CS-C3240 MACHINE LEARNING - PROJECT - STAGE 1
Running pace preditcor

Aalto University

March 7, 2022

1 Problem formulation

The purpose of the derived machine learning algorithm is to yield a prediction model for the next running pace during the training activity of one individual athlete in minutes per kilometre at the instance of each completed kilometre. This then provides a guideline for the runner to either increase or decrease the current pace, based on the training goals. During the project the python libraries [1] and [2] were used.

2 Description of the dataset: Datapoints, label and features

The data set used in the project consists of 42 "running activities" obtained from a "Garmin 735XT" [3] smart watch, which was provided by the athlete. Besides high resolution time series the watch provides a template of numerous metrics with average values for every kilometre after finishing one training session. The latter can be exported as an .csv- file for every activity and is used in the project. To only take into account the current fitness condition of the athlete, it was opted to utilize a training period of the last 6 months, hence from the 1^{st} of August 2021 to the 1^{st} of February 2022
One datapoint is the exact instance in time after a full kilometre has been finished, leading to a total amount of 395 datapoints before the data cleaning and manipulation process.
The data of the first kilometre, during which the athlete is still in a warm up phase, is most often not very representative for the whole run. Therefore, it is omitted from the dataset. A similar justification can be made for the neglect of the last kilometre of an activity, which rarely constitutes a "fully completed kilometre".
As can be read from section 1, the label y_i of the machine learning problem is the average pace during one kilometre in minutes per kilometre.

3 Feature selection

Each activity of the above mentioned dataset contains 11 columns of metrics stored for each kilometre. After preselecting the most important metrics using expert knowledge, the basic feature set for the prediction is composed of the following data:

- X_1: Cumulative completed kilometres during one activity prior to the datapoint in km
- X_2: Cumulative ascent during one activity prior to the datapoint in m
- X_3: Average pace of all previously completed kilometres prior to the datapoint in min per km
- X_4: Average heart rate during all of the previous kilometres prior to the datapoint in bpm
- X_5: Pace of the previous kilometre in min per km prior to the previous datapoint
- X_6: Ascent of one kilometre in m prior to the previous datapoint
- X_7: Average heart rate during the previous kilometre in bpm prior to the previous datapoint

The metrics of the kilometer prior to the previous datapoint ($X_5...X_7$) allow the consideration of the recent physical load of the athlete, whereas the average and cumulative features ($X_1...X_4$) provide knowledge about the whole training activity. However, the various features was combined to one single feature by multiplying their numeric values for each datapoint. This yields a unitless, empirical feature, which is called "physical load" from now on. It is not important to obtain the precise correlation of all the listed features with the label, but rather to predict a valid numeric value for the label. Therefore, this procedure is justifiable.

4 Data cleaning and manipulation of the dataset

The following data cleaning process was conducted before implementing a machine learning model. For the features X_5 and X_6 some of the datapoints are filled with a "–" string (NaN), when the corresponing integer equals 0. These values are replaced by a integer of 0.1. the value is not set to 0 as otherwise the physical load of all the datapoints would be set to zero during the multiplication step. This would not be realisitc. Additionally, some unrepresentative datapoints are removed at a certain threshhold, after looking at the scatter plots of all of the features. These upper boundry values werde determined with the athelete and expert knowledge and can be summarized as follows:

$$X_2 = 250m; X_3 = 6\frac{min}{km} \tag{1}$$

Furthermore, some minor changes like changing the pace format from min:sec per km to xx,xx min per km, Calculating the cumulative values for each activity and Shifting the pace and ascent data to obtain the "previous" values X_5 and X_6 were done. After plotting the scatterplot feature X_6, however, it can be observerd that it does not provide any value to the machine learning problem due to lack of correlation. Therefore, it is dropped again. Finally, all single features ($X_5...X_7$) were multiplied as described above, yielding the physical load X. This must be normalized to obtain a parameter range from 0% to 100%. The normalization is conducted as follows [4]:

$$X_{i,norm} = (X_i - X_{min})/(X_{max} - X_{min}) \tag{2}$$

5 Polynomial models, Loss-function and validation

It can be comprehended easily that there must be a nonlinear dependence of any kind between the label y and the feature X (physical load). However, the model is not expected to be highly nonlinear. These assumptions can also be vailidated with the scatterplot in the appendix.
Therefore it was opted to investigate the polynomial hypothesis space as follows [5]:

$$\mathcal{H}^n_{poly} := \{h(\mathbf{X}) = \Phi(\mathbf{X})\} \tag{3}$$

Where

$$\Phi(\mathbf{X}) = (1, X, ...X^n) \tag{4}$$

Hence, the linear hypothysis h is used in combination a feature map $\Phi(\mathbf{X})$. The maximum degreed of the polynomial models n is sset to 6, in order to be able to investigate many models with little extra effort.
The corresponing sk-class in python [2] uses the ordinairy mean squared error (MSE) as a loss function, which is also common in numerical machine learning problems. The MSE provides good statistical properties and is convex. Therefore the empirical risk can be minimized via gradient descent. [book]
In addition, it was analysed, if regularization can provide any improvement of the accuracy of the models. It can be seen from the appendix, that regularization only has little to no effect on the model. This was done, because the

6 Sources

[3] Garmin (2022): Forerunner® 735XT. Garmin Ltd. Available online at https://www.garmin.com/fi-FI/p/541225#specs,updatedon2/7/2022, checked on 2/7/2022.