ML4QS Assignment 1

Haohui Zhang[1 [2722930]], Yilin Li[1 [2737659]], and Yiming Xu[1 [2696855]]

Vrije Universiteit, Amsterdam, Netherlands

1 Theoretical Part

1.1 Chapter 2

Q1. 1) People could live in different places and engage in various activities, which leads to different accelerations, different magnetometers etcetera.
2) Some individuals may have abnormal or faulty sensors which leads to different outliers for different individuals. 3) Different individuals may use different devices to collect data, some may use a smartphone, and some may use a smartwatch or smart bracelet. Smartphones make significantly more noise than smartwatches when collecting heart rate data.
Q2. 1) Task: Different machine learning tasks may lead to requiring different granularity. Focusing on supervised learning, we have two tasks separately classification task and regression task. A classification task which aims to predict a label, apparently does not need a very small granularity compared with the regression task, which may need to predict sensory values.
2) Noise level: When the dataset is really noisy, aggregating the data using a high granularity can remove some of the noise and reduce the noise level. But if a small granularity is applied, the aggregated data still contains a lot of noise. In the scenarios of different missing values, if the dataset has a lot of missing values, selecting a high granularity will be a good choice.
3) Available memory and Cost of storage: We also need to decide on the granularity based on the available memory and cost of storage. Normally, the smaller the granularity, the more instances contained by the dataset which means the more information. However, more instances also lead to more cost. When the available memory or cost is limited, we had better use the higher granularity.
4) The shape of the dataset: This criteria is also an important factors. If the dataset is too small, a high granularity cannot be applied. This is because the high granularity will cut the dimension of the dataset, and using a small dataset to train a machine learning model will normally lead to overfitting [1].
Q3. 1) Unsupervised learning task: clustering. Clustering is the task of grouping a set of objects so that objects in the same group are more similar to each other than objects in other groups[2]. Tackling for the crowdsignals dataset, we can identify a task to cluster the locations often visited by Bruce to find out what impact the different locations have on his mood. For Arnold, we can cluster his physical state to see in what situation he may have a very good state, and in what situation he will have a bad state.
2) Reinforcement learning task: reinforcement learning techniques different from

other machine learning method, mainly aims to find the optimal policy for different states and to final achieve the ultimate goal. For Arnold, his ultimate goal is to finish the line during the Hawaiian IRONMAN championship. So in this task, the reward may be the improvement of Arnold's running performance, and what we need to learn is to give appropriate daily advice based on his state. This advice, which can also be named the optimal policy, will maximize the total rewards of Arnold. The reinforcement learning task can directly support the user in achieving his ultimate goal.

1.2 Chapter 3

Q2. In some cases, it is not known whether a particular attribute follows a normal distribution, and as a result, some normal data points will be mistakenly detected as outliers. Therefore, we need to perform extensive testing to find a mathematical distribution that fits the attribute. Because distribution-based algorithms mainly target at single attribute, when the dataset has very large dimensions, using the distribution-based algorithms is not practical. Compared to distribution-based algorithms, distance-based algorithms are nonparametric. Nonparametric methods generally make fewer assumptions about the data and thus can be applicable to more scenarios [3]. Besides, distance-based algorithms are also more scalable for large datasets. So when the dataset is large or high-dimensional, distance-based algorithm will definitely be a better choice.
Q4. The local outlier factor (LOF) and the related local outlier detection algorithms are nearest neighbor based algorithms. The computational complexity of this type of algorithms is $O(N^2)$. Specifically, LOF uses Euclidian distance and kNN to estimate local density [4]. Since the LOF method traverses the surrounding K points, it needs to find the LOF value for all data points, so it will generate a huge amount of computation. Another reason is that LOF requires the entire dataset and distance values to be stored in computer memory. Because the algorithm has the problem of high time complexity, it is not suitable for large data sets and high-dimensional data sets. We can consider reducing the K value or loading the algorithm running on the GPU to accelerate the computational time. Local Correlation Integral (LOCI) can be used to improve the scalability and efficiency of detecting anomalies, however, it requires more execution time because the radius expansion needed to be repeated. Another fatal flaw of the LOF algorithm is that it needs to be recomputed from scratch if any modification occurs in the dataset. To improve this, the ILOF algorithm applies a landmark window to update and calculate the LOF score when a new data point arrives, and this algorithm has a computational complex of $O(NlogN)$ [4].

1.3 Chapter 4

Q1. 1) Mean: Similar to the example given in the book, assume the dataset collected from the smartwatches, which record the heart rate per second. We can use the mean value to summarize the history of this numerical attribute.

