Senior Project Proposal

Adam Martin
Winter 2017

Project Objective

The project objective is to use a supervised machine learning algorithm to predict World Cup Nordic skiing results. More specifically the system will be trained to predict the outcome of time trial events where each athlete races individually against the clock.

Part of the task is to implement this without a machine learning software package. The calculations will be done with a linear algebra library to optimize speed.

For training data, the website www.fis-ski.com hosts World Cup skiing results going back to 1924. A portion of the project is to automate the collection and transformation of this data into a usable form.

The final output of the algorithm may be either a classification of the place (1^{st} through 30^{th}) a racer is likely to attain or his predicted percentage behind the winner.

Finally, as a prerequisite for the above objectives, this project requires basic understanding of machine learning and familiarity with the libraries used to implement the algorithms.

Technologies

The project will be implemented in Python (2.7). This choice will minimize coding effort so that the project emphasis can remain on machine learning. In addition, the numpy and mathplotlib python packages will be useful.

Learning Objectives

This will be real world practice in machine learning. Since the data is not already picked and prepared, the project will be more complex than a prescribed data set for a machine learning practice exercise and accordingly will offer greater educationally opportunity. Furthermore, the machine learning algorithms will be implemented without a machine learning library. Using the linear algebra python library numpy will facilitate better understanding of the machine learning material.

Proposed Elements

Prerequisites

Take Andrew Ng's Coursera course for a basic understanding of machine learning (5 pts)

Learn basic python (1 pts)

Section total (6 pts)

Data collection

Use regex to efficiently record ski results from www.fis-ski.com (3 pts)

Store race result data in .csv files using numpy (1 pts)

Build matrices of training data (2 pts)

Section total (6 pts)

Machine learning

Use linear regression to predict each racer's percentage behind – by hand (2 pts)

Use a neural network to predict each racer's place – by hand (3 pts)

Implement regularization to prevent overfitting of the training data (1 pts)

Programmically eliminate outliers and determine if this improves performance (1 pts)

Section total (7 pts)

Presentation

Interpret the algorithm's predictions and output the predicted results (1 pts)

Total (20 pts)

Grading Scale

A – score >= 18

B – 16 <= score < 18

C – 14 <= score < 16

D – 12 <= score < 14

F – score < 12