Least Square Analysis on Exercise Duration of Recorded Data from the Endomondo Fitness App

Fang Qiao
Computer Science and Engineering Department
University of California, San Diego
faqiao@ucsd.edu

Abstract

The Endomondo is social network focused on user fitness. The technology is featured as a mobile app that collects data for physical exercises. Each recorded exercise of a user, or "workout" instance, consists of data for various spots and with a number of exercise-related statists including but not limited to speed, duration, weather, etc. In this study, I focus on analyzing the data for the purpose of identifying experimenting on one aspect of the data indicative of fitness. We then extrapolate the relation of the rest of the data to that fitness aspect, and attempt to build a model that best represents the relationship.

1 The Endomondo Data Set

For this study, we focus our efforts on the workout data itself rather than user data.

A web scrapping of the site in the mid-2014 has turned up 5.6 million workouts from 1.5 million users[1].

Each workouts contains a subset of 36 properties (Appendix A), and consists of one of 55 different kinds of sports (Appendix B).

Experiment Data Set Construction

We construct our experimental data set with the goal of:

- Reducing data size to make our tasks more tractable
- Avoid the missing data problem
    - Our data set is very large, we can cut out large chunks and still have a representative sample.
    - Some instances are "tests" by the user, and are therefore not real workouts

Workout instances are discarded if it does not have the set of essential features: Avg. Speed, Max. Altitude, Total Descent, Distance, Max. Speed, Hydration, date-time, Duration, Total Ascent, Min. Altitude, Calories, and Weather.

Several features were eliminated to make analysis tractable:

- Large lists – these are data points describing a particular time during a workout: speed, duration, lat, lng, alt, cad, hr, distance
- Features with low availability across all data: Fitness Level, HR After Test, Temperature, etc.

[1] http://jmcauley.ucsd.edu/data/endomondo1.txt.gz

I also removed any user and its respective workouts if the user have less than 4 instances of workouts recorded.

Finally, because sports type appears to be very much correlated with our prediction task (see section 3), we further crop our data set to include only 7 types of most common workouts:

- Running
- Aerobics
- Mountain biking
- Walking
- Cycling (transport)
- Hiking
- Cycling (sport)

Users who have recorded more than one type of workout are removed.

With the above operations, the dataset is reduced to 106,198 users and 786,621 workout instances.

2 Prediction Task

To start, 5 features were identified as potential labels for prediction:

1. hr/ HR after test
2. fitness score/fitness level
3. avg. speed/max. speed
4. calories
5. duration

Due to the lack of availability, heart rate and fitness were ruled out. Calories count was seriously consider until it was confirmed that Endomondo manually computes the number using an imprecise formula derived from *"The Compendium of Physical Activity"* [1] and would therefore have direct relation to the workout properties. Furthermore, calories is said to be computed from both the stats of the individual user (body weight, age, etc.) as well as the per workout instance numbers, but for this study we are only focusing on workout data.

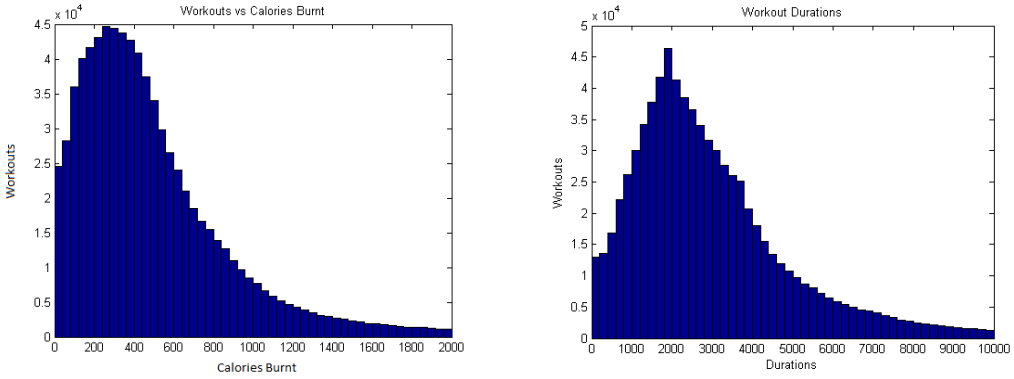

Figure 1. Distribution of workouts over Calories (kcal) and Durations (seconds).