2) Standard deviation: As for the acceleration attribute, only the standard deviation value is a good choice to summarize the history. Acceleration is a attribute having positive and negative values on both sides of 0. The mean value of this attribute can be 0, the max and min values cannot reflect the variation of acceleration. Therefore, only the standard deviation can provide a suitable value to reflect the change in acceleration over each period of time.
3) Minimum: The minimum value of the temperature attribute have an impact on exercise performance or heart rate. It can be used to find the relationship between temperature and physical status. It can also be used to summarize the attribute in the regression task to predict heart rate.
4) Maximum: The maximum value of the speed or gyroscope attribute is an important feature to determine the movement status (walking, running, etc.), which can also be used to summarize the attribute in the classification task.
Q6. In order to determine the mood of the user, we can collect the time or duration of screen activity, the duration of usage of social apps and location of the user. The longer time you speed on the phone, the higher likelihood of negative emotions is. The duration of usage of social apps can reflect how long the user spent on the social activity, the more time are used, the more depression may be felt. If the user located in the park or in the shopping mall, he or she is probably happy. In contrary, if the user in the school or in the working place, he or she has a higher likelihood of negative mood.
Q7. Pros: Stemming shortens the vocabulary space, reduces the size of index significantly, and the speed of stemming process is fast.
Cons: Stemming cannot relate words that have different forms based on grammatical structure. For example, better should be resolved to good and worse should be resolved to bad. However, stemming fails. Also with stemming, there is a lot of ambiguity.

2 Practical Part

2.1 C2Q1: Creating Our Dataset and Plot Figures in Various Δt

In this section, we intended to generate our own dataset through utilizing Phyphox[5] on smartphone to monitor and collect subject's behaviors. In detail, various sensory data are collected by this application, our experiment focused on analyzing the following six characteristics: `(L)accelerometer, Gyroscope, Location, Magnetometer and behavior labels` with different levels of granularity (i.e., summarize data periodically $\Delta t = 0.25s/60s$). Details are shown on Fig.1 and 2. We illustrated the fluctuation of the different features over time. The value of each time point (i.e., results of measurement on the Y-axis) in the figure represents the average value of all monitored points within the time window λ. Specifically, the top panel of Fig.1 displays the value of `acceleration` on X, Y and Z-axis. Besides, the only two different sub-figures are: the sub-figure describes `Location` contains data in three different axis of measurement, that are `Laitiude, Longtitude and Velocity`; When the smartphone was in the "On Table" state, the subjects did not perform specific behaviors. Hence, most of the

sensory data had a value of zero, but the `Magnetometer` value was perturbed due to the interference of the ambient magnetic field.

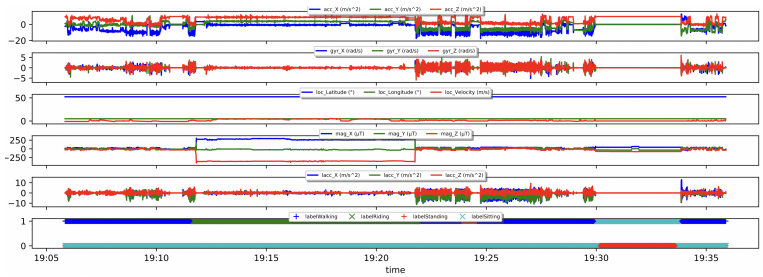


Fig. 1. Plotting features with $\Delta t = 0.25s$

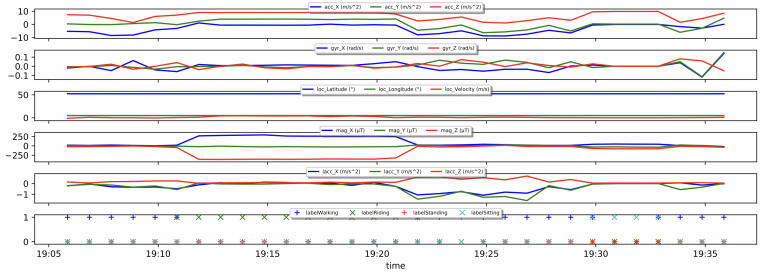


Fig. 2. Plotting features with $\Delta t = 60s$

Table 1. An overview of differences between owned dataset and crowdsignals

Crowdsignals Data Set $\Delta t = 0.25s$	Mean in λ	Stanard Deviation	Minimum	Maximum
acc_phone_x (m/s^2)	1.1	4.7	-11.8	17.1
gyr_phone_x $(^\circ/s)$	0.0	0.6	-4.0	5.7
Owned Data Set $\Delta t = 0.25s$	Mean in λ	Stanard Deviation	Minimum	Maximum
acc_phone_x (m/s^2)	2.7	6.9	-7.6	9.3
gyr_phone_x (rad/s)	-0.009	0.617	-5.2	4.4
Owned Data Set $\Delta t = 60s$	Mean in λ	Stanard Deviation	Minimum	Maximum
acc_phone_x (m/s^2)	2.4	6.4	-7.2	7.9
gyr_phone_x (rad/s)	-0.004	0.046	-4.0	0.14

