Project idea

The goal of the project is to give you an opportunity to apply your knowledge on a real data set and to get experience in extracting useful information from the data. To this end, the project is very open-ended: there is no definitive task to be achieved. Your goal is to analyse the dataset and extract as much useful knowledge from it as you can.

You may choose to perform the project individually, or in groups of up to two students.

Domain description

This is a data set from an online exercise diary http://www.funbeat.se/, where people register their trainings. Athletes may enter information manually, but we will focus on data uploaded using sports watches. Data entered includes at least the duration of the training and the type of sport. If people have a more advanced sport watch they can also upload data on heart rate, speed, GPS position, altitude and even more specific characteristics like cadence, power output (in cycling).

Not all files contain all of this information, depending on the configuration and the model of the sports watch. According to our contact person, most training sessions at Funbeat do not have an attached file (i.e. no sports watch is used). The complete database consists of roughly 45 million training sessions, registered between 2005 and 2016. Approximately 3 million of them include data from sports watches. The popularity of different sports also varies greatly. For example, among those 3 million, there are roughly 1 million training sessions in the sport “running”.

One possible question for you to explore is whether it is possible to predict finish times on particular distances, such as 10 km, 21.1 km and 42.2 km (marathon/half marathon) based on factors such as training volume, age or gender. For example:

- What is the correlation between marathon finish time and the training volume (total hours) per week? What is the correlation with age?
- Is there a significant difference between finish times for males and females? How large is it?
- Which factors (individually or combined in different ways) give the best prediction of marathon finish times? E.g. gender, age, total training volume, total running training volume, average training speed, sessions per week, number of different sports?

These questions can all be asked for different distances (e.g. 5 km, 10 km and 21.1 km). There are, however, several caveats that need to be taken into account:

- Athletes probably pay extra attention to starting and stopping the time when they race, but in general neither distance nor time not guaranteed to be accurate. One could assume that the distance can deviate a little bit (maybe 1%).
- It is not necessarily clear how to distinguish training session from actual competition.
- People sometimes enter the wrong sport, and it can be difficult to define “sport”
- Data from the watches can be noisy
- Athletes forget to stop the watch at the end of training
- People other than the owner sometimes use/borrow the watch
- …

However, in your project you will need to address more than just this one question.

Data specification

We have access to approximately 1.7 million files. The total size of the data is over 500GB. You do not need to analyse all of it in this project, of course. Some statistics concerning available data can be found in TrainingStats.csv file.

Initially a small sample of ~200 files have been put on the Blackboard. Those files come in .json or .tcx (which is a variant of XML) formats. Both should be quite easy to parse and have a pretty regular structure.

Those files are named using the following syntax.
`fileId_parentTrainingType_trainingType_personId_gender_birthYear_year_month_duration.format`

`fileId`

An uninterrupted series from 1 to 1829498

`parentTrainingType and trainingType`

Identifiers of different sports, explained in TrainingTypes.csv file

`personId`

Unique id per person
Obfuscated so it is not traceable to FunBeats member ids, but still useable for you to group the data.

`gender`:

F = Female
M = Male
X = Unknown

`birthYear`:

0 = unknown

`year` and `month`:

When the training session took place

`duration`:

Duration in minutes
If this differs from the duration parsed from the files it means that the member has altered the duration manually at FunBeat. This is common when the member forgot to stop the watch after the training session, i.e. the file shows a faulty duration. You can use this duration to "cut off" data in the files if you want to use only data from the training sessions.