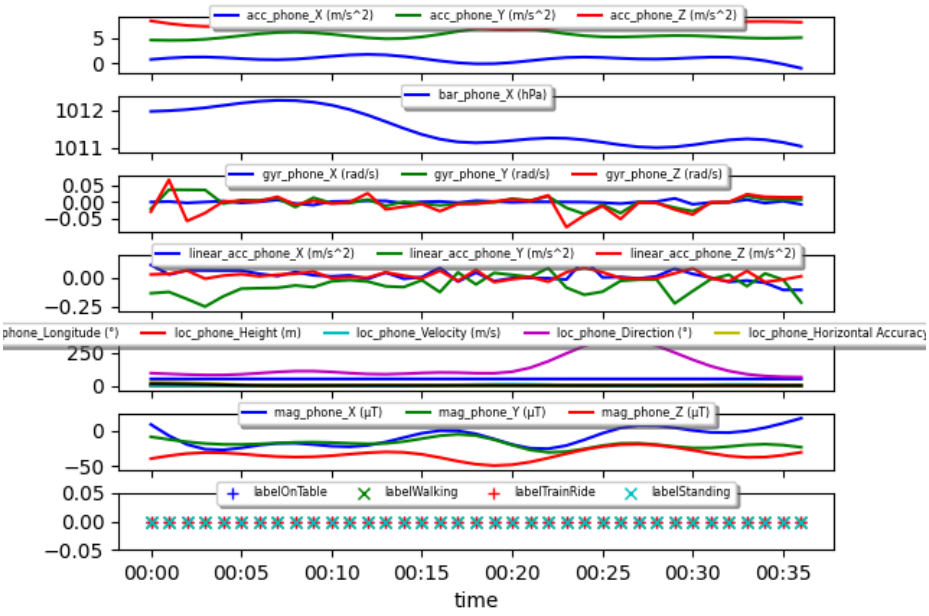


Figure 3: Total Features After Preprocessing with $\Delta t = 60$s