Do You Even Lift, Bro?
Matthew Katzman, Samuel Sowell, and Christina Ramsey
{mkatzman, alex4936, cmramsey}@stanford.edu

**PREDICTING:**
- We created a tool that will enable users to choose the best sport for them.
- The tool will also predict the likely success of a user in their optimal sport.
- We base the prediction on relevant user information, including gender, age, and workout history in selected exercises.
- The data we used were gleaned from the “Fitocracy” social media fitness application [1]. We used 105 total sports in our model, with around 1200 exercises.

**DATA AND FEATURES:**
- Data comes from three separate datasets
- Rows denote one exercise per user/group
- Columns included information on age, height, gender, each exercise performed (paired with a tuple of rep and weight values), and the sport(s) the user plays.
- In our proficiency step, our features were “calculated one-rep maxima” for each exercise. For the supervised learning, we used indicators on exercises performed, as well as on age and gender buckets.

**MODELS:**
- In order to select the optimal sport for a given user, we use softmax regression. We employ the stochastic update rule:

\[
\theta_j := \theta_j + \alpha (y_i = j - h_{\theta_j}(x_i)) x_i
\]

where \(\theta_j\) is the weight vector for sport \(j\), \((x_i, y_i)\) is the \(i^th\) training example, and \(h_{\theta_j}(x) = g(\theta_j^T x) = \frac{1}{1+e^{-\theta_j^T x}}\).
- After solving for the weights, we assign \(y_i = \arg \max_j h_{\theta_j}(x_i)\).
- Finally, within each group, we create and average Gaussians on each exercise’s 1-rep-max to predict proficiency.

**RESULTS:**
- Since users could play more than one sport, we considered a successful trial any trial in which the user played the predicted sport. With 26,274 training examples and 4,722 testing examples, 5 epochs, and step size \(n^{-1/2}\), we achieved

<table>
<thead>
<tr>
<th>TRAINING ERROR</th>
<th>TESTING ERROR</th>
</tr>
</thead>
<tbody>
<tr>
<td>37.26%</td>
<td>43.69%</td>
</tr>
</tbody>
</table>

- For the Gaussians portion, accuracy can only be tested with more access to more specific data.

**DISCUSSION:**
- This was an interesting classification project because the training examples could be classified under multiple sports, not just one. This slightly affected the softmax algorithm.
- Because of this, we did expect rather high errors to result. However, given 105 classes, these errors are quite acceptable to us.
- The Gaussian algorithm was only an approximate attempt, and did not fully utilize our machine learning tools.

**FUTURE:**
- In the future, in addition to improving the time complexity of our algorithms, we could employ EM to improve the proficiency estimation portion.

**REFERENCES:**
[1] The Fitocracy dataset was provided by David Jurgens, who collected the data for [2].