Finally, speed was ruled out in favor of duration by the intuition that people directly micro manage their intensity during exercise, but directly/indirectly macro manages duration either due to fatigue or pre-planning. Analysis shows that the workout instances have a similar distribution over duration as it does over calories (Figure 1), and that suggest perhaps it has similar implications on the effort of the user as indicated per workout.

As described in the previous sections, the task is a supervised learning problem to use a set containing both discrete and continuous features to predict a discrete label (duration). Two obvious approaches comes to mind:

1. Regression directly on the features and the label

2. Cluster the labels, and train classifiers that maps the features to a certain range of duration.

The two approaches [2] I examine are

- Basic linear regression:

$$\forall x \in Training\ Data\ X, \forall y \in Training\ Label\ Y, N = Training\ size$$

$$\min_{\theta} \arg \sum_{i}^{N} (y_i - \theta x_i)^2 \Rightarrow \theta = (X^T X)^{-1} X^T Y$$

- Linear regression with L2 regularization with gradient descent.

$$\min_{\theta} \arg \frac{1}{N} \sum_{i}^{N} (y_i - \theta x_i)^2 + \lambda \|\theta\|_2^2$$

where

$$f(\theta) = \frac{1}{N} \sum_{i}^{N} (y_i - \theta x_i)^2 + \lambda \|\theta\|_2^2$$

and

$$\frac{\partial f}{\partial \theta_k}(\theta) = -\frac{2}{N} \sum_{i}^{N} (y_i - \theta x_i) x_{ik} + 2\lambda \theta_k$$

3 Background Literature

For the task of predicting performance based on user and item data, a recommender systems based on users vs. workout sport type and intensity would be most fitting. Recommendation systems such as Microsoft Matchbox [3] and Etsy [4] are capable of dealing with very large scale data any describes additional methodologies for dealing with sparse data, the cold start problem, and topical modeling, etc.

Moreover, since the intuition is that people who exercises have different fitness level and aptitude with sports, the ideal is to temporal pattern should considered by deriving dynamic expertise levels from users based on their sequence of workouts in addition to user, workout, and user-workout latent features. State of the art techniques for such tasks are described in the work of Julian McAuley [5] [6].

In this study, however, due to time limitations and because Endomondo data set is a new set that has no known previous studies, we focus on a few basic techniques as did Kowalski et al. for the Chinese Pinyin education data [7]. The goal is to derive some preliminary patterns.

4 Analyzing Features

We start with the set of available properties picked during experimentation data set construction (Section 1): Avg. Speed, Max. Altitude, Total Descent, Distance, Max. Speed, Hydration, date-time, Duration, Total Ascent, Min. Altitude, Calories, and Weather.

Hydration is ruled out first because Endomondo disclosed that it is computed based on the workout intensity, duration, temperature, and user weight[2].

The features are then evaluated against duration to test for correlation.

4.1 Speed vs. Duration

All speeds are converted to mph, if any were minute/mile previously. Workouts with infinite and 0 speeds are discarded.

[2] https://support.endomondo.com/hc/en-us/articles/201869007-Hydration-

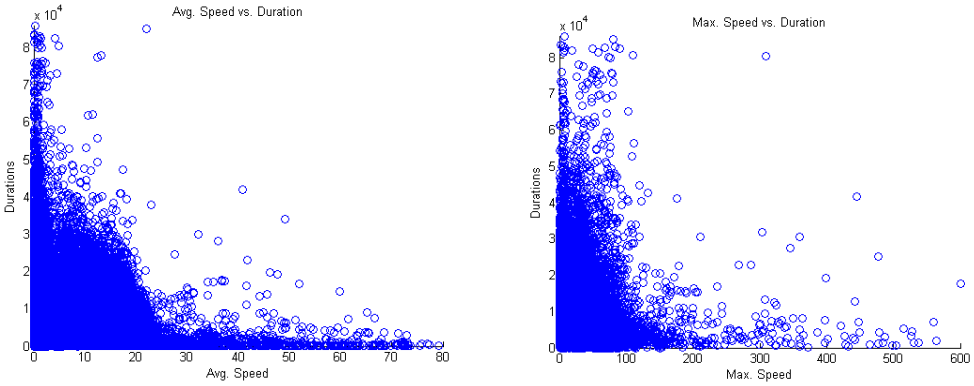

Figure 2 - Speed vs. Duration.

