ML4QS - Assignment 1

Group 62

Sander Brinkhuijsen[2685424], Tommaso Castiglione Ferrari[2673807], and Simone Korteling[2671463]

Vrije Universiteit, Amsterdam

Chapter 2

Theoretical part

Question 1
The substantial difference in sensory values between multiple users can be caused by a variety of factors [1], we will discuss some of them here.

Subject variability First of all, different users have different bodies and psychological characteristics. One person might be in very good shape causing a lower heart rate while another person is in bad shape causing a higher heart rate. This has consequences for the generated sensory data.

Measurement errors Second, since every sensor is unique, data might systematically differ from sensors used by other users. It is impossible to create perfectly identical sensors, there always will be some discrepancies that arise due to sensor manufacturing or product usage.

Environmental influences A third factor is the environment of the subject in which the sensory data was gathered. The environment directly influences the sensors and the reaction of users' bodies. For example, measurements in a cold environment are very different from measurements in a warm environment. Also temporal factors such as the time of day, or (previous) activity of the subject might influence the sensory data.

Question 2
We will briefly discuss four criteria that play a role in deciding on the granularity (step size) for the measurements of a data set.

The first criteria for deciding granularity is the level of detail that is required for the task at hand. The example in the book shows that a higher granularity complicates the detection of the motion state because the data points are smoothed too much (i.e. too much data is lost) [2]. Therefore, when you want to measure a feature that has a short duration, such as eye-blink, a small step size is needed to detect the signal. On the other hand, when you want to measure a task that takes longer time, such as distinguishing between someone walking or running, you'll need a higher step size.

Another criteria on granularity is performance. A low granularity level causes more available data points per time unit, all of this data has to be stored and processed on a computer. In some instances the hardware might not be able to deal with the increased data size and computational complexity.

Third, in case your data contains a lot of random variability (noise), it might be preferable to increase the granularity. Since the aggregation of the data within every time point (i.e. interval) may filter out random variability. Therefore it is useful to increase the step size for summarizing the data and depicting the overall data-trend of interest.

Last of all, the level of granularity depends on the measurement sampling rate, this determines the time intervals at which the measurements are taken. Having a high sampling rate means there are a lot of data points, whilst a low sampling rate provides fewer data points. A granularity level beneath the sampling rate causes problems it is difficult to fill the intermediate time steps, for example, when trying to use a granularity level of an hour on a data set with a daily sampling rate, there is no valid method to determine what each hour contributed to the final measurement.

Question 3

We have identified two tasks we are going to tackle for the crowdsignals data. Here we discuss two other machine learning tasks that could be performed on the crowdsignals dataset, which could be relevant to support a user.

An alternative to the proposed machine learning tasks is clustering, this is an unsupervised learning method that tries to obtain a discrete set of separable groups. This algorithm can be used to distinguish different levels of activity [6], this allows the user to gain statistical information on his measurements during a specific activity, like running [3]. This information could be beneficial as it might give insight into the success of their training routine. So an athlete could use this unsupervised clustering algorithm to find interesting correlations between high performance and other factors such as type of diet.

A second alternative could be reinforcement learning, this paradigm is also unsupervised that tries to find actions that would optimize an agent's cumulative reward [3]. This method could provide a user with advice to improve their physical health. Lets consider the case where the user wants to improve their stamina. The algorithm would advice the user with the optimal action to obtain this goal, this could be to exercise or to rest.

practical part

Question 1

For this exercise we collected data with an Android smartphone using an application called Sensor Record. This application allows for the collection of various types of sensory data. We selected all sensors that overlapped with the sensors used in the crowdsignal dataset, resulting in the following measured features: accelerometer, gyroscope, light and pressure with recording intervals (in ms) of 10, 10, 100 and 100 respectively. Our dataset was collected in a semi-continuous manner, all activities were recorded individually, but they followed each other without notable interval. This was necessary to prevent technical difficulties that arose whilst saving larger amounts of data. The selected activities were: Walking (x2), Cycling, Sitting (x2), OnTable (no interaction to study background noise of sensors) and Yardwork. The sensory data spans a total of 2 hours and 15 minutes.

Table 1 shows the statistics of the numerical attributes. The first observation of interest is the number of missing values for low granularity, the sensor recording interval is too low to allow for such a small granularity. Table 2 provides insight in the label distribution, Sitting is by far the most common label as it occurs in 42.5% of the cases while ScreenTime is the least occuring label.

Table 1: Statistics of processed dataset for numerical attributes (first number listed is for Δ t = 60 s, second value for Δ t = 0.25 s), all values are rounded to the third decimal

Attributes	Missing (%)		Mean		Standard deviation		Minimum		Maximum	
acc_phone_X	0.045	0.916	2.731	0.617	3.980	2.676	-7.575	-8.004	9.309	10.985
acc_phone_Y	0.045	0.916	-2.366	-0.701	5.617	6.631	-11.487	-12.689	7.603	8.964
acc_phone_Z	0.045	0.916	-0.787	4.021	5.328	5.367	-8.996	-10.726	9.886	11.733
gyr_phone_X	0.030	0.910	0.040	0.001	0.590	0.257	-2.288	-2.515	3.106	3.123
gyr_phone_Y	0.030	0.910	0.045	-0.006	0.275	0.243	-0.836	-1.756	1.067	1.673
gyr_phone_Z	0.030	0.910	-0.000	-0.034	0.322	0.213	-1.375	-1.696	1.898	2.383
light_phone_Illuminance	0.022	0.905	9.390	24.054	58.212	199.338	0.000	0.000	664.700	3798.700
press_phone_Millibars	0.022	0.905	1021.848	1021.767	0.232	0.241	1021.332	1021.309	1022.138	1022.185

Table 2: Statistics of processed dataset for categorical attributes (first number listed is for Δ t = 60 s, second value for Δ t = 0.25 s), all values are rounded to the third decimal

Attribute	Value	Percentage of cases (%)	
Label	Cycling	0.134	0.132
Label	OnTable	0.097	0.086
Label	ScreenTime	0.090	0.082
Label	Sitting	0.425	0.418
Label	Walking	0.134	0.134
Label	Yardwork	0.142	0.135

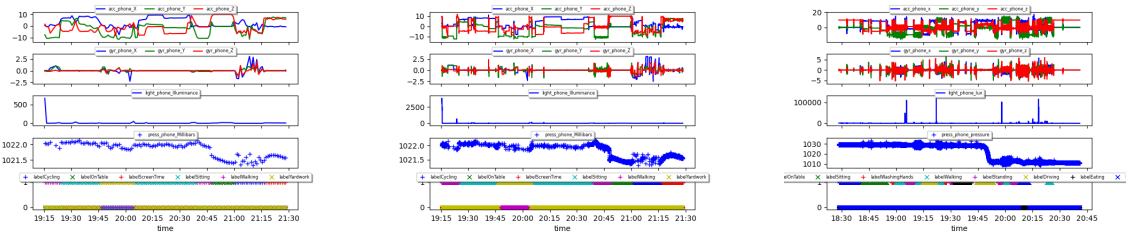
Complementing the table with Fig. 1a and 1b, it appears allows for a more natural inspection of the data. The accelerometer does not show remarkable observations. The gyroscope sensor indicates that the phone was kept in the same orientation for most passive activities, like sitting, but the orientation changed in more physically active activities like Cycling. The light sensor contains very little variation, this is likely to be caused by the fact that the phone was carried around the a pocket most of the time.

A higher granularity requires more data points to be aggregated into a single discrete timestamp, meaning that information is lost. On the contrary, a lower granularity means that fewer data points have to be aggregated and more details of the data are preserved. Fig. **??** and Fig. 1b show recorded values of the different attributes with granularity 60s and 0.25s respectively. The described effect can be observed in these figures as the high granularity data is more smoothed than the low granularity data. On the other hand, the low granularity data contains large variability.

Question 2

The crowdsignal dataset and our dataset have the following activities in common: OnTable, Sitting and Walking. It is difficult to compare our measurement to the crowdsignal dataset as the latter dataset contains more activities that alternate in shorter time periods. This is probably also the cause for the overall increased variation across all attributes of the crowdsignal dataset, Fig. 1c compared to our dataset, Fig. 1b. Interestingly, the scale of the sensor activations for the crowdsignal dataset are different for all attributes but pressure.

As described in the theoretical section, there are various factors which can cause structural deviations in the data. An important step for comparing the two datasets would be to standardize the data, to make sure that the attributes are compared on the same scale.



(a) Δ t=60s (b) Δ t=0.25s (c) Δ t=0.25s

Fig. 1: Descriptive information, sensor values over time

Chapter 3

Question 2

There are generally two approaches for outlier detection: distribution based and distance based

algorithms. We will briefly discuss two situations in which a distance based approach would be preferred over a distribution based approach.

Distribution based methods for outlier detection assume a certain underlying distribution, a normal distribution for example. When the datapoints are outside the boundaries of this distribution, they are considered as outliers. However, when you are working with a population which doesn't follow a certain distributed, there is a change this algorithm will falsely classify a data point as outlier. Hence, in a situation where you're uncertain about the normality of the distribution in your data, an algorithm which looks at features, instead of looking at the distribution, is preferred.

In such case one should use another type of algorithm to detect outliers: the distance based algorithm. This algorithm uses features of the data to identify whether a value is an outlier. More specific, it considers the distance between multiple points in the data set and thereby indicate whether the level of (ab)normality of an datapoint [4]. This approach is preferred in a situation in which the outlier is caused by an entire measurement error, thereby affecting multiple variables within the timepoint. By combining the deviations from other instances for several features, a distance based method might provide a better approach to detect these errors.

Question 4

Since the high computational complexity of local outlier factor (LOF) algorithm, we will discuss some ways to improve the scalability of the approach.

Essentially the LOF algorithm computes the local outlier factor in two steps: first it finds all the values that are in the k-neighbourhood. Then, for all k-values in the neighbourhood the k-distance kdist of a point xi is defined [2]. So instead of taking a global look at points, the LOF approach only takes points into account that surround it. So it computes LOF values for all the datapoints. When you have to detect outliers in a huge dataset, this complexity can result in a lot of computations and therefore long running time. To decrease the running time one could think of priori (to the computation) exclude datapoints that are likely not outliers anyway. This could be done by taking into account the density of the datapoints before the LOF values are computed by the algorithm [5].

Question 3 - compulsory

Imputation is the process of replacing missing values in a dataset. Model-based imputation tries to impute missing values by creating a predictive model. This exercise tries to create a model that is able to predict the missing values for the *heart_rate* attribute. This attribute contains 31.838 data points 24.327 of which are missing values.

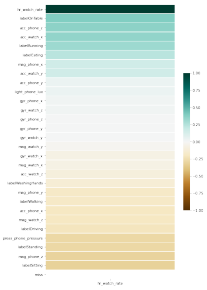
The model of choice is multiple linear regression, this is a simple algorithm that tries to explain the *heart_rate* based on the remaining attributes. Fig. 2a shows the correlation between the attribute of interest, *heart_rate*, against the other attributes. Multiple attributes appear to have some correlation to *heart_rate*, which means that a multiple linear regression model can be used to predict the missing values. The model was trained with all feature, the dataset was reduced to all instances that did not miss a value for *hearth_rate*. Some of the explanatory variables also contained missing values, these have been filled with the median value of the explanatory variable. The model obtained an average R2 score of 0.62 [1] after 10 fold cross validation.

Finally, the missing *heart_rates* are predicted by the model. Fig. 2b shows the distribution of both the imputed and the original (non-imputed) values per label. The figure shows that the model is capable of predicting the *heart_rates* because the distributions of the imputed values are similar to the distribution of the original values.

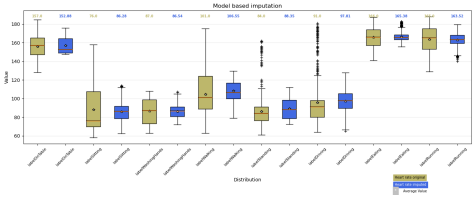
Question 2

To generate Figs. 3.8 and 3.9 we have used the parameter settings described in Sect. 3.5.1. We varied the constant c (smaller and larger values) of the Chauvenet's criterion and study the dependency of the number of detected outliers on c. We then repeated this for the other three methods presented for outlier detection. For this exercise we used the following parameters to see whether the effect of different algorithms with different parameters have an effect on the number of outliers detected in the variable 'acc phone x':

[1] The best possible score is 1



(a) Correlation of *heart_rate*



(b) Distribution of imputed values per label

Fig. 2

- Chauvenet's criterion: $c \in [1, 2, 10]$
- Mixture model: $ncomponents \in [1, 3, 10]$
- Simple distance-based approach with $[dmin, dmax] \in [[0.01, 0.99], [0.1, 0.99], [0.5, 0.99]]$
- Local outlier factors with $factors \in [2, 5, 10]$

As we can see from the figures below (Fig. 4a, 4b, 4c), we can see a small difference in the identified outliers for light_phone_lux, where for *k=1* and *k=2* the outliers shown are the same, while in *k=10* there are few elements which are not considered outliers. This shows that an higher value of k equals to a more lax identification of outliers, which implies that the higher the k value, the fewer, but more certain, are the outliers found.

Conversely, for acc_phone_x (Fig. 5a, 5b, 5c) there seems to be no identified outlier, for either low nor high values of k.

Using the mixture model with different parameters, as we can see from the figures, the results for both the light_phone_lux (Fig. 6a, 6b, 6c) and acc_phone_x (Fig. 7a, 7b, 7c) do not differ while varying the mixture model parameter, resulting in a somehow consistent behaviour. Even so, higher values of k result in higher probabilities of observing a datapoint (as expected, since this allows for a better fit to the dataset). The resulting dataset however hardly differs. Choosing low value is in this case fine, as this saves some computation.

Analyzing the euclidean distance, we can see that, similarly to what happened for the Chauvenet's comparison, for both light_phone_lux (Fig. 8a, 8b, 8c) and acc_phone_x (Fig. 9a, 9b, 9c) there is a clear increase of the number of identified outliers decreasing the value of dmin, while, increasing dmin, the confidence of the fewer outliers detected is higher.

Finally, we can consider the difference that varying the number of factors of the local outlier factor can lead to. Actually, analyzing the figures for light_phone_lux (Fig. 10a, 10b, 10c) and acc_phone_x (Fig. 10a, 10b, 10c), there is no apparent difference while changing the number of factors considered.

Chapter 4

Theoretical part

Question 1

There are different functions that summarize numerical values within the time domain to a single number and create features [6]. Here we will discuss four of them and provide examples in which situations they can be useful.

Mean The mean is used as a measure of central tendency when you want to include all the values in the data set and any change in any of the scores will have an affect on the value of the mean. So for example in case you want to make predictions about someones progress in losing weight you can use the mean of physical activity on a day and include calorie intake, to predict how many weight someone will lose in a couple of weeks.

Standard deviation The standard deviation is used as a measure of the amount of variation or dispersion of a set of values. Data from an accelerometer with contain both positive and negative values because movements are measured back and for over the same axis. Since you don't want to approximate to zero (in case of using the mean), one could better use the standard deviation over multiple time points to summarize the amount of activity during these time points.

Minimum A dietitian is helping a client who is diagnosed with eating disorder Anorexia. So in case the dietitian want to make sure the calories taken are above a certain threshold, an indicator of whether that person succeeded might be helpful. Taking the minimum calories during a certain time interval would be a good indicator whether the client succeeded in gaining weight or not.

Maximum An athletic with heart problems might want to make sure his heart rate is not too high, to prevent himself for serious health problems. Taking the maximum heart rate during a certain time period could be a good indicator whether he is not taking risks for his health.

Question 6
Imagine that we want to learn a model that predicts someones mood based on the amount of social activity. We here define three dedicated features that can be useful in this context based on measurements we can potentially collect from the mobile phone.

Time First we could use the mean of time a day a person is active on social applications like: whatsapp, instagram, snapchat and facebook in a certain time period. With 'active' we mean these applications are 'open'. It it important to mention that we only use the screen time of applications which contain social activity, so we wouldn't include spotify for example.

Amount Since whats-app is one of the most commonly used communication applications, we could use the amount of send whats-app messages on a day to make predictions about someones mood.

Location Since social activity can also be activity beside our mobile phone, so real life social activity, one could think of using GPS location trackers to someones location. We could use this location to see whether that person is at a location which is beforehand classified as 'social', such as: bar, clubs, terrace and sport clubs.

Question 7
One reason to perform stemming on the words is to make sure all conjugates of verbs or plural forms of nouns are considered as the same word. Here we explain one advantage and one disadvantage of using stemming,

Stemming refers to the process of identifying the stem of each word to reduce words to their stem and map all different variations of to a single term [2]. So this means that words with a slightly different form but the same meaning are mapped to one single word. The advantage of this transformation is that words which are rarely used are easily recognized by the system. This will therefore increase the predictive power in such a way that the relationship between words are faster recognized than when those words a considered as separate words.

