Conformal Prediction and Explanation of a Fitbit Dataset

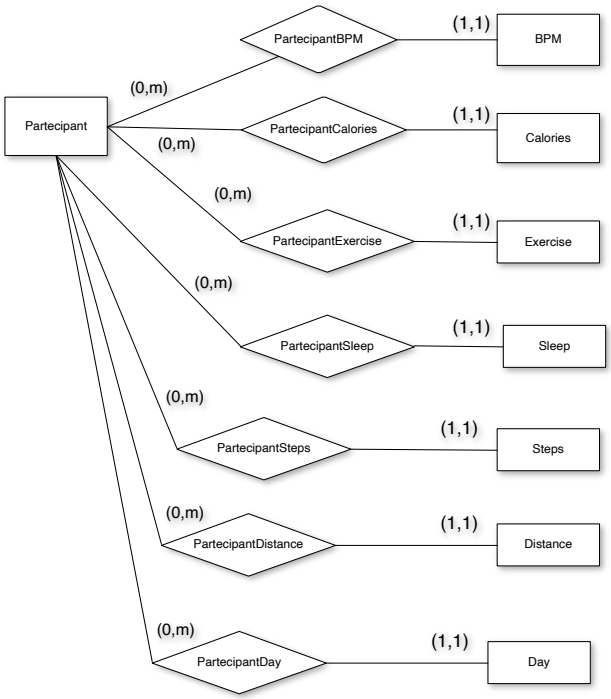
Pietro Sala
Department of Computer Science
University of Verona, Italy
pietro.sala@univr.it

✓ Checking the format

1 Goal

Forecast the (range) of sleep score for a fitbit user for the current date at given hours during the day (i.e., 8 A.M., 12 A.M., and 16 P.M.). Moreover, explain which features contribute to the prediction.

2 Data Profiling Task



Data Sources `https://datasets.simula.no/pmdata/`:
Select which features may be functional for the main goal from the following files (in the following we have $DD \in \{01, \ldots, 16\}$):

1. **Partecipant**, partecipant data from *participant-overview.xlsx*;
2. **BPM**, partecipant heart_rate from *pDD/fitbit/-heart_rate.json*; ✓
3. **Calories**, partecipant calories from *pDD/fitbit/calories.json*;
4. **Exercise**, partecipant exercise from *pDD/fitbit/exercise.json*;

data format is not ok

5. **Sleep**, partecipant sleep from from *pDD/fitbit/sleep.json*; start and end time what to do
6. **Steps**, partecipant steps from from *pDD/fitbit/steps.json*; ✓
7. **Distance**, partecipant distance from from *pDD/fitbit/distance.json*; ✓
8. **Day**, partecipant daily data from files:
   - *pDD/fitbit/lightly_active_minutes.json*, ✓
   - *pDD/fitbit/moderately_active_minutes.json*, ✓
   - *pDD/fitbit/resting_heart_rate.json*, ✓
   - *pDD/fitbit/sedentary_minutes.json*, ✓
   - *pDD/fitbit/time_in_heart_rate_zones.json*, ✓
   - *pDD/fitbit/very_active_minutes.json*. ✓

Input Parameters:

1. path: the path to the main folder of the dataset;
2. partecipants: a non-empty list of integers denoting the selected partecipants;
3. prediction time: an integer in $[0, 24]$ which provides the hour when the prediction occurs;
4. observation window: an integer denoting the number of hours before the prediction time to be considered for providing the prediction.
5. test size: a real in $[0, 1]$ represent the number of samples that are hold out for testing.

Output:
A collection of tables (e.g., pandas dataframes) that contains the values for selected features splitted into samples according to the partecipants, prediction time, observation window provided by the parameters. Moreover, this step produces a table providing, for each sample, the sleep score. The sleep score may be directly transferred into the sample value (in this case we are dealing with regression) or it may be turned into a class according to the following function:

$$sleep\_class = \begin{cases} excellent & sleep\_score \in [90, 100] \\ good & sleep\_score \in [80, 89] \\ fair & sleep\_score \in [60, 79] \\ poor & \text{otherwise.} \end{cases}.$$

Finally, here you select randomly a test size portion of all the samples which are not transferred to the Fit Task for the obvious purpose of doing the test set.

3 Fit Task

parameters: calibration size (real in $[0, 1]$)

Before fitting the task remove randomly from the train dataset a portion of calibration size samppleas that wil constitute the calibration set
This task is divided in the following substeps.

3.1 Run Length Encoding of Time Series

parameters: $ts(timeseries), sw(integer), step(integer)$ (+ cluster algorithm hyperparameters).

For each time-series $ts : [0, N] \to \mathbb{R}$ in a sample create a clustering function $cts : \mathbb{R}^{sw} \to [1, k]$ which maps each sliding-window of length sw of the time-series into a cluster index (e.g., $cts(ts(i) \dots ts(i + sw)) = k'$). Let $ts(i, sw) = ts(i) \dots ts(i + sw)$, given a step value $step \in [1, +\infty)$, the *run-length encoding of* ts (according to $(cts, sw, stwp)$) is the sequence $rts = h_1^{e_1} \dots h_r^{e_r}$ such that $N = sw + step * \sum_{1 \le j \le r} e_j$, $cts(ts(0, sw)) = cts(ts(step, sw)) = \dots = cts(ts(e_1 * step, sw)) = h_1$ and for every $1 < j \le r$ we have: (i) $h_j \neq h_{j-1}$; (ii) $h_i = cts(ts(s_j, sw)) = cts(ts(s_j + step, sw)) = \dots = cts(ts(s_j + e_j * step, sw)) = h_j$ with $s_j = sw + step * \sum_{1 \le j' < j} e_{j'}$.

3.2 Merge Features

parameters: None.

Transform each sample in an array x_i in $\mathbb{R}^M$ containing:

1. the run-length encoding rts provided at the previous step eventually padded with items 0^0 at the end for reaching the one with the maximum length;
2. the sequence of events $(e_1, vt_1), (e_h, vt_h)$ that occur in the sample where $vt_1 \le \dots \le vt_h$ are integers and represent the offset (e.g., in minutes) from the start of the observation window.

Output A dataset of pairs (x_i, y_i) where y_i is the sleep score to be predicted for the sample y_i.

3.3 Train

parameters: classifier/regressor hyperparameters.

Train the selected classifier/regressor on the training data then calibrate it on the calibration set.

4 Prediction Task

parameters: $\epsilon \in [0, 1]$

Using the test set check prediction accuracy or regressor mse, test for miscalibration using ϵ and explain the test set using SHAP.