As expected, we see that duration generally decreases as the speed increases.

4.2 Ascension/Descend/Altitude vs. Duration

We take the difference of ascension – descend in feet.

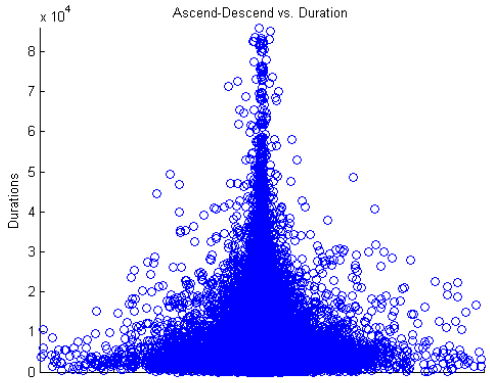

Figure 3. Ascension - descend in ft vs. Duration

Duration is highest when users are exercising mostly on flat ground, and drops off when there's greater ascension or descend.

This may be due to the fact that most of the selected sports (running, aerobics walking, cycling for transportation/sport) do not involve great changes in altitude.

4.2 Sport Type vs. Duration

Some correlations appear to exist between the different type of sports and duration. People can take long walks but tend to spends less time in intense aerobics. But then again, running having the highest duration is probably because there are all sorts of different runners. This would have been captured if we had taken into account user features and expertise.

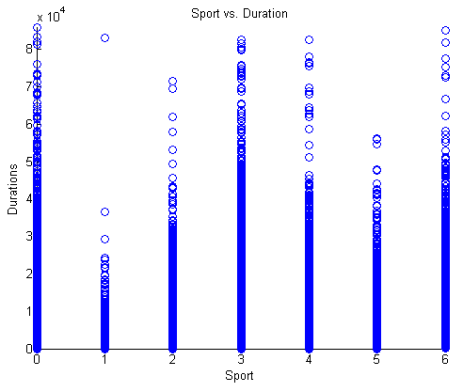

Figure 4. Sport type vs. duration - 0. Running, 1. Aerobics, 2. Mountain biking, 3. Walking, 4. Cycling (transport), 5. Hiking, 6. Cycling, sport

4.2 Weather vs. Duration

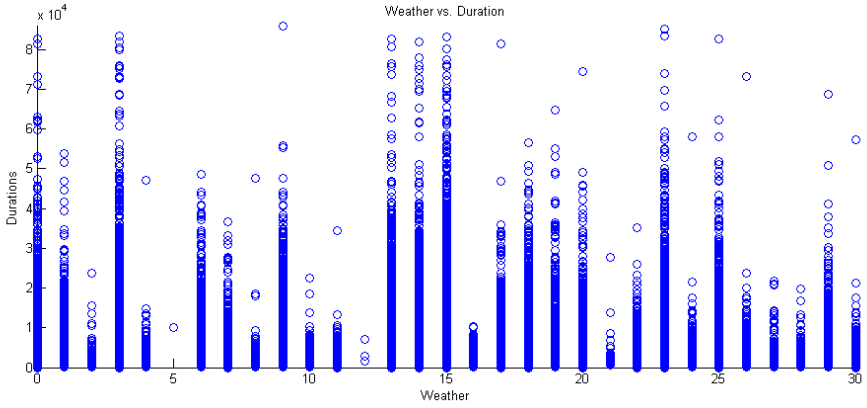

Figure 5. Weather condition vs. Duration. – See Appendix C.

People exercise more/longer if it's mostly sunny or cloudy, and much less if there's rain. Few people persists or even do anything at all through snow or dreary weather. We see this pattern because most of the sport we've selected are outdoor activities, with the exception of perhaps aerobics. Also, people may just be less willing to get out of the house when there's bad weather.

