Assignment 1 - Group 32*

Bas van der Bijl (sbl206), Max de Bruijne (mbe345), and Tom de Valk (tvk550)

Vrije Universiteit Amsterdam, 1081 HV Amsterdam, The Netherlands

1 Theoretical Exercises

Exercise 2.1 Three causes for substantial differences between sensory values obtained from multiple users:

1. The use of the sensory devices may differ substantially. Some users may wear their sensory device around their arm or carry it in their pocket or bag, leading to divergent measurements.
2. The location of measurement likely affects the measurements of sensory devices as travel distances and amounts of social interaction can differ per location, for example.
3. The different users of the sensory devices likely have different lifestyles with varying activities, leading to different measurements.

Exercise 2.2 Four criteria for choosing the granularity of the measurements for a dataset, as discussed in Machine Learning for the Quantified Self [4]:

1. Available memory and cost of storage
2. Available computational resources for machine learning tasks
3. The task for which the measurements are used
4. The level of sensory noise

Exercise 2.3 Next to the classification and regression tasks mentioned in Machine Learning for the Quantified Self [4], two alternative machine learning tasks that could be performed on the *crowdsignals* dataset are identified:

1. Time series can potentially be used to predict serious health conditions such as future epileptic seizures or heart attacks.
2. Clustering can be used to group users with similar activities. These groups can then be used for the recommendation of activities or training programs.

Exercise 3.2 One of the advantages of distance-based outlier detection compared to distribution-based outlier detection is that the distance-based method does not assume a distribution of the data. Secondly, distance-based outlier detection methods often use a finer granularity of analysis compared to distribution based methods [1]. Therefore, it is possible to distinguish noise and true outliers, since noise has a lower k-nearest neighbor distance compared to true outliers.

* Part of the 2021 Machine Learning for the Quantified Self course at VU Amsterdam, The Netherlands

Exercise 3.4 The Local Outlier Factor (LOF) algorithm identifies instances with a lower local density compared to its neighbours, therefore focusing on the local density. The algorithm is complex and can be split up into the following four steps:

1. The largest distance among the distances of the k closest points of a point x_i is defined as $k_{dist}(x_i)$. The set of points (neighbours) that is within k_{dist} is defined as,

$$k_{dist\_nh}(x_i) = \{x | x \in \{x_1, ..., x_{i-1}, x_{i+1}, ..., x_N\} \wedge d(x, x_i) \leq k_{dist}(x_i)\} \quad (1)$$

2. The reachability distance, which is the real distance when x_i is not within the k nearest points of x, is defined as,

$$k_{reach\_dist}(x_i, x) = \max\{k_{dist}(x), d(x, x_i)\} \quad (2)$$

   where $d(x, x_i)$ is the real distance between the points x and x_i.
3. The local reachability density around a point x_i is then defined as,

$$k_{lrd}(x_i) = \frac{|k_{dist\_nh}(x_i)|}{\sum_{x \in k_{dist\_nh}(x_i)} k_{reach\_dist}(x, x_i)} \quad (3)$$

4. Overall, to determine how big of an outlier a point is compared to its neighbors the Local Outlier Factor is calculated using the following formula,

$$k_{lof}(x_i) = \frac{\sum_{x \in k_{dist\_nh}(x_i)} \frac{k_{lrd}(x)}{k_{lrd}(x_i)}}{|k_{dist\_nh}(x_i)|} \quad (4)$$

A large value for the local reachability density (k_{lrd}) implies the point exists very close to its neighbors. Therefore, a low local reachability density score implies a high LOF. The LOF algorithm mainly has two disadvantages as it is not accurate enough to detect outliers in large datasets and the performance of the algorithm depends on the scale of the local neighborhood (k) [3]. In order to improve the scalability of the LOF algorithm a Graphics Processing Units (GPU) implementation of the k-nearest neighbors can be implemented [2]. Alshawabkeh et al. found a CUDA-based GPU implementation of the k-nearest neighbor algorithm to accelerate LOF classification and therefore making it more appropriate for large datasets. Another way to improve the scalability of the LOF algorithm is introduced by Yan et al., using the TOLF method. Instead of computing the LOF score for all data points, which is costly when working with large datasets, the dataset is partitioned into regular shaped cells with a carefully designed size, a cell that contains more than k points can be pruned immediately without other computations [7].

