Can you *on-sight* it?

Victor Zhang (zhangvwk)

Department of Mechanical Engineering, Stanford University

Objectives

Definition

***On-sight**: completing a climbing route in one go with no prior information.*

- Predict whether one can *on-sight* a route or not, given a set of features of the climber and the route.
- Test different classifiers and measure the performance according to F_1 score.

Predicting

Why care about *on-sighting*?

1. proof of mastery;
2. rewarded in competitive rock climbing;
3. absence of formal statistical studies on the subject.

Input/Output

1. Input: vector of 8 features taken from preprocessed raw data.
2. Output: prediction of label `is_onsight` $\in \{0, 1\}$.

Summary of results

1. Tested 5 different classifiers (see Methods).
2. Beat the baseline not by a lot (12 % increase).

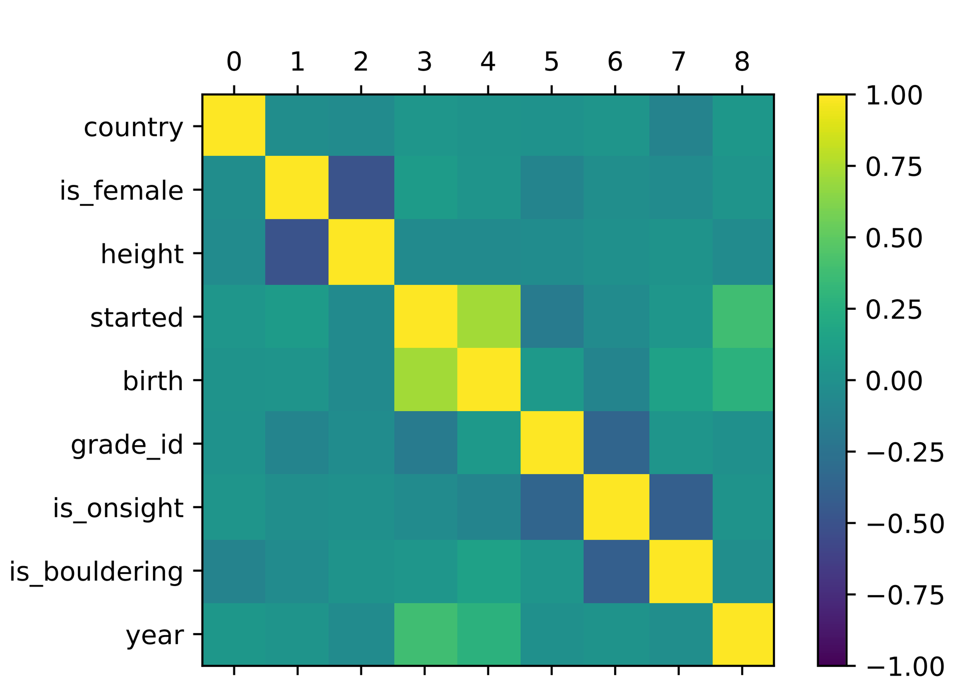

Figure 1: Correlation matrix of the features.

Data

- **8a.nu**: the world's largest climbing logbook, used to track completed rock climbs all over the planet.
- Raw data:
  1. **users** (≈ 63k rows, 22 columns)
  2. **ascents** (≈ 4.1 million rows, 28 columns)
  3. **climbing methods** (5rows, 4 columns)
  4. **climbing grades** (83 rows, 14 columns)

Set type	Set size (10^6)	`is_onsight`
Original	≈ 1.76	32.02%
Imputed	≈ 2.17	31.81%
Balanced	≈ 1.38	50.00%
Test	≈ 0.433	30.14%

Table 1: Training and test sets size and label statistics.

Metrics

- Dataset skewed towards the minority class of **non**-on-sighted climbs.
- Accuracy is not appropriate a metric.
- Precision and recall:

$$\text{precision} = \frac{tp}{tp + fp}$$

$$\text{recall} = \frac{tp}{tp + fn}$$

- F_1 score:

$$F_1 = \left(\frac{\text{recall}^{-1} + \text{precision}^{-1}}{2} \right)^{-1}$$

Features

`country, is_female, height, birth, grade_id, is_bouldering, started, year`

Methods

1. Regularized logistic regression:

$$-\sum y^{(i)} \left[\log \sigma(x^{(i)}) + (1 - y^{(i)}) \log(1 - \sigma(x^{(i)})) \right] + \lambda \|\theta\|_p^p, \qquad p \in \{1, 2\}$$