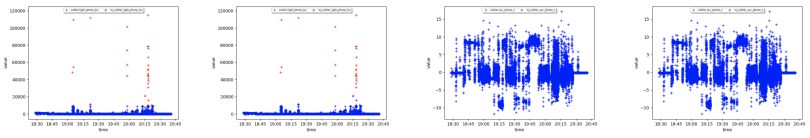
The differences of plot with various Δt we observed are: 1)Different Δt lead to different values for the same calculated metric). As known from the Table.1, we observe that the extreme values of metrics such as standard deviation and mean are higher than the result of $\Delta t = 60s$ when $\Delta t = 0.25s$ is chosen in analyze. 2) Although the result of $\Delta t = 0.25s$ shows the specific behavioral patterns of the subjects, the narrow size of rolling window, which leads to processed data is susceptible to noise. The result of $\Delta t = 60s$ circumvents the effect of noise as

much as possible. However, the oversized rolling window value causes difficulties in reflecting the characteristics of different behaviors. In total, as the level of granularity increases, the influence of outliers will decreases, but the specificity between various behaviors will be harder to recognize.

2.2 C2Q2: Comparison between Owned Dataset and Crowdsignals

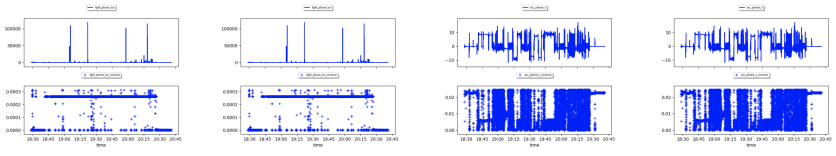
Considering that data from different platforms use different measurement metrics (e.g., The Crowdsignals selects ($°/s$) as unit of `Magnetometer` while (rad/s) is imported by The Owned Dataset), the different data from multiple platforms cannot be used directly for EDA and subsequent procedures. The best solutions to achieve comparison are: 1) Normalization of data using relevant algorithms. This could be done by pandas or sklearn; 2) Generating relevant features from existed features in different metrics. For example, the result of standard deviation on various dataset can be used in the following steps, since its value is not affected by different measurement methods (i.e., The two panels at the bottom of Table.1 shows new features are independent of the measurement units) **Differences:** In addition to using different measurement units for some of the data, Crowdsignals continuously collected approximately 2 hours of behavioral data while we recorded only half an hour of behavioral activities of the subjects. The former contains data collected by optical sensors (i.e., `Light_phone`) while ours count changes in `Laccelerometer`. The Crowdsignals dataset contains a richer set of behaviors while our data only monitors basic daily behaviors.

2.3 C3Q2: Test of the Chauvent's Criterion



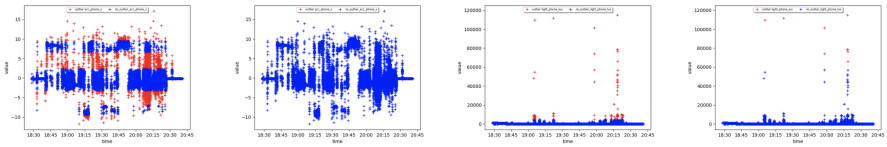
(a) light_phone_lux c=2 (b) light_phone_lux c=10 (c) acc_phone_x c=2 (d) acc_phone_x c=10

Fig. 3. Chauvenet, as C increases, the outliers for light_phone_lux decrease, but the accuracy of the detected outliers improves. On acc_phone_x, however, we do not observe obvious outlier changes as C increases.



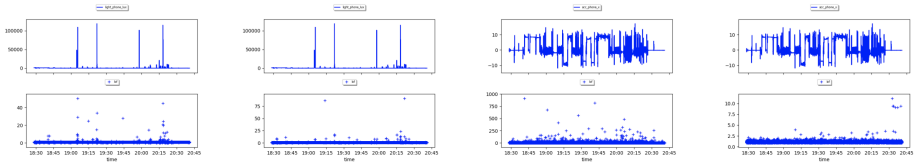
(a) c = 2 (b) c = 10 (c) c = 2 (d) c = 10

Fig. 4. Mixture, with the parameter increases, we do not observe a significant difference in the outlier of light_phone_lux and acc_phone_x.