Exercise 4.1 Mean, standard deviation, minimum and maximum are functions that are used to summarize numerical values within the time domain. Examples related to a user's heart rate are given for each of these functions:

- **Mean**: Can be used to define an expected value of the heart rate at any given time.
- **Standard deviation**: Can be used to describe the average variations in the heart rate.
- **Minimum**: Can be used to define resting heart rate
- **Maximum**: Can be used to show the heart rate at peak performance

Exercise 4.6 In the prediction of a person's mood based on the amount of social activity, the following three features may be useful. These features can potentially be extracted from the person's mobile phone.

1. Amount of time spent using different social media applications
2. Calling activity
3. Amount of words spoken to others

Exercise 4.7 One of the **advantages** of stemming is the shortening of the vocabulary space, and therewith reducing the size of the feature space. The related words can be identified quickly and there is no need for dictionary data for every language.

Two **disadvantages** of stemming, as discussed by Jivani et al. [5], are over-stemming and under-stemming. Over stemming occurs when two words, with different stems, are stemmed to the same root. Under stemming occurs when two words that should be stemmed to the same root, are not.

2 Coding Exercises

Exercise 2.1a The dataset that was created contains information originating from a Fitbit Charge 3 and the Sensor Recorder iPhone app [6]. The Fibit provides data about heart rate, burned calories, steps, exercise labels, weight, BMI and device temperature. The Sensor Recorder iPhone app provides gyroscope and accelerometer data.

Figures 1 and 3 show a plot and boxplot of the gathered gyroscope, accelerometer, heart rate (in beats per minute (BPM)) and activity. The remaining data that was gathered had a too low frequency to be of any value. Therefore, the features steps, burned calories, weight and BMI have been disregarded. In Figure 1, it can be seen that there is a period of missing values between 9:00 and 11:00. Moreover, it can be seen that the heart rate yields higher values when the activity label shows "Sport". In Figure 3, it can be seen that the spread of the gyroscope data is larger than that of the accelerometer. Considering that the range of the gyroscope and accelerometer data was much lower than that of hear rate, the heart rate was not plotted in the same boxplot. The spread of the heart rate is shown in Figure 5. It can be seen that the median recorded heart rate is approximately 80 and that the spread of the heart rate is approximately between 60 and 130 beats per minute.

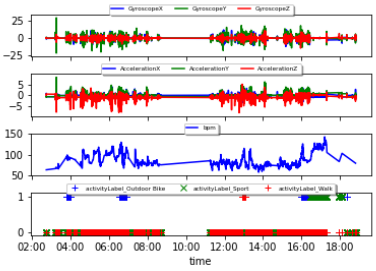

Fig. 1. Plots of gyroscope, accelerometer, heart rate (BPM) and activity data.

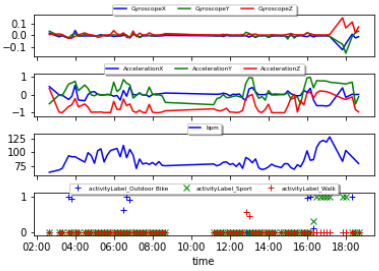

Fig. 2. Plots of gyroscope, accelerometer, heart rate (BPM) and activity data with a Δt of 10 minutes.

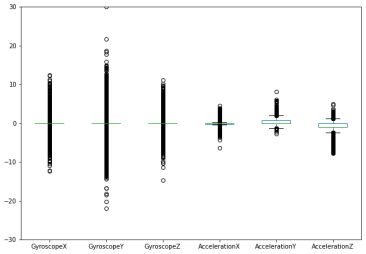

Fig. 3. Boxplot of gryoscope and accelerometer data.

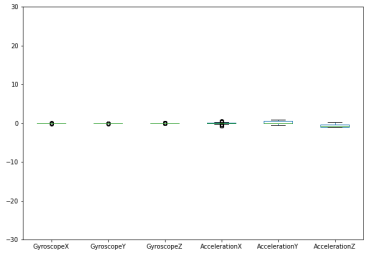

Fig. 4. Boxplot of gryoscope and accelerometer data with a Δt of 10 minutes.

