Predicting Cycling Times

Anna Revinskaya
Stanford University

Goals

- Estimate how long it takes to bike a road segment.
- Analyze which factors influence cycling speed the most.

Data

Cycling data sources:
- Page Mill data: pagemilling.com
- Baylands trail data: ridewithgps.com

Weather data source:
- worldweatheronline.com
- 3 hour granularity

Features:
- Heart rate
- Attempt count in last 2 months
- Air temperature
- Humidity
- North wind (cos(wind angle) * wind speed)
- East wind (sin(wind angle) * wind speed)

Normalize both X and y: $X_{norm} = \frac{X - \mu_X}{\sigma_X}$

20% test, 80% train/CV data split

Challenges

- Small amount of data (30-40 data points for some cyclists).
- Unobserved factors influencing performance: fixing flat tire, different bike, other rides etc..
- Noisy measurements: unreliable heart rate sensors, weather differences between measurement station and segment location.

Least Squares Ridge regression

- Fitting separate models for each person and average out prediction errors.
- Using normal equations. $\theta = (X^T X + \lambda I)^{-1} X^T y$
- LOOCV for feature set and regularization parameter tuning.

LSTM

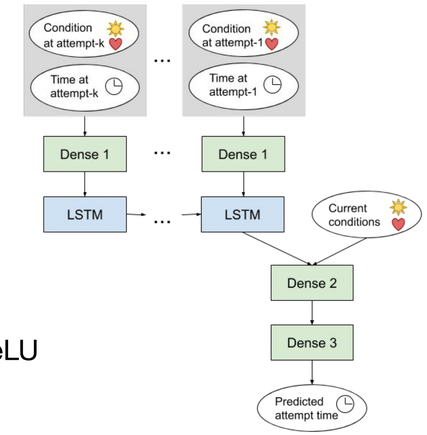
- Fit separate models for each person and average out prediction errors.
- 4-fold CV for hyperparameter tuning

Architecture details:
- Huber loss to reduce sensitivity to outliers.

$$L = \frac{1}{|D|} \sum_{x,y \in D} \begin{cases} \frac{1}{2} r^2, & \text{if } |r| < \delta \\ \delta |r| - \frac{1}{2}\delta^2, & \text{otherwise} \end{cases}$$

where $r = \theta^T x - y$

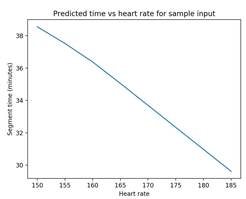
We use $\delta = 1$
- Dense layer activation: LeakyReLU
- LSTM layer activation: ReLU

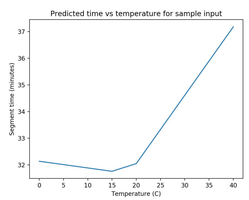


Generalized LSTM

Single model trained for all cyclists in the dataset. The same architecture as above.

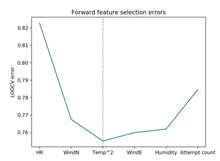
Analysis:
Predicted times plotted against HR and temperature while keeping other variables fixed (based on Page Mill segment):





Feature selection

We ran forward feature selection for Ridge regression model.

Page Mill (Moody to DF)	**Baylands Park West**
• Runs North to South • Uphill  *Image source: strava.com*	• Runs East to West • Flat  *Image source: strava.com*
Selected features: • HR • Attempt count • Temperature	Selected features: • HR • Wind north • Temperature
	

Results

RRMSE error on test dataset for Least Squares Ridge Regression and LSTM models.

Page Mill (Moody to DF)

	LS	LSTM
Test	0.0824	0.0816
Train	0.0826	0.0802

Baylands Park West

	LS	LSTM
Test	0.0680	0.0671
Train	0.0529	0.0535

RRMSE errors for generalized LSTM model:

	Page Mill (Moody to DF)	Baylands Park West
Test	0.0916	0.0636
Train	0.0895	0.0512

Conclusions: Neural network does not significantly outperform linear regression but enables a more general model that can make predictions for cyclists without data in the training set. More data would help get more conclusive results.

Predicing Cycling Times

Anna Revinskaya
Stanford University

Dataset Sizes:
Page Mill per-cyclist attempt counts:
33, 173, 96, 35, 58, 83, 93, 53, 37, 68, 369, 31, 70, 81, 91, 36, 79, 59, 81, 148, 140, 35, 35, 114

Baylands West per-cyclist attempt counts:
77, 44, 114, 56, 33, 30

Page Mill total attempt count used for Generalized LSTM model:
2272

Baylands West total attempt count used for Generalized LSTM model:
466

Video link:
https://photos.app.goo.gl/eDWMA6BqErAKKtew9