EEE4114F: Lab Assignment 1

Jarryd Son

1 Introduction

In this assignment you will be working with the MotionSense dataset. As mentioned before the ML lab assignments are designed to form a basis for the course project.

In this lab you will be using some machine learning and data science libraries to understand and manipulate the data. Getting to grips with the data is a crucial step in developing machine learning solutions.

The MotionSense dataset includes the following description on the Github page:

> This dataset includes time-series data generated by accelerometer and gyroscope sensors (attitude, gravity, userAcceleration, and rotationRate). It is collected with an iPhone 6s kept in the participant's front pocket using SensingKit which collects information from Core Motion framework on iOS devices. All data collected in 50Hz sample rate. A total of 24 participants in a range of gender, age, weight, and height performed 6 activities in 15 trials in the same environment and conditions: downstairs, upstairs, walking, jogging, sitting, and standing. With this dataset, we aim to look for personal attributes fingerprints in time-series of sensor data, i.e. attribute-specific patterns that can be used to infer gender or personality of the data subjects in addition to their activities.

Question 1 [10 marks]

Based on the information you can gather from the Github page and other related sources please answer the following questions with regards to the (A) DeviceMotion_data set:

a. How many input features are present in this dataset? (1)

b. Are these raw sensor values or have there been pre-processing steps? (2)

c. Are there any input features that you think would not be necessary or are redundant for human activity recognition? (2)

d. How many different target classes are provided in this dataset for classification? What are they? (2)

e. How many examples are in this dataset if you were to consider each trial an example? (1)

f. Suppose you want to increase the number of examples by slicing the trials into fixed windows of time. e.g. instead of one long recording per activity, divide it into multiple smaller recordings. What do you think potential issues with this approach might be? (2)

Question 2 [10 marks]

For this question you may make use of existing libraries (such as scikit-learn) for resampling data into different splits. Correct resampling is crucial to providing unbiased estimates of a model's performance.

a. Implement a function to divide the (A) DeviceMotion_data set into smaller windows with a fixed length. (5)

b. Implement a function that takes in the windowed dataset and a splits it in a three-way holdout with variables for validation and test sizes. (2)

c. Implement a function that takes in the windowed dataset and splits it multiple times for k-fold cross-validation. (3)