096224: Distributed Database Management

Background

In this project you are called to help the "Distributed Human Health Association".

The "Distributed Human Health Association" is a health association that aims to improve the lifestyle of humans in different ways. Recently, the association have started a project that works on improving the health of people and combining different activities in their lifestyle.

In order to investigate different lifestyles, the association gathered information about the daily activities of different people using smartphones and smartwatches.

You are requested to preform several tasks to help the "Distributed Human Health Association".

Data

The dataset contains the readings of motion sensors commonly found in smartphones of several users. The readings were recorded while users executed activities scripted in no specific order carrying smartwatches and smartphones.

The data consists of the following columns:

- Index: The row number.
- Arrival_Time: The time the measurement arrived to the sensing application.
- Creation_Time: The timestamp the OS attaches to the sample.
- x,y,z: The values provided by the sensor for the three axes, x,y,z.
- User: The user this sample originates from, the users names range from a to i.
- Model: The phone/watch model this sample originates from.
- Device: The specific device this sample is from. They are prefixed with the model name and then the number, e.g., nexus4_1 or nexus4_2.
- Gt: The activity the user was performing: bike, sit, stand, walk, stairsup, stairsdown, and null.

The timestamps were measured according to unix time (number of seconds passed since 1970).

You will be working in this project with 2 types of the data: static and dynamic. In parts 1 and 2 you will be working on the static data, meaning that you have all the data beforehand and the data will not change while you are working on it.
In part 3 you will work with dynamic data, meaning you will need to use spark streaming in order to read the data and the data will change while you are working on it.

The dataset was published by:
Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen "Smart Devices are Different: Assessing and Mitigating Mobile Sensing Heterogeneities for Activity Recognition" In Proc. 13th ACM Conference on Embedded Networked Sensor Systems (SenSys 2015), Seoul, Korea, 2015.

Assignment

Data Analysis (35 Points)

Investigate the data and derive 3 insights from the data. Try to think of insights based on different aspect (time dimension, space dimension, etc). Use visualizations where applicable.

What to submit?

1. Jupyter notebook (.ipynb) containing your analysis and explanations (use Markdown cells to write text). File name: project2_part1_[ID1]_[ID2].ipynb
2. html file (.html) of the notebook. File name: project2_part1_[ID1]_[ID2].html

[ID1] and [ID2] are the ids of the students doing the project.

Learning Task (30 Points)

Use the insights from the first part and build a Machine Learning model which utilizes the data from your data store.

1. The task is to predict the activity variable ('gt').
2. Define your training set and testing set (from your data store).
3. Create (build and train) a model using Spark MLlib. You can use a model that is composed of several different models and utilize different learning techniques.
4. Evaluate your model using Accuracy (on train and test sets). If possible, through visualizations. Your model should get at least 40 percent accuracy when training on 70 percent of the data and testing on 30 percent of the data.

What to submit?

1. PDF file containing your learning task description. File name: project2_part2_[ID1]_[ID2].pdf
2. Python file (.py) containing your model implementation. File name: project2_part2_[ID1]_[ID2].py

Extract and Learn (35 Points)

Use Spark Streaming to extract the data from the data source. While streaming, train the ML model you built in the previous task, while simultaneously testing it. Show the model's performance throughout the streaming process.

Constraint: You can not predict data that your model were trained on and you must predict all the data (i.e. you can't train the model on streamed data and then test the model on the same data).

What to submit?

1. Python file (.py) containing all the process you did. File name: project2_part3_[ID1]_[ID2].py
2. PDF file containing your training process and accuracy throughout the streaming. File name: project2_part3_[ID1]_[ID2].pdf

Bonus - Competition (10 Points)

You will be measured for the accuracy you get in the "Extract and Learn" part.
Feel free to use any method you like. Note that rather than choosing the perfect ml model, you can improve your score by: Feature Engineering, Data Augmentation, Adjusting Hyperparameters, etc.
Training a model with the latest data can also improve your results, but keep in mind the previous part constraint: you can not predict a data that your model were trained on.

What to submit?

No need to submit another file here, your results in the previous part will determine your score.

General Guidelines

- Use *Spark 3* and above.
- Your final submission should be a zip file (.zip) named project2_[ID1]_[ID2].zip. If the project is done in a group of 3 students, replace [ID1]_[ID2] in the submitted files and folder with [ID1]_[ID2]_[ID3]. The zip file should contain 3 folders: part1, part2, part3. Each folder should contain the matching part's files. If you chose to implement to bonus task, you should submit another folder named bonus in addition to the 3 other folders.
- Write clean code and document functions when necessary.
- Questions related to the project will be answered in the forum, via Moodle, exclusively.
- Only one of the team members need to submit.
- Submission is due to June 30, any delay in submission will result in a reduction of 20 points from the final score of this part.

Good luck,
Course staff.