A disadvantage of stemming is that it might be the case that important information is lost. When for example the small difference between words have great impact on the meaning. The additional information will lost when the word is transformed to its basic form.

practical part

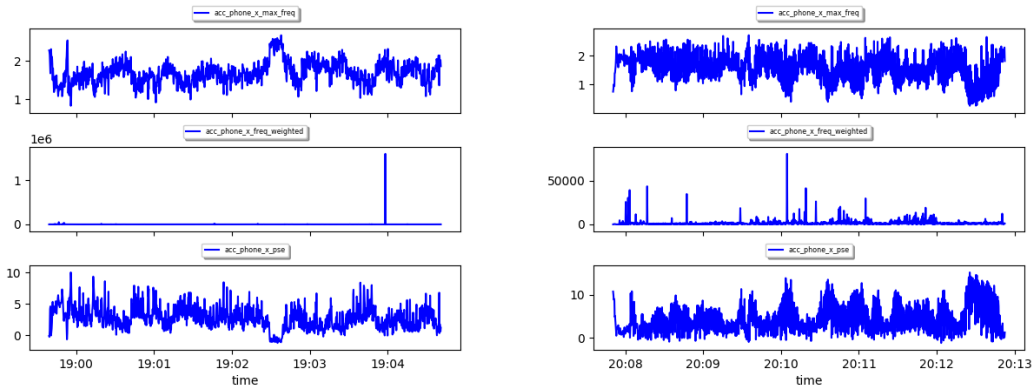
Question 1 In order to better identify possible interesting patterns, let's consider the attribute acc_phone_x, as it is more consistent, for the labels of "walking" and "running", since sitting vs walking or ontable vs walking should not show any interesting pattern.

In order to better compare "walking" and "running", the number of instances of "walking" is 4694 and the number of instances of "running" is 1208, we consider only the first 1208 values for the label "walking".

For both activities, the metrics of the frequency-domain considered are the following: maximum frequency, weighted frequency, and power spectral density estimation (PSE). All three of these metrics are found by "windowing" the signal, meaning that the signal was divided in "slices" by a sliding time window, so these metrics can be computed by passing the signal through a discrete Fourier transform.

In the images below (Fig. 3a), we can see a consistent behaviour with the walking activity, where the weighted frequency is zero for almost the whole activity, and both the maximum frequency and the pse show the consistent periodical behaviour of the walking activity, for every time window throughout the whole activity, with exception for the time going between 19:02:30 and 19:03:00, where there is a "strange" activity, that may represent an outlier, perhaps caused by an error of the signal acquisition from the phone.

Considering the "running" activity (Fig. 3b), we can see that both the maximum frequencies and the PSEs are more dense, which cna imply a much higher number of changing frequencies and their overall power, which can be traced back to the more "chaotic" activity. This can also bring forth the possibility of more outliers, and for this we can analyze the weighted frequencies graph, in which few instants appear to have weighted frequencies exceedingly too high to be normal.



(a) Maximum frequency, average frequency and power spectrum estimate of the walking activity considering the parameter acc_phone_x

(b) Maximum frequency, average frequency and power spectrum estimate of the running activity considering the parameter acc_phone_x

Question 2 To better study the given dataset, we can also consider few more metrics both in the time-domain and in the frequency-domain: Mean, Standard Deviation, Skewness, Kurtosis, and maximum estimate of the power spectrum density. For studying the effects of this new metrics, we consider the attribute acc_phone_x pf the crowdsignals dataset.