Exercise 2.1b When considering a standardized Δt of 10 minutes for the frequency of measurements, the plots in Figures 2 and 4 can be made. Figure 2 shows a plot of the different features over time. When comparing this plot to the plot seen in Figure 1, it can be seen that the gyroscope and accelerometer data yield less extreme values. It can also be seen that the heart rate curve is smoother as a result of the fewer measurements. Figure 4 shows a boxplot of the gyroscope and accelerometer data. It can clearly be seen that the amplitude of these features is much smaller than the amplitude seen in Figure 3. Figure 6 shows a boxplot of the heart rate in beats per minute with the Δt of 10 minutes. It can be seen that the spread of the heart rate is smaller than the spread seen in Figure 5.

Exercise 2.2 In Figures 7 and 8, a comparison between the gathered date from Exercise 1 and the provided *crowdsignals* data has been made. This comparison is between the phone gyrometer and accelerometer data and the smartwatch heart rate and activity label data. The first observation that can be made is that the time frame, in which the data is collected, is larger for the Exercise

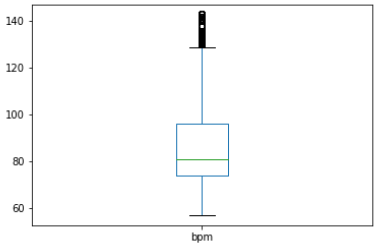

Fig. 5. Boxplot of heart rate data.

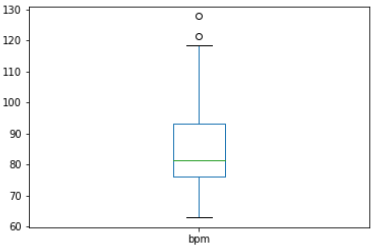

Fig. 6. Boxplot of heart rate data with a Δt of 10 minutes.

1 data. It can also be seen that the accelerometer amplitude is similar for the two datasets, where the gyrometer amplitude seems to be larger for the Exercise 1 dataset. Another observation that can be made is that the heart rate ranges of the two datasets seem to be similar. Lastly, it can be observed that the the *crowdsignals* dataset contains a larger array of activity labels than the Exercise 1 dataset.

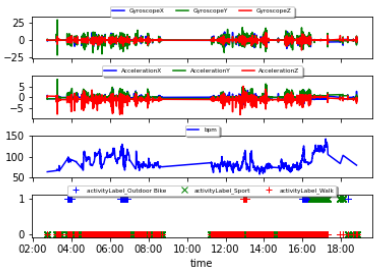

Fig. 7. Plots of the gyroscope, accelerometer, heart rate and activity label data that was gathered in Exercise 1.

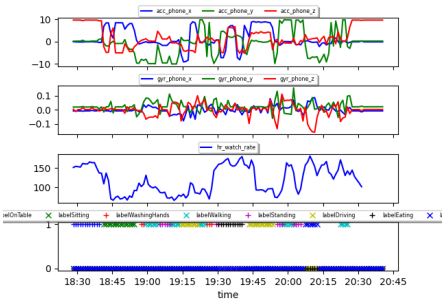

Fig. 8. Plots of the gyroscope, accelerometer, heart rate and activity label *crowdsignals* data set.

Exercise 3.2 Using Chauvenet's criterion, outlier detection is performed for *accelerometer phone - x*. Figure 9 shows the results of this outlier detection. It can be seen that a larger value of c leads to the labelling of less data points as outliers.

Outlier detection is also carried out for *accelerometer phone - x* using mixture models. The result of this is shown in Figure 10. It can be seen that there are no clear outliers.

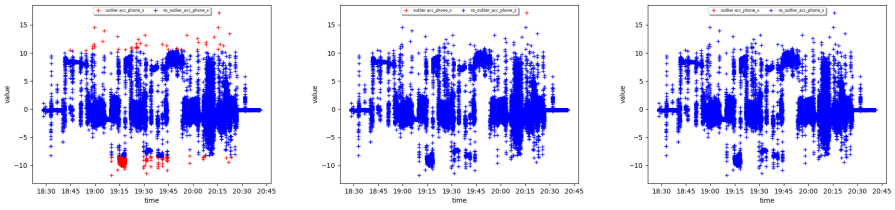

Fig. 9. Plot of the outliers for *accelerometer phone - x* according to the Chauvenet's criterion with different values of c. Left: $c = 0.5$, Middle: $c = 1$, Right: $c = 2$

