Do you even Lift Bro?

Predicting Athletes' Sports and Proficiency based on Exercise Regimes

Matthew Katzman - mkatzman,
Christina Ramsey - cmramsey,
Samuel Sowell - alex4936

December 16, 2016

Contents

1 Background

According to the Bureau of Labor Statistics, only 16% of people in the United States over the age of 15 participate in some form of physical exercise daily [3]. This number is quite low, and is often cited as a large factor in the high obesity rate within the United States. It is becomming crucial to motivate people to exercise.

Whether through activities such as running, weightlifting, cycling, or playing more traditional sports like basketball, people in the United States need to be more active. However, a primary obstacle for many people is the lack of motivation to exercise and lack of goals to set. We believe that, if people were training for a sport able to compare their performance to others, then they might have more motivation to exercise. Herein lies the inspiration for our project.

2 Introduction

The high-level aim of this project is to accept user data to determine the most likely sport a user might enjoy (and might already be preparing for). This prediction would be based off of information containing details about age, gender, height, weight, and workout history. Training this predictor on a dataset containing over 100 sports and over 1000 exercises yields a fairly comprehensive set of weights to analyze.

These weights, in turn, provide valuable information on exactly which exercises and characteristics might be more, or less, suited to certain sports. Using these weights, and approximating performance by a normal distribution on "caluclated 1-rep-maxima", we then are able to take the weighted average of percentiles in each exercise to predict how a user might compare to others in their sport.

3 Related Work

While the field of athletics is rich with predictive algorithms, the focus is on the outcome of a specific sport. As such, we found next to no work similar to our project. We instead intended our results as an expansion of the work done by David Jurgens, James McCorriston, and Derek Ruths in their paper "An Analysis of Exercising Behavior in Online Populations" [2].

Using data from the fitness-based social media platform "Fitocracy", their paper analyzes a wealth of data, ultimately concluding that patterns do emerge in the dataset, particularly involving motivation, biased toward gender and age. We base our work on these results, hoping to reverse engineer motivation from the data.

4 Dataset and Features

4.1 Dataset

We used the Fitocracy dataset provided by David Jurgens covering the Fitocracy usership between February of 2011 and January of 2015 [2]. Because the population is self-selected, we decided that it was reasonable to treat the data as factual. The dataset was extremely large and contained the following relevant information on each user:

- Age
- Gender

- Height
- Weight per rep for each set of every exercise performed (or similar information for weightless exercises, such as length of time per run)
- Groups the user participated in (each group was associated with specific sports the user, in turn, must play)

An example of data format showing demographic information for each user ID:

```
1       FRED    11      13,700  7,577   29      5'0"    m16
577     29      5'0"    m1000   kethonna        1       34
1000091 scubed427       2       109     0       21      Nor
str_pwr 6       3,656   0       None    5'6"    m1000210
er      7       4,167   3       41      5'5"    f1000368
11      11,692  5       25      5'5"    f100049 daisygirl97
helle75 1       43      0       39      5'2"    f100062 fre
3       397     0       37      6'2"    m1000748        fit
36      306,379 83      30      6'1"    m1000891        zer
0       37      5'10"   m100100 curvvy63        2       165
1141    MistyMeadows    5       2,541   0       None    Nor
eBennie 5       1,833   0       28      16'8"   f1001371
1       5'7"    f1001525        erin_messmer    2       238
1001680 amandalynn13    5       2,261   0       None    Nor
6,378   68      32      5'10"   m1001774        Alfa_w  7
menlala 5       2,246   2       20      5'3"    f1001889
0       60      5'5"    f1002022        andrewalt2511   5
22      5'3"    f1002188        hawkeyecowgirl  32      168
2"      f1002397        AmandaSchmidt   19      42,383  0
5"      f1002514        dejesusj        4       917     0
5"      f1002639        leahloo 3       472     0       20
5       2,430   2       26      5'9"    m1002775        emi
erocket 5       1,755   0       23      5'9"    f1002879
6       5'6"    f1003058        alschemehorn    2       324
one     f1003251        mckaela_27      2       154     0
ina     5       1,953   0       30      5'4"    f1003439
f100357 kimcicle        5       2,196   15      30      5'8
100371  marchlight      3       417     0       None    5'7
an      11      11,851  0       25      5'4"    m1003858
3,597   5       31      5'5"    f1003950        alliehumm
```

While analyzing our data, we determined that each user, on average, participated in between one and two sports. The complete dataset contains 441,034 users, but we limited the scope of our analysis to those who had recorded at least one exercise and had completely filled out all the relevant information that we required.