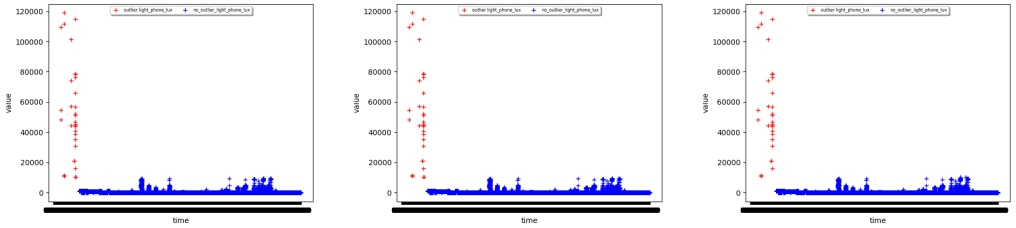
As shown in Fig. 12, we can see clear correlations between the mean and standard deviation values of the signal and the different labels, which suggests that these metrics may contain possibly interesting information that can be added to the dataset. Moreover, the skewness metric, which represents the magnitude and direction of the distribution's deviation from the normal distribution, seems to be - perhaps weakly - related to the different labels through time, which may also indicate possible information that can be added to the dataset.

However, kurtosis, which is another technique that, similarly to the skewness, tries to explain the shape of the distribution with regards to the normal distribution, does not seem to be able to deliver interesting and/or useful information to be furtherly studied.

Finally, the maximum estimate of the power spectrum density does not show any particularly interesting features or patterns from a first look, so we may either discard this metric, or look more closely for any correlation or information that may be useful.

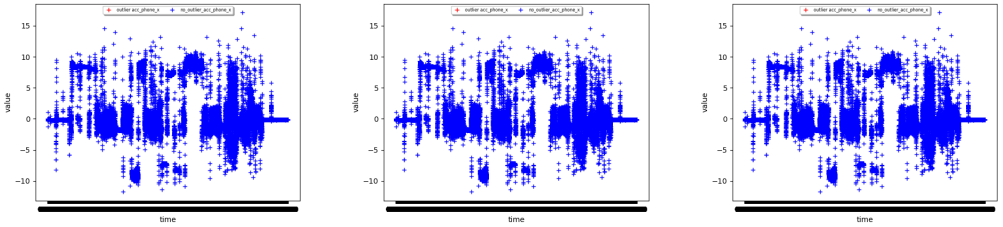
References

1. C. Dobbins and R. Rawassizadeh, "Clustering of Physical Activities for Quantified Self and mHealth Applications," 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, 2015, pp. 1423-1428, doi: 10.1109/CIT/IUCC/DASC/PICOM.2015.213.
2. M.Hoogendorn and B.Frunk. Machine learning for the quantified it self (2007) Amsterdam, p:66-83.
3. Lloyd, S., Mohseni, M. & Rebentrost P. (2013). Quantum algorithms for supervised and unsupervised machine learning.
4. M. M. Breunig, H. Kriegel, R. T. Ng, and J. Sander. (2012). LOF: Identifying Density-Based Local Outliers. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data (SIGMOD '00). ACM, New York, NY, USA, 93-104.
5. V. Hautamaki, I. Karkkainen and P. Franti, "Outlier detection using k-nearest neighbour graph," Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004., 2004, pp. 430-433 Vol.3, doi: 10.1109/ICPR.2004.1334558.
6. Gravetter FJ, Wallnau LB. Statistics for the behavioral sciences. 5 th ed. Belmont: Wadsworth - Thomson Learning; 2000.



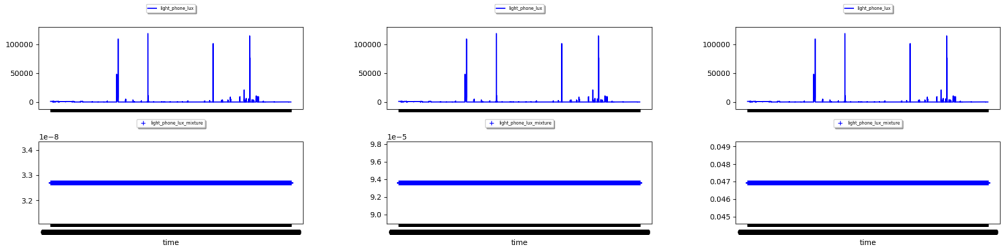
(a) Chauvenet's criterion of "light_phone_lux" with *k=1* (b) Chauvenet's criterion of "light_phone_lux" with *k=2* (c) Chauvenet's criterion of "light_phone_lux" with *k=10*

Fig. 4: Chauvenet's criterion with varying k values for the attribute "light_phone_lux"