4.2 Date-time vs. Duration

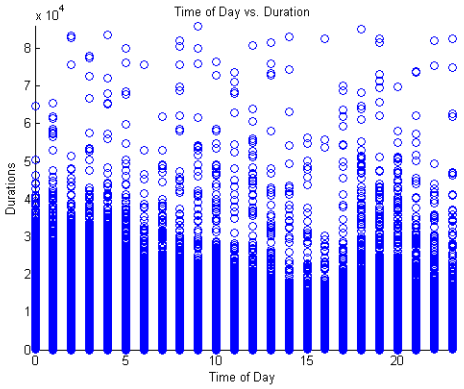

Figure 6. Time of day vs. Duration.

Surprisingly, people workout at any god forsaken hour of the day. There appears to be very subtle correlations between time of day and workout duration. People appears to exercise out least in the early afternoons, and most at night.

This may have something to do with the demographics of Endomondo users, but we have no such data at this point.

Graphing the month of year vs. duration shows that the vast majority of the exercises were recorded during April. Brief research into the timeline has not revealed the cause. Which can be due to anything from when the mobile apps where released, to the app's hosted events (a good number of them happened in April, 2014), when it became viral, etc.

Due to this unexplained irregularity, month of year is not included among the feature set for this study.

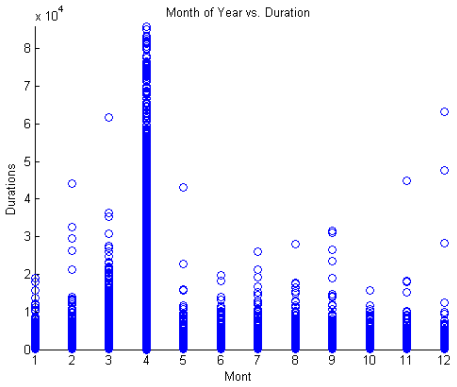

Figure 7. Month of Year vs. Duration

5 The Model

As described in the previous sections, we aim to train a linear model θ on the feature set. The initial model for basic linear regression consists of the following nine features: max. speed, avg. speed, total ascent, hour, min. altitude, max. altitude, total descent, sport type, weather. Out of these, hour, sport type and weather are discrete, the rest are continuous.

The data vector of each workout consists of the numeric values of each feature plus a constant element, 1:

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_8 x_8$$

After the preliminary results were obtained, each discrete features are further converted to an array of binary features, since we do not measure the variation of duration over these any of these features. For instance, if a workout occurred at 9 pm, then its array of 24 binary features consists of all 0s with a 1 at index 20. Since there are 24 hours, 31 weather conditions and 7 sports, this yields a total of 68 features plus a constant:

$$y = \theta_0 + \theta_1 x_1 + \cdots + \theta_{68} x_{68}$$

Finally, we have to tune a parameter, λ, for our L2-regularized model, so our final model is simply $\{\theta, \lambda\}$.

6 Experiments and Results

6.1 Baseline

We establish a baseline by running randomized 5-fold cross-validation (629297 training data and 157324 test data per iteration) with the mean of the training set label as the predicted label.

The cross-validation average sum of square deviation of training label mean as prediction is:

5.6257e+12 for the training set and 1.4064e+12 for the test set.

Averaged against the training/testing sample size, we have:

8.9396e+06 for the training set and 8.9396e+06 test set

	Training error	**Test error**
Sum-of squared deviation, cross-validation average	5.6257e+12	1.4064e+12
Sum-of squared deviation averaged by sample size	8.9396e+06	8.9396e+06

As expected, the training error and test error are alike (exactly the same in our case).

6.2 Basic Linear Regression

Basic linear regression with 8 features plus one constant feature with 5-fold validation yields some improvement over our fixed baseline:

	Training error	**Test error**
Sum-of squared deviation, cross-validation average	4.3208e+12	1.1047e+12
Sum-of squared deviation averaged by sample size	6.8660e+06	7.0220e+06

As expected, per sample test error is larger than training error.

Sample result model parameter:

constant	2645.54698395046
max. speed	0.0837270047229225
avg. speed	-0.256701852371684
total ascent	1.28860714956148
time of day	-42.4509581369517
min. altitude	0.146401708946244
max. altitude	-0.114250151278743
total descent	0.554825597877418
sport type	221.07801346099
weather	-9.2578660252033

The above sample output parameters indicates that in a linear model, there's a very high baseline constant (2645 sec ~= 45 minutes). No one feature stands out greatly given the range of the values of the features. For instance, there's 7 different sport types, 24 hours and 31 different weather conditions, and their respective coefficients are 221, -42.5 and -9.2.