4.2 Features

In order to perform our analysis, our feature extraction applied indicator features to age, gender, and height buckets:

- Age: under 25, between 25 and 35, over 35
- Gender: male or female
- Height (in inches): under 60, between 60 and 66, between 66 and 72, over 72

Additionally, each exercise in the dataset mapped to an indicator feature measuring the specific workout recorded. Exercise success was ultimately not stored as a feature; because of this, we were able to regress solely on the workout regimen and not on the skill level.

5 Methods

5.1 Predicting the Sport

After considering a number of possibilities, we ultimately settled on a slight variant of the softmax algorithm for our prediction. Because some users played a number of sports, we considered training output $y^{(i)}$ as a set of elements, and used the slightly modified stochastic gradient ascent update rule

$$\theta_j := \theta_j + \alpha \left(\mathbf{1} \left[j \in y^{(i)} \right] - h_{\theta_j} \left(x^{(i)} \right) \right) x^{(i)}$$

for training example $\left(x^{(i)}, y^{(i)} \right)$, where θ_j is the weight vector for sport j and

$$h_\theta(x) = g\left(\theta^T x\right) = \frac{1}{1 + e^{-\theta^T x}}.$$

We decided to iterate through the weights 5 times and we used

$$\alpha = \frac{1}{\sqrt{n}}$$

on the n^{th} training example, as these two decisions displayed a strong pattern of convergence.

Once training was completed, we assigned

$$y^{(i)} = \arg\max_j h_{\theta_j} \left(x^{(i)} \right).$$

After the training, note that prediction follows the classic softmax model.

5.2 Predicting Proficiency

In order to create the separate Gaussians, we used

$$f_k\left(x^{(i)}\right) = w_k^{(i)}\left(1 + \frac{r_k^{(i)}}{30}\right)[1]$$

to calculate the approximate 1-rep-maximum for user i on exercise k, with $w_k^{(i)}$ representing their most recent weight for exercise k and $r_k^{(i)}$ representing the number of reps at that weight. Using this formula, we calculate

$$\mu_k = \frac{\sum_{i=1}^{n} f_k\left(x^{(i)}\right)\left[\phi\left(x^{(i)}\right)_k = 1\right]}{\sum_{i=1}^{n} \mathbf{1}[\phi\left(x^{(i)}\right)_k = 1]}$$

$$\sigma_k^2 = \frac{\sum_{i=1}^{n}\left(f_k\left(x^{(i)}\right) - \mu_k\right)^2\left[\phi\left(x^{(i)}\right)_k = 1\right]}{\sum_{i=1}^{n} \mathbf{1}[\phi\left(x^{(i)}\right)_k = 1] - 1}$$

estimating the parameters for normal distributions based on the sample means and variances. If we define m as the number of exercises, creating the $m \times 1$ vector $p^{(i)}$ with

$$p_k = \int_{-\infty}^{f_k\left(x^{(i)}\right)} \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}\, dx$$