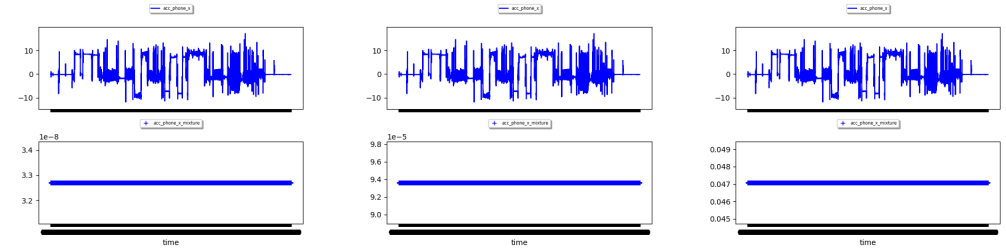
(a) Chauvenet's criterion of "acc_phone_x" with *k=1* (b) Chauvenet's criterion of "acc_phone_x" with *k=2* (c) Chauvenet's criterion of "acc_phone_x" with *k=10*

Fig. 5: Chauvenet's criterion with varying k values for the attribute "acc_phone_x"



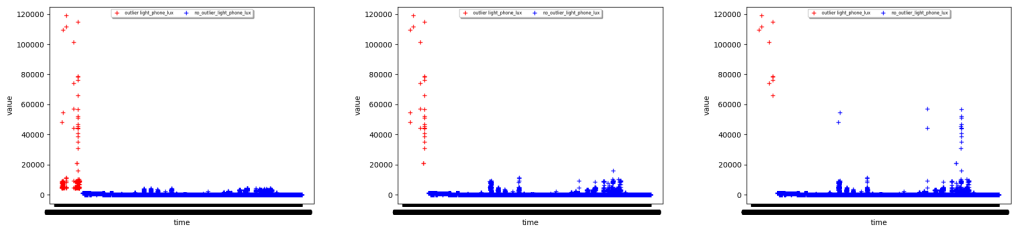
(a) Mixture models of "light_phone_lux" with *k=1* (b) Mixture models of "light_phone_lux" with *k=3* (c) Mixture models of "light_phone_lux" with *k=10*

Fig. 6: Mixture models with varying k values for the attribute "light_phone_lux"



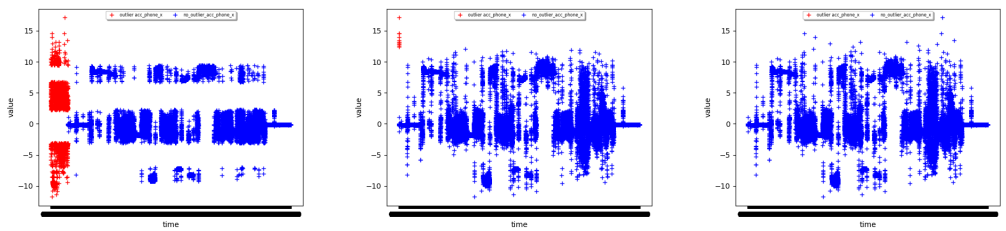
(a) Mixture models of "acc_phone_x" with *k=1* (b) Mixture models of "acc_phone_x" with *k=3* (c) Mixture models of "acc_phone_x" with *k=10*

Fig. 7: Mixture models with varying k values for the attribute "acc_phone_x"



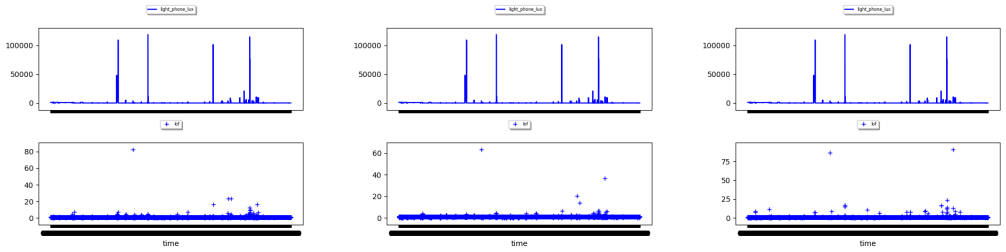
(a) Euclidean Distance of "light_phone_lux" with *dmin=0.01* and *dmax=0.99* (b) Euclidean Distance of "light_phone_lux" with *dmin=0.1* and *dmax=0.99* (c) Euclidean Distance of "light_phone_lux" with *dmin=0.5* and *dmax=0.99*