Some surprises from this results include:

- Average speed is negatively correlated with duration from the baseline constant. This was not what was observed in the exploration phase.
- Total ascent increases duration whereas total descent reduces duration from the baseline constant.
- There appears to be a strong negative correlation between duration and time of day, from midnight till 11pm. No such relation was observed in exploration.

6.2 Basic Linear Regression with improved features

In the next experiment, discrete variables are expanded into vectors of binary features. Results demonstrates trivial improvement over linear regression.

Furthermore, repeated experimentation shows that the error is related to how the dataset is partitioned into training/test sets. With repeated runs, there appears to be no apparent improvement of the improved features over basic features.

	Training error	**Test error**
Sum-of squared deviation, cross-validation average	4.2042e+12	1.0726e+12
Sum-of squared deviation averaged by sample size	6.6808e+06	6.8178e+06

One issue in computing the parameter for the improved features is that the product of transpose on the data array was close to singular. To remedy that, a 69 by 69 identity matrix multiplied by a constant c was added to the product. Tuning with c cross validation yield no significant improvement/deterioration to prediction result as long as $c \geq 0.01$.

For sample parameters of training result see Appendix D.

The resulting parameter from the improved features sheds insight into the relationship of individual values of sports, weather, and time of day vs. duration:

- Mountain biking, hiking, and cycling (sport) have the longest duration.
- Running, cycling (transportation) reduces duration from the baseline constant. This was not observed during exploration, but should be due the fact that the majority of the people run less and uses bike only to get to close places. But there are more variations.
- People exercise more and longer when it is mostly sunny or cloudy.
- Freezing rain drastically reduces duration.
- We had only one data point for ice and that's not enough. Perhaps if I included ice skating the parameter would be none zero.
- People exercise less around the mid/early afternoon, and at night.

6.2 Basic Linear Regression with L2-regularization

For regression with L2-regularization, one third of the data set is first randomly selected for testing. Then, 5-fold validation is implemented on the non-test data to train θ and tune λ.

Unfortunately gradient descent converges back to the input parameter no matter what input parameters were specified and how large the step size. There might be either problem with the implemented function and gradient, or bugs in the gradient descent code.

Further investigation is required to discover what is wrong.

References

[1] L. R. Keytel, J. H. Goedecke, T. D. Noakes, H. Hiiloskorpi, L. R, M. L. van der and V. E. Lambert, "Prediction of energy expenditure from heart rate monitoring during submaximal exercise.," *J Sports Sci 23,* p. 289–297, 2005.

[2] J. McAuley, *CSE 255 - Lecture 1,* University of California, San Diego, Winter 2015.

[3] D. H. Stern, R. Herbrich and T. Graepel, "Matchbox: large scale online bayesian recommendations," in *Proceedings of the 18th international conference on World wide web*, Madrid, Spain, 2009.

[4] D. Hu, R. Hall and J. Attenberg, "Style in the long tail: discovering unique interests with latent variable models in large scale social E-commerce," in *KDD '14 Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* , New York, NY, 2014.

[5] J. Yang, J. McAuley, J. Leskovec, P. LePendu and N. Shah, "Finding Progression Stages in Time-evolving Event Sequences," in *WWW'14*, Seoul, Korea, 2014.

[6] J. McAuley and J. Leskovec, "From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise through Online Reviews," in *WWW 2013*, Rio de Janeiro, Brazil, 2013.

[7] J. W. John W. Kowalski, Y. Zhang and G. J. Gordon, "Statistical Modeling of Student Performance to Improve Chinese Dictation Skills with an Intelligent Tutor," *Journal of Educational Data Mining,* vol. 6, no. 1, 2013.

Appendix A – Property Counts over all Workouts

Note: excludes user_id, workout_id, sport, and an unknown distance property.