2. Regularized linear support vector classification:

$$\begin{aligned} \underset{\gamma,\theta}{\text{minimize}} \quad & \frac{1}{2}\|\theta\|_2^2 + C \sum_{i=1}^{m} \xi_i \\ \text{s.t.} \quad & y^{(i)} \left(\theta^\top x^{(i)} \right) \geq 1 - \xi_i, \quad i = 1, \ldots, m \\ & \xi_i \geq 0, \quad i = 1, \ldots, m \end{aligned}$$

3. Random forests [1] ($\alpha_b = 1/B$) and AdaBoost [2]:

$$\hat{f}(x) = \sum_{b=1}^{B} \alpha_b f_b(x)$$

4. Multi-layer perceptron: 4 layers $(5, 5, 5, 5)$.

Results

Model	Training (%)	Test (%)
Logistic regression	71.06	56.34
SVM	75.95	62.80
Random forests	75.25	61.73
Adaboost	76.23	62.82
MLP	77.15	63.11

Table 2: Comparison table of F_1 scores after training on the balanced/imputed dataset.

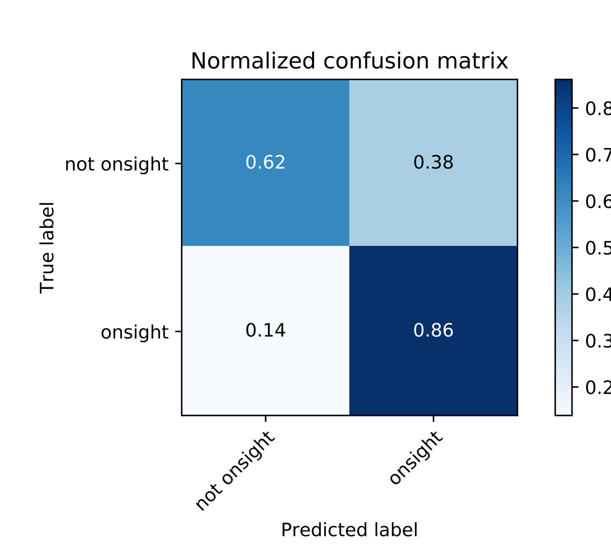

Figure 2: Confusion matrix for the MLP on the test set.

Discussion

- Close to the highest F_1 score with the baseline already.
- Challenging to find suitable non-linearities that would greatly outperform a simple linear model.
- Might need additional features to make the dataset more "learnable" (see Figure 3).

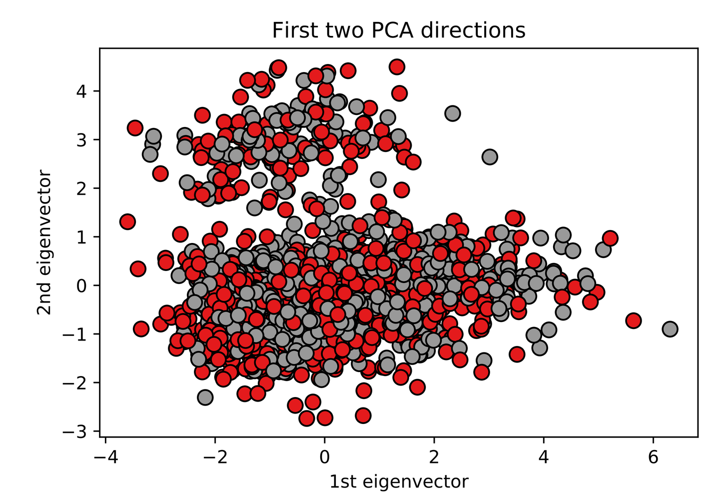

Figure 3: Subset of training set projected on the first two principal eigenvectors.

Future work

1. Try out other non-linear classifiers: kernel-induced random forests, boosting with more advanced base estimators...
2. Try out other reweighting schemes [3].
3. Include additional useful features.

Note: the majority of this dataset comes from seasoned climbers. Thus, if we truly want to generalize this to general climbing population, we would need to find a dataset that better captures the statistics of said population.

References

[1] Leo Breiman.
Random forests.
Mach. Learn., 45(1):5–32, October 2001.

[2] Robert E. Schapire and Yoav Freund.
Boosting: Foundations and Algorithms.
The MIT Press, 2012.

[3] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie.
Class-Balanced Loss Based on Effective Number of Samples.
arXiv e-prints, page arXiv:1901.05555, Jan 2019.