Exercise 3: Feature extraction and selection

Data Description:

Exercise 3 deals with Smartphone-Based Recognition of Human Activities and Postural Transitions Data Set. More information about the dataset here: http://archive.ics.uci.edu/ml/datasets/smartphone-based+recognition+of+human+activities+and+postural+transitions

All the participants were wearing a smartphone (Samsung Galaxy S II) on the waist during the experiment execution. This dataset includes 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz using the embedded accelerometer and gyroscope of the device. The dataset is divided in two parts: Raw Data folder includes the raw acceleration and gyroscope files for each subject and each experiment, Training and Testing folders include a ready-made machine learning dataset, where the subjects have been randomly divided to training and testing sets, and many features have been calculated for each annotated activity.

Human activities and postural transitions mean, that the following activities have been recorded and annotated:

1. walking
2. walking upstairs
3. walking downstairs
4. sitting
5. standing
6. laying
7. stand to sit
8. sit to stand
9. sit to lie
10. lie to sit
11. stand to lie
12. lie to stand

Acceleration describes the rate of change of velocity in three dimensions (x, y, z) with respect to time.

Acceleration and gyroscope tell many details about physical activities, for example about the symmetricity of the activity, which direction there is the most changes, how “steady” (= less acceleration) the activity is etc. You may consider basic physics courses when thinking about how each activity type would show in acceleration or gyroscope signals.

The dataset (HAPT Data Set.zip) may be downloaded from Moodle or from the original source archive.ics.uci.edu/ml/datasets/smartphone-based+recognition+of+human+activities+and+postural+transitions. The file structure is described in README.txt.

Tasks

1. Randomly choose one of the triaxial acceleration data files. Make a plot of the raw acceleration data (x, y, and z channels) vs time (use either sample number from the data file or transform it into the real time in s).

   Question: Is it possible to visually extract different activities from the raw data? Why do different activities have similar or different features in the raw data? Suggest a method to identify activities from the particular plotted data file.

   Suggested implementation:
   - use *pandas* for data manipulation (https://pandas.pydata.org/docs/user_guide/index.html)
   - use *DataFrame.loc* propertie (https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.loc.html)
   - use *Matplotlib* for plotting (https://matplotlib.org/stable/tutorials/introductory/pyplot.html)
   - use *pyplot.vlines* for visualizing beginning of the activieties (https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.vlines.html)

2. Let's extract some features that could be useful in separating walking segments (classes 1-3) from other activities (classes 4-12). Purely data-driven machine learning usually employs statistical features, such as mean, median, minimum, maximum, etc. in tasks for classification or regression. It may be hard to know beforehand which features are the best ones. Thus, usually machine learning projects calculate a lot of features at first, and then do a feature selection algorithm to pick the features that give the highest accuracy without over-fitting.

   Select a few features (= 3-4) that you want to calculate and make a script that calculates them for each segment concerning the activities of one single subject and channel in the example file. You can for example select features that have been calculated in the dataset (see background info) or choose other features you find somewhere else.

   Question: What features did you choose, why? No need to present any scientific proof but more of your own reasoning.

   Suggested implementation:
   - Create a function to implement the task, as in the next question the same will be calculated for all the subjects (https://docs.python.org/3/tutorial/controlflow.html#defining-functions)
   - Calculate the statistical feature for each of the spacial coordinates of raw data from the start point to the end point of the activity.

3. Extract the same features as in task 2 of the current exercise from all the ACC files in the Raw Data. Concatenate your results, e.g., create one big dataframe.

   Question: Are the features extracted for different activities the same at different subject? What trends may be observed? What may be the reason for variations?

   Suggested implementation:
   - use *pandas.concat* to concatenate pandas objects (https://pandas.pydata.org/docs/reference/api/pandas.concat.html)