Steps	142776	Avg. Steps/Min	140937	Distance	5590495	pace	2159421
Wind	339024	speed	2675286	lat	4290807	alt	496547
Temperature	339024	duration	4859231	Max. Speed	4005721	Duration	5295767
distance	5110588	Max. Altitude	4389917	Max. Heart Rate	159310	Total Ascent	4276126
cad	140452	Fitness Score	8105	Fitness Level	8105	Min. Altitude	4389917
Cadence	155356	Humidity	339024	Hydration	4567032	Calories	5346504
Avg. Speed	4420628	Avg. Heart Rate	151707	lng	4290807	Weather	4297949
hr	115557	Total Descent	4276126	date-time	5590495	HR After Test	1075

Appendix B – Sports Types

Gymnastics	Polo	Pilates
Golfing	Badminton	Snowshoeing
Snowboarding	Surfing	Walking, transport
Basketball	Elliptical training	Yoga
Step counter	Boxing	Indoor cycling
Cycling, transport	Handball	Fencing
Skiing, cross country	Rowing	Fitness walking
Circuit Training	Tennis	Athletics - Sprints
Climbing stairs	Wheelchair	Roller skiing
Windsurfing	Swimming	Skating
Baseball	Football, soccer	Kite surfing
Skiing, downhill	Brisk walking	Aerobics
Table tennis	Walking	Dancing
Cricket	Skateboarding	Volleyball, beach
Football, American	Cycling, sport	Treadmill running
Hockey	Riding	Weight training
Martial arts	Sailing	
Scuba diving	Football, rugby	

Appendix C -Weathers

0	Mostly sunny	16	Dreary
1	Mostly cloudy with showers	17	Rain
2	Partly cloudy with thunderstorm	18	Partly cloudy night
3	Partly sunny	19	Mostly clear night
4	Snow	20	Mostly cloudy night
5	Ice	21	Partly sunny with flurries
6	Intermittent clouds night	22	Hazy sunshine
7	Partly sunny with showers	23	Cloudy
8	Mostly cloudy with flurries	24	Mostly cloudy with thunderstorm
9	Fog	25	Clear night
10	Hazy night	26	Thunderstorms
11	Flurries	27	Partly cloudy with showers
12	Freezing rain	28	Rain and snow mixed
13	Mostly cloudy	29	Showers
14	Intermittent clouds	30	Partly sunny with thunderstorm
15	Sunny		

Appendix D – parameters of basic linear regression with improved features

constant	2295.81	Sunny	250.69	hour 7	275.47
Running	-542.08	Dreary	-170.94	hour 8	192.92
Aerobics	-169.46	Rain	-41.62	hour 9	148.45
Mountain biking	1286.55	Partly cloudy night	187.99	hour 10	74.50
Walking	200.07	Mostly clear night	185.53	hour 11	9.57
Cycling, transport	-421.23	Mostly cloudy night	122.44	hour 12	-99.08
Hiking	945.29	Partly sunny with flurries	659.93	hour 13	-201.27
Cycling, sport	996.67	Hazy sunshine	-139.70	hour 14	-313.46
Mostly sunny	343.77	Cloudy	130.33	hour 15	-345.43
Mostly cloudy with showers	80.67	Mostly cloudy with thunderstorm	-26.47	hour 16	-303.81
Partly cloudy with thunderstorm	5.64	Clear night	61.42	hour 17	-207.91
Partly sunny	272.05	Thunderstorms	-35.69	hour 18	-39.37
Snow	-89.01	Partly cloudy with showers	13.71	hour 19	76.15
Ice	0.00	Rain and snow mixed	-222.87	hour 20	14.49
Intermittent clouds night	353.35	Showers	6.60	hour 21	-80.14
Partly sunny with showers	368.66	Partly sunny with thunderstorm	-205.69	hour 22	-78.40
Mostly cloudy with flurries	379.65	hour 0	53.39	hour 23	-20.79
Fog	142.40	hour 1	197.78	max. speed	0.07
Hazy night	-42.12	hour 2	482.71	avg. speed	-0.13
Flurries	-212.11	hour 3	756.92	total ascent	1.19
Freezing rain	-602.61	hour 4	765.39	min. altitude	0.13
Mostly cloudy	196.08	hour 5	562.36	max. altitude	-0.12
Intermittent clouds	323.71	hour 6	375.37	total descent	0.59