The Fundamentals of a Text-based Soccer Simulator

Sergei Chevtsov (chevtsov@stanford.edu)
December 12, 2010

Introduction

Since the 2010 FIFA World Cup, it's been our goal to develop a realistic as well as entertaining soccer simulator. Our demo [1] generates goals from a probability distribution that is based on the ELO ratings of national teams [2]. Encouragingly, the results of this simple engine don't deviate far from what actually happened in the tournament.

This paper proposes a path to a much more sophisticated simulator that can achieve a level of entertainment similar to those of real soccer matches. With several years of experience in following soccer games online, e.g. through Yahoo! live commentary [3], we decided to base this new soccer simulator on the existing textual descriptