CS221 Project Proposal

Matt Millett and TJ Melanson

1 Introduction

Runners are a curious, anxious bunch. Many of them record and plan out their weekly pace, workouts, and mileage. However, right now, there's no rigorously data-driven way to tell someone how they're going to race based on this data. The runners will also have data from different races that might be able to give them a much better idea of how they will perform.

2 Inputs and Outputs

Inputs: Previous race and workout data for individual athletes. E.g. Marc Scott has raced the 8K many times and has race times of 23:35, 23:30, 24:43, and has many other times for other races. He does workouts like mile repeats (4x1600 at 4:50 mile pace). We hope to aggregate these inputs to figure out a level of fitness which will help generate our outputs.
Outputs:
1) What race times this athlete could run right now for particular distances (e.g. Scott might be able to run a 1:55 for the 800m)
2) Workout recommendations based on a specific distance goal and your previous training (e.g. Hal Higdon's marathon plan for beginners or advanced, but more detailed: on Monday, run a 5x1600 at 5:50 mile pace with 1 minute rest. Tuesday, do a 4 mile distance run. Etc.).

3 Topics/Techniques

For the time-prediction problem, we plan to classify runners into different levels of fitness based on their performances in races and workouts. Fitness involves many things– a sprinter and marathoner can be equally fit, but in different ways (speed/power vs endurance). We plan to use different features such as race and workout distance and speed to differentiate these different modes of fitness and run a classification algorithm on them (perhaps linear regression). Once we have separated the runners into these buckets, we can some kind of continuous regression on the runners' times in each bucket to see what times the "typical" runner in each bucket runs. This will ultimately be our prediction, where we will attempt to predict what someone in each bucket will run for typical races of length 800m - marathons.

The other part of this project is finding a way for runners to jump from one level of fitness to the next via training and workouts. To do this, we plan to use a state-based prediction model, where we model states as levels of fitness and actions as workouts. The state space could include things such as age, weight, height, race time, estimated vO2Max, and other additional features apart from simply fitness level.

Actions would be a series of workouts (which we would suggest to eventually reach a state of better fitness for a particular distance). Running workouts can be classified by intensity, which is modulated principally by 3 things: 1) distance, 2)speed, and 3) rest time between intervals. Distance runs also fit in here "rest between intervals" = 0. We would essentially find a way to take a series of actions (including runs, workouts, and rest) to reach a desired race time/level of fitness. We may have build this as a Markov problem because there are also probabilities of injury involved, which could greatly reduce fitness.

4 Data and baselines

4.1 Preliminary Data

We scraped data from Flotrack's website on the top 16 runners in the running NCAA Division I Cross-Country and took all their race times from the past 4 seasons. Interestingly, their fitness for various distances varies more once we move away from the relatively standard 8k.

4.2 Baselines

Time prediction: Peter Riegel, an engineer, developed a widely used time predictor based on a single race: Take the race you want to run of distance D2 and the most recent time you have run T1 for a race of distance D1. The time T2 can be described as $T2 = T1(\frac{D2}{D1})^{1.06}$. In other words, $t = ad^{1.06}$, where t is the time, d is distance, and a is some constant factor. This works well with races that are close in distance to each other (say within 1000 meters or about 20 to 30 percent), but fails to work accurately for disparate distances (for example, the 800 meter and the 8000 meter run). This proves problematic when track runners want to train for longer distances such as the marathon (42000 meters).

The baseline we use bases the new race off of the most recent race run, indirectly setting a to whatever would make the most recent distance be proportional at the most recent time.

5 Oracle and Baseline Results

To test our baseline and oracle, we had each predict times and recommend workouts for each athlete. We gave them all the race data from one athlete, excluding data from one distance, and asked them to predict the athletes' times in the new distances the predictors had not seen.

5.1 Baseline

Average error:11.71%

The baseline would recommend workout intensities based on their most recent race- i.e. if they ran a 1:45 800m (hard-coded to be considered "fast"), even if they were training for a 10k, they would still be seen as fast and recommended a "hard training plan". We have excel spreadsheets documenting these suggestions, but one example is Amar Moussa was predicted to run 15:15.91 for the 5k and was given a training plan level of "medium; intermediate training plan".

5.2 Oracle

Average error:9.73%

Our oracle was the Running Club coach, who unfortunately didn't have time to recommend workouts.Her average error was 9.73%.The oracle here is googling different workouts for particular distances. For instance, someone training for the 1600 could use the Oregon training program to train for that distance, putting in their times so the paces could be adjusted.

Also, for workouts, we will plan different paces based on race times for different lengths: hard could be 2-mile or 5K race pace, moderate could be 10K, and long run would be marathon/ultramarathon pace.