Details of the real-world problem

Currently, more than 5 billion mobile devices with several sensors (e.g., accelerometer and GPS) are in use, capturing detailed, continuous, and objective measurements on different aspects of our life, including physical activity. Such widespread smartphone adoption provides unparalleled potential for data collection to study human behaviour and health. Smartphones, with appropriate storage, strong processors, and wireless transmission, may collect massive amounts of data about huge groups of people over long periods of time without the use of extra hardware or instruments.

This project requires proposing a solution using ML to perform human activity recognition (HAR). HAR is a process aimed at the classification of human actions in a given period of time based on discrete measurements (acceleration, rotation speed, geographical coordinates, etc.) made by personal digital devices. In order to make an informed decision, you should develop and test several ML models before suggesting a final approach to solve the task. You will be supplied with a database that consists of data collected from 36 different users performing six types of human activities (ascending and descending stairs, sitting, walking, jogging, and standing) for specific periods of time. These data were acquired from accelerometers, which are able of detecting the orientation of the device measuring the acceleration along the three different dimensions. They were collected using a sample rate of 20 Hz (1 sample every 50 millisecond) that is equivalent to 20 samples per second.

Database Description

The database contains information about 36 users who performed six different human activities, including ascending and descending stairs, sitting, walking, jogging, and standing. The data was collected using accelerometers, which measure acceleration in three dimensions and determine device orientation. The time series data was divided into 10-second snippets and statistical features or metadata were extracted from each snippet. You will have access to both the metadata and the time series, so you can choose to build ML models based on either one. Additionally, you may consider extracting additional features from the time series.

Description of the files and folders

• ‘signals.csv’ file – contains the time-series data organised by users and snippets.

• ‘signals\_test’ file – contains the time-series data corresponding to the test subset. It follows the same structure as the ‘signals.csv’ file.

• ‘signals\_kaggle’ file – contains the time-series data corresponding to the kaggle subset. It follows the same structure as the ‘signals.csv’ file.

3

• metadata.csv –contains features extracted from the signals and the target column (activity)

• metadata\_test.csv – contains the same columns as ‘metadata.csv’.

• metadata\_kaggle.csv – contains the same columns as ‘metadata.csv’ but without the target column.

• predictions\_example.csv - A sample of the predictions file using the correct format for the Kaggle competition.

1. The “signals.csv” file has the following columns:

• user\_snippet: User id + snippet id.

• timestamp: in milliseconds.

• x-axis: The acceleration in the x direction as measured by the phone's accelerometer. Values are floating-point between -20 and 20. A value of 10 = 1g = 9.81 m/s^2, and 0 = no acceleration. The acceleration recorded includes gravitational acceleration toward the centre of the Earth, so that when the phone is at rest on a flat surface the vertical axis will register +-10.

• y-axis: same as x-axis, but along y axis.

• z-axis: same as x-axis, but along z axis.

2. The “metadata.csv” file is a CSV file with the following columns:

• user\_snippet – the identifier of the user who acquired the data (integer) + the snippet identifier (integer).

• activity – the activity that the user carried out at the corresponding snippet. This is the target, which could be one of the following class labels:

o walking

o jogging

o sitting

o standing

o upstairs

o downstairs

• Extracted features – the rest 30 columns correspond to 10 features extracted from the x, y and z acceleration time series. Each column name has the format D-axis\_\_FEATURE, where D is either x, y or z; and FEATURE is one of the following:

o sum\_values

o median

o mean

o length

o standard\_deviation

o variance

o root\_mean\_square

o maximum

o absolute\_maximum

o minimum