we calculate total percentile for user i with predicted sport j as

$$\text{Percentile}^{(i)} = \frac{\theta_j^T p^{(i)}}{||p^{(i)}||_2}.$$

6 Results and Analysis

6.1 Results

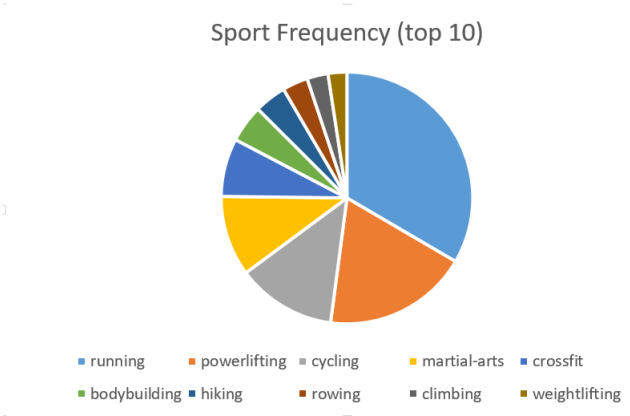
Because the proficiency estimate is not testable with the data available, we discuss only results of the sport prediction algorithm. As the datset was very large and the training took a long time, we ultimately elected to use 26,274 training examples and 4,722 testing examples. Using the softmax model discussed above, with 5 epochs and annealing as described, we classified a trial as a success if the predicted sport was one of the sports played by the user. Under this definition of success, we obtained the following results:

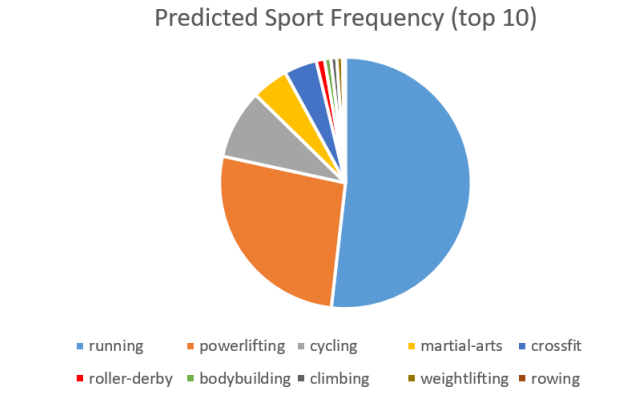
Training Error	Testing Error
37.26%	43.69%

6.2 Analysis

Although at first discouraging, these results proved to be more successful than they might initially have appeared. As per our dataset, classification takes place over 105 possible labels with an average of fewer than two correct labels.

Furthermore, the following distributions show promising similarities:





Each color represents the same sport in each of the above charts. Our predictor certainly over-predicted the more common sports, but not by so much as to abolish credibility.

7 Conclusions and Future

7.1 Conclusions

Given demographic details and exercise history, our prediction algorithm was able to fairly accurately assign a suggested sport to any user hoping to learn what they might be suited for. It would appear that our modified softmax algorithm trains the weights fairly well in order to both predict a user's sport and proficiency therein. As a subproblem, the algorithm is also adept at determining how important certain exercises are for training in specific sports.

7.2 Future Work

Given more time, there are many more interesting tests to run. One idea that we toyed with, were not able to implement to a fully functioning finale, is clustering sports together and training/predicting on clusters instead of specific sports. This would have the have the effect of steering a user towards a certain group of sports, from which they would be better able to choose than they would from the total set of sports.

Along those lines, predicting the k best sports for a user would have a similar effect, and would reduce the effect of incorrect predictions.

Regularization in the original softmax loss function might reduce the overwhelming frequency of the more common sports (like running and powerlifting), allowing the pie charts above to look a bit more similar to each other.

Finally, instead of the current weighted average algorithm currently in place to predict proficiency within a given sport, an EM algorithm to predict proficiency might improve proficiency predictions even when sport predictions are inaccurate (which they are a significant enough amount of the time).

Overall, there is certainly a lot left to study in this area. Nevertheless, our results are promising, and we are hopeful that this algorithm might be put to good use some day.

References

[1] Epley, B. Poundage chart. *Boyd Epley Workout.* Lincoln, NE: Body Enterprises, 1985. p. 86.

[2] Jurgens, David; McCorriston, James; Rhuts, Derek. 2015. *An Analysis of Exercising Behavior in Online Populations.* McGill University Department of Computer Science. In *Proceedings of ICWSM.*

[3] United States Dept. of Labor. Bureau of Labor Statistics. *Sports and Exercise.* Web. 14 Dec. 2016. https://www.bls.gov/spotlight/2008/sports/.