**Generating Digital Markers from Wearable Sensor Data**

1. Collect input data. There are 16 features for each data point that correspond to raw sensor values.
   1. Sampled at 10Hz
   2. yaw
   3. pitch
   4. roll
   5. rotation rate (x, y, z)
   6. acceleration (x, y, z)
   7. location (latitude, longitude, altitude)
   8. accuracy (horizontal, vertical)
   9. course
   10. speed
2. Add a feature to each data point corresponding to whether the data point is an instance of a given class (feature value = 1) or not (feature value = 0).
   1. Each activity model is learned from all available labeled wearable data to date. If data were collected at a different sample rate, they are upsampled or downsampled as needed to be uniform for training the models. We will continue to refine these models as we receive more labeled data. Eventually, as labeled data are received for a particular user, we will want to weight them more heavily than data from others and automatically retrain the models on a regular basis as more data are collected and labeled.
   2. Current activity categories are (learned as individual oneclass classes, listed in this order in the file):
      1. Airplane
      2. Art
      3. Bathe
      4. Beach
      5. Biking
      6. Bus
      7. Car
      8. Chores
      9. Church
      10. Computer
      11. Cook
      12. Dress
      13. Drink
      14. Eat
      15. Entertainment
      16. Errands
      17. Exercise
      18. Groom
      19. Hobby
      20. Hygiene
      21. Lunch
      22. Movie
      23. Music
      24. Relax
      25. Restaurant
      26. School
      27. Service
      28. Shop
      29. Sleep
      30. Socialize
      31. Sport
      32. Travel
      33. Work
3. Add a feature to the current data point indicating the primary activity from among the list above. The output of this step is 52 features per data point: 16 raw sensor values + 33 current activities with 0/1 labels + date + time + activity. Current primary activity categories are (they are listed in this order in the file):
   1. chores
   2. eat
   3. entertainment
   4. errands
   5. exercise
   6. hobby
   7. hygiene
   8. relax
   9. school
   10. sleep
   11. travel
   12. work
4. Impute missing values. This ensures that there are continuous data round the clock. Values are computed by averaging over available data at the same time of day for most features. In the case of activity labels and location, the value is computed as the most frequent value. This will need to be updated because for the STTR study, data are not collected at night. We will need to decide whether to impute missing nighttime data, leave it blank, or infer they are sleeping in bed during that time. I am also not sure how we will know the watch is being charged versus being worn.

The result of this step is one feature vector for each minute of the day, starting with midnight on the date of the first raw sensor reading and ending with 11:59pm on the date of the last raw sensor reading.

1. Generate daily behavior markers. There is one set of 102 features for each day for each person and this is stored in file <ID>.day.
   1. Total rotation, summed over each minute of the day
   2. Total acceleration, summed over each minute of the day
   3. Number of missing values, determined in step #4
   4. Distance travelled
      1. Euclidean distance from one location to the next
      2. summed over the day
      3. Use latitude and longitude, not altitude
   5. Time spent on each oneclass activity (33)
      1. Number of minutes in which binary value for activity feature == 1
   6. Time spent on each primary activity (12)
   7. Time of first occurrence that day for each oneclass activity category (33)
      1. Minute of day in which binary value for an activity == 1 for the first time that day
   8. Time of first occurrence that day for primary activity (12)
   9. Time spent at each location type (4)
      1. Number of minutes in which binary value for location type feature == 1
      2. We map each (latitude, longitude) location onto a location type. To do this, we use
         1. The Nominatum open street map library, when this is accessible. This library returns an address and a location type
         2. A learned model that maps a location and surrounding sensor values onto a location type. This is learned from all previously-collected wearable sensor data for which we have the Nominatum information. This is continuously updated.
         3. Nominatum returns many locations types (>50). Right now we map these onto:
            1. House
            2. Road
            3. Work
            4. Other
   10. Time of first visit for each location type (4)
       1. Minute of day in which binary value for a location == 1 for the first time that day
2. Generate hourly behavior features. There is one set of 53 feature values for each hour of collected data and this is stored in file <ID>.hour.
   1. Total rotation, summed over each minute of the hour
   2. Total acceleration, summed over each minute of the hour
   3. Distance travelled
      1. Euclidean distance from one location to the next
      2. summed over the hour
      3. Use latitude and longitude, not altitude
   4. Number of missing values, determined in step #4
   5. Time spent on each oneclass activity (33)
      1. Number of minutes in which binary value for activity feature == 1
   6. Time spent on each primary activity (12)
   7. Time spent at each location type (4)
      1. Number of minutes in which binary value for location type feature == 1
3. Generate global set of 952 behavior markers.
   1. Statistics for each of the features generated in step #5 for day data, followed by statistics for each of the features generated in step #6 for hour data
      1. mean (102 values for day data, 53 values for hour data)
      2. median (102 values for day data, 53 values for hour data)
      3. standard deviation (102 values for day data, 53 values for hour data)
      4. max (102 values for day data, 53 values for hour data)
      5. min (102 values for day data, 53 values for hour data)
      6. zero crossings (1 value)
      7. mean crossings (1 value)
      8. interquartile range (1 value)
      9. skewness (1 value)
      10. kurtosis (1 value)
      11. signal energy (102 values for day data, 53 values for hour data)
   2. regularity index (9)
      1. Compute using only the hour data generated in step #6
      2. Calculated within weeks, within weekdays, and between weeks
      3. Applied to total acceleration, total rotation, total distance
      4. The regularity index assesses the difference between the same hours across two different days. First rescale the data to [-0.5, 0.5]. The product of two rescaled values is positive if the original values are close and negative if they are not similar. The regularity index between days a and b is then defined as , where *T* = 24 hours, and is the rescaled value in hour *t* of day *a*.
   3. circadian rhythm (3)
      1. Compute using only the hour data generated in step #6
      2. Applied to total acceleration, total rotation, total distance
      3. Circadian rhythm measures the strength with which a person follows a 24-hour rhythm as exhibited in the behavioral data. We estimate a power spectral density using a periodogram, which generates values for a large set of possible cycle lengths (in numbers of hours). The circadian rhythm value is the normalized periodogram-derived value for 24 hours.
4. Generate set of behavioral change scores comparing the baseline (first week of day after removing first day which is typically incomplete) with the following weeks of data. There are two values for each week.
   1. Change score calculated using small-window Permutation-based Change in Activity Routine.
   2. Binary value indicating whether change is significant (1) or not (0).