Fig. 8: Euclidean Distance with varying dmin and dmax values for the attribute "light_phone_lux"



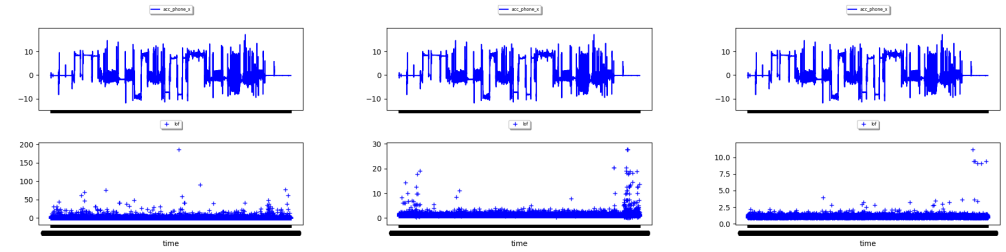
(a) Euclidean Distance of "acc_phone_x" with *dmin=0.001* and *dmax=0.99* (b) Euclidean Distance of "acc_phone_x" with *dmin=0.001* and *dmax=0.99* (c) Euclidean Distance of "acc_phone_x" with *dmin=0.001* and *dmax=0.99*

Fig. 9: Euclidean distance with varying dmin and dmax values for the attribute "acc_phone_x"



(a) Local Outlier Factor of "light_phone_lux" with *factors=2* (b) Local Outlier Factor of "light_phone_lux" with *factors=5* (c) Local Outlier Factor of "light_phone_lux" with *factors=10*

Fig. 10: Local Outlier Factor with varying the number of factors for the attribute "light_phone_lux"



(a) Local Outlier Factor of "acc_phone_x" with *factors=2* (b) Local Outlier Factor of "acc_phone_x" with *factors=5* (c) Local Outlier Factor of "acc_phone_x" with *factors=10*

Fig. 11: Local Outlier Factor with varying the number of factors for the attribute "acc_phone_x"

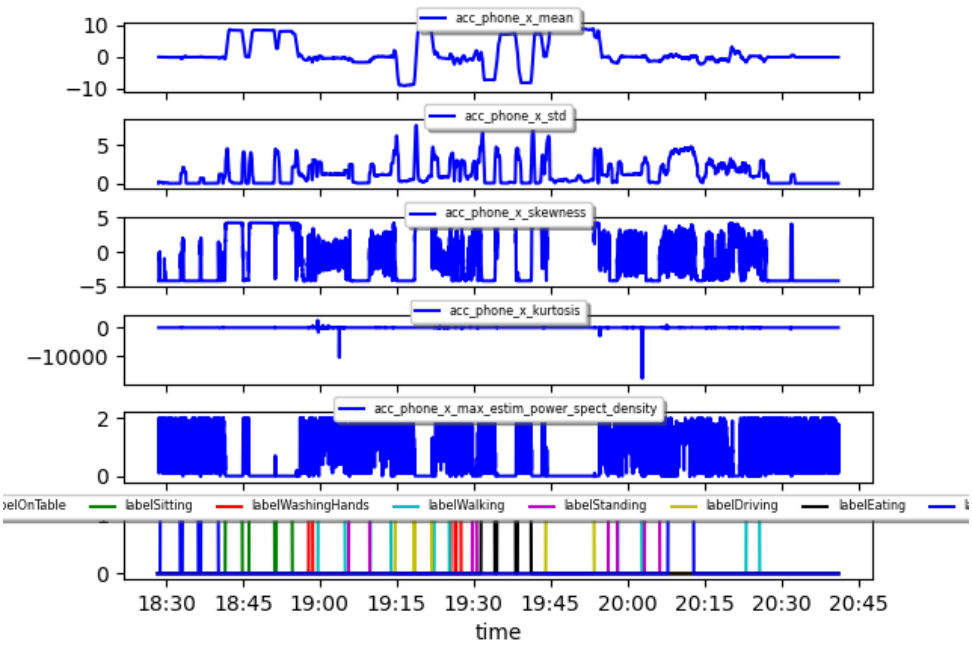


Fig. 12: Mean, Standard Deviation, Skewness, Kurtosis, and maximum estimate of the power spectrum density of the attribute "acc_phone_x" considering every label