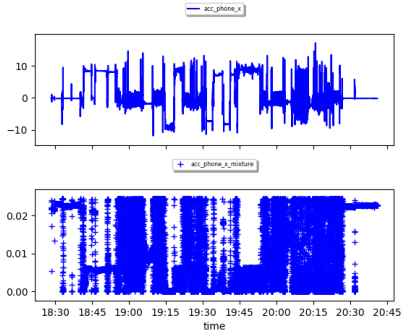

Fig. 10. Plot of the outliers for *accelerometer phone - x* using mixture models.

Simple distance-based outlier detection has also been applied. The result of this for multiple values of d_{min} and f_{min} are shown in Figure 11. In general, smaller values for d_{min} and f_{min} lead to the labelling of more data points as outliers.

Another method that has been applied for outlier detection is Local Outlier Factor (LOF). Figure 12 shows this for multiple values of k. Generally, a large value of k leads to a lower LOF factor, leading to the labelling of less data points as outliers. This is a result of higher robustness in the detection of outliers. Larger values for k were not computed due to limitations of computational resources.

Exercise 3.3 Figure 13 shows the heart rate values with no, mean, median and interpolated imputation, respectively. It can be seen that no imputation leaves a number of gaps in the data. For mean and median imputation, it can be seen that the missing values are filled with a value of approximately 121. These methods do not seem to be suitable for the imputation of the heart rate. It can also be seen that the missing values have been filled in an adequate manner using linear interpolation, approximately keeping the nature of the curve intact.

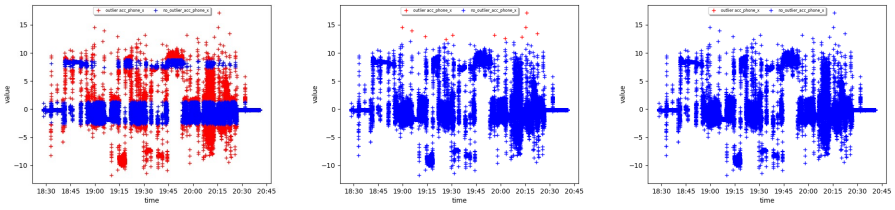

Fig. 11. Plot of the outliers for *accelerometer phone - x* using simple distance-based outlier detection. Left: dmin=0.05 fmin=0.8, Middle: dmin=0.1 fmin=0.99, Right: dmin=0.2 fmin=0.999

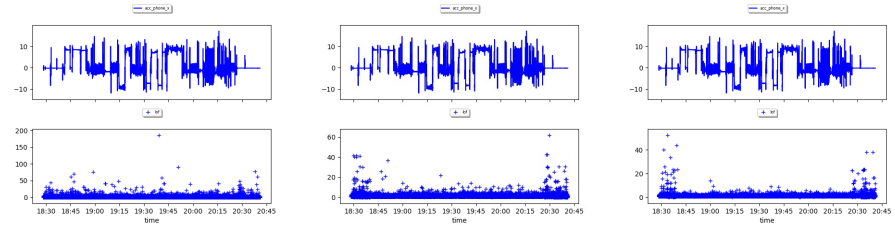

Fig. 12. Plot of the outliers for *accelerometer phone - x* using local outlier factor (LOF). Left: $k = 2$, Middle: $k = 3$, Right: $k = 4$.

Exercise 4.1 Figure 14 shows the frequency and amplitude of *accelerometer-phone-x* measurements from the *crowdsignals* dataset. The first plot of Figure 14 shows the most frequently-occurring measurement amplitudes. The second plot shows the weighted signal average and the third plot shows the power spectrum entropy. The last plot provides the labeled activities corresponding to the measurements.

It can be seen that the amplitude is low to zero for sedentary activities such as driving and sitting. These sedentary activities also yield low measurement frequencies. It can also be seen that the amplitude of the measurements is generally highest for the running activity. During running, the measurement frequency seems to be incredibly inconsistent. The weighted signal average provides little additional information, other than a few outliers.