(a) dmin=0.01, fmin=0.99 (b) dmin=0.5, fmin=0.99 (c) dmin=0.01, fmin=0.99 (d) dmin=0.5, fmin=0.99

Fig. 5. Distance, as dmin increases, the number of detected outliers for acc_phone_x and light_phone_lux decreases, but the detected outliers are more accurate.



(a) K = 1 (b) K = 10 (c) K = 1 (d) K = 10

Fig. 6. LOF, We did not observe any effect of changing the value of K.

For Chauvenet's criterion, we set the value $c \in [2, 5, 10]$, and we set $C \in [1, 5,$ in Mixture models. In Simple distance-based approach, we set the $[dmin, dmax] \in [[0.01, 0.99], [0.1, 0.99], [0.5, 0.99]]$ and we set $k \in [1, 3, 10]$ for Local outlier factor.

2.4 C3Q3: Imputation of Heart Rate

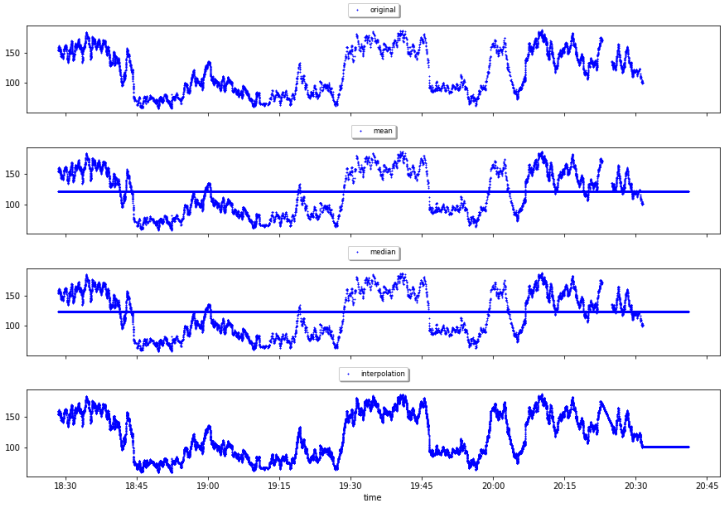


Fig. 7. We supplemented the missing value with median, mean, and interpolation, resulting in the following heart rate.

2.5 C4Q1: Explore the Frequency Domain

Power spectral entropy (PSE)[6] demonstrates whether a given behavior can be represented by one or several discrete frequencies. Multiple behaviors show their pattern based on the value of PSE over time. Since various behaviors of subjects lead to differences in the pattern of PSE, we can incorporate it into the set of features. Fig.8 summarizes the patterns if we consider individual frequencies. Besides, we did actually observe the consistent amplitudes of certain frequencies during the same activities. Compared to other metrics, the maximum frequency of acceleration on the X-axis stays with high and consistent amplitudes for a long time if we ignored the effect of noise. However, because of the complexity of the subject's behavior and the occurrence of noise during the test, a specific amplitude cannot be associated with a behavior of the subject.

What we should also note is that different activities differ for the amplitudes. In detail, Fig.8 shows the difference of amplitudes on various activities, which illustrates the effect of different activities on the PSE of the same feature (i.e., `Accelerometer`). As shown in the Figure, the amplitude of the subjects was significantly higher when they were running than when they were walking and sitting still. Also, the PSEs of different features under the same behavior (i.e., Running) have some of the same characteristics. `Acceleration, gyroscope and magnetometer` measurements had similar amplitudes when the subjects were running. This leads to the preliminary conclusion that, in general, the amplitude of the PSE increases to a certain extent with the increase of behavior intensity. However, considering the complexity of the experimental environment and the generation of noises such as the one shown in the third panel of Fig.8, we still need to investigate further.