Exercise 4.2 The two additional metrics that have been implemented are the median and slope. The result of this can be found in Figure 15. It can be seen that the median shows larger deviations than the mean, indicating that there are many extreme values, which are smoothed by the mean. This is the case for the three different window sizes. This characteristic may be useful when comparing different activities as the difference in amplitude will be clearer when using the median than the mean.

The slope shows the difference between windows, which can be used to identify transitions between activities or fluctuations within a certain activity. The

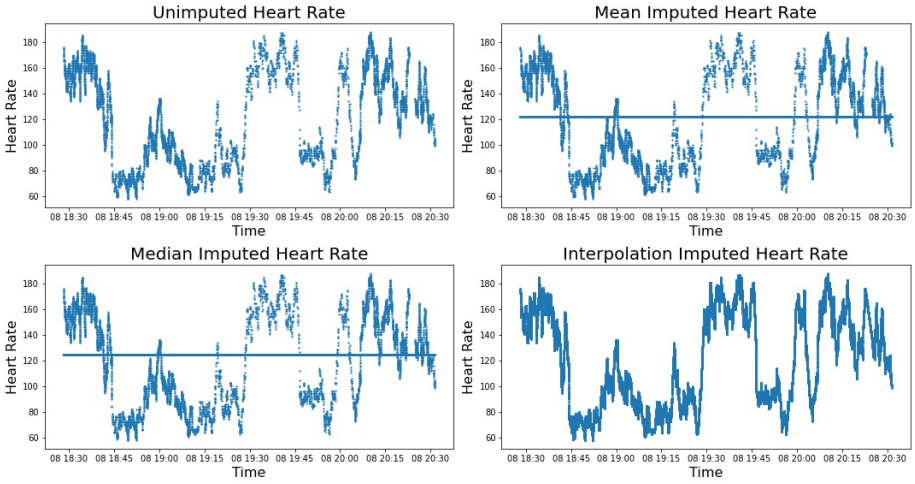

Fig. 13. Plot of the heart rate with different types of imputation for missing values.

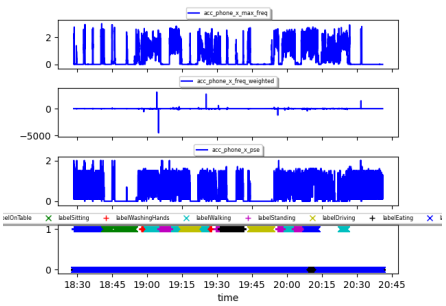

Fig. 14. Plots to show the frequency of *accelerometer-phone-x* measurements.

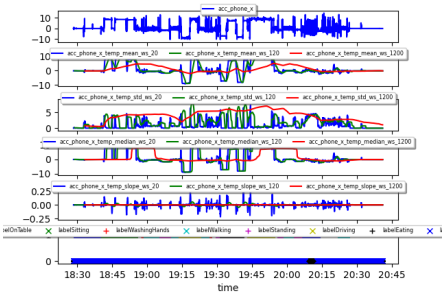

Fig. 15. Plots to show the frequency of *accelerometer-phone-x* measurements with two additional metrics: median and slope.

slope seems to be fairly consistent, especially when a larger window size is used. When consulting the graph for the window size of 1200 measurements, it can be seen that the line is approximately straight.

References

1. Aggarwal, C.C.: Outlier analysis. In: Data mining. pp. 237–263. Springer (2015)
2. Alshawabkeh, M., Jang, B., Kaeli, D.: Accelerating the local outlier factor algorithm on a gpu for intrusion detection systems. pp. 104–110 (01 2010). https://doi.org/10.1145/1735688.1735707
3. Gao, J., Hu, W., Zhang, Z.M., Zhang, X., Wu, O.: Rkof: Robust kernel-based local outlier detection. In: Huang, J.Z., Cao, L., Srivastava, J. (eds.) Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg (2011)
4. Hoogendoorn, M., Funk, B.: Machine learning for the quantified self. On the art of learning from sensory data (2018)
5. Jivani, A.G., et al.: A comparative study of stemming algorithms. Int. J. Comp. Tech. Appl **2**(6), 1930–1938 (2011)
6. Nils Ackermann: Sensor data recorder, http://www.sensordatarecorder.digitalmoom.de
7. Yan, Y., Cao, L., Rundensteiner, E.A.: Scalable top-n local outlier detection. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1235–1244 (2017)