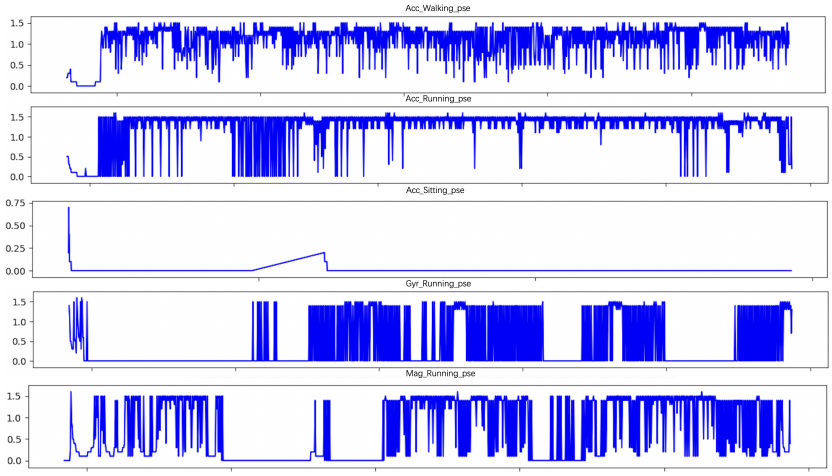


Fig. 8. Differences of patterns based on analyzing various behaviors and metrics

2.6 C4Q2: Implementation of Additional Metrics

With the aim of generating additional metrics in the time domain, we have already realized the implementation of *std, sun, variance and median* as metrics. We calculated them by designing functions in TemporalAbstraction.py and summarized the results in Fig.9. We observed that those additional metrics hardly illustrate patterns of behavior, except for *sum*. It shows the sum of the acceleration of the X-axis over time within the rolling window. When $ws = 1200$, the curve of *sum* can reflect the subject's behavior to some extent (i.e., the higher the behavior intensity, the larger the value of *sum*). We believe that choosing a smaller ws can achieve more significant results. In stead of designing for time domain, *mean and variance* are also implemented as metrics for frequency domain. However, consider both generated metrics cannot represent the fluctuation of data and fit in any pattern, we concluded they are not useful in this section.

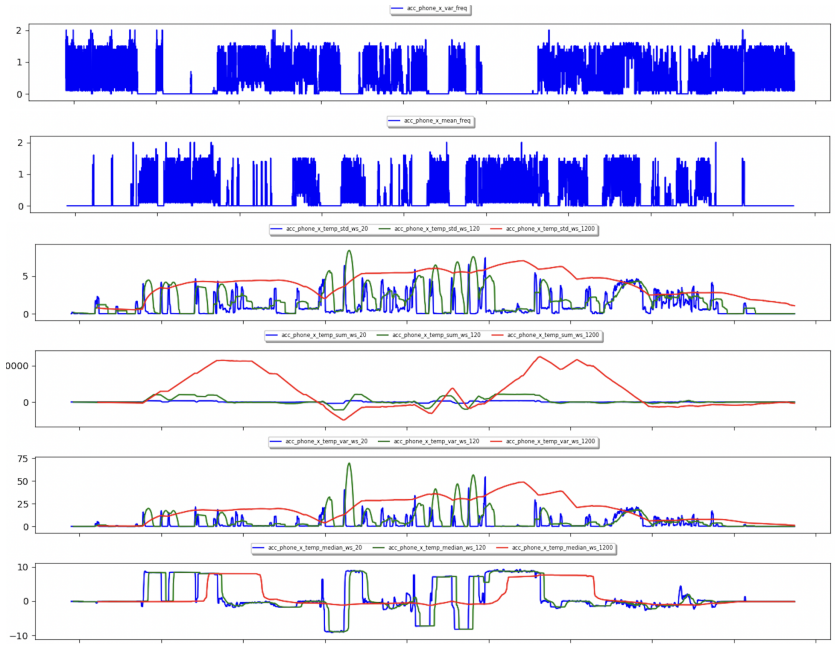


Fig. 9. Frequency Domain: Implementing the variance and mean of frequency; Time Domain: Implementing the std, sum, variance and median of acc_phone_x.

References

1. Rafael Alencar, Dealing with very small datasets, Kaggle, https://www.kaggle.com/code/rafjaa/dealing-with-very-small-datasets/notebook. Last accessed 12 June 2022.
2. Cluster analysis, https://en.wikipedia.org/wiki/Cluster_analysis#Applications. Last accessed 12 June 2022.
3. Daniel Chepenko, A Density-based algorithm for outlier detection, Towards Data Science, https://towardsdatascience.com/density-based-algorithm-for-outlier-detection-8f278d2f7983#: :text=Distance%2Dbased%20outlier%20detection%20method,reasonable%20neighbourhood%20of%20the%20object. Last accessed 12 June 2022.
4. Alghushairy O., Alsini R., Soule T., Ma X.: A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Stream. In: Big Data and Cognitive Computing 2021, 5(1):1. https://doi.org/10.3390/bdcc5010001
5. Staacks, S., Hütz, S., Heinke, H., Stampfer, C. (2018). Advanced tools for smartphone-based experiments: phyphox. Physics education, 53(4), 045009.
6. Hoogendoorn, M., Funk, B. (2018). Machine learning for the quantified self. On the art of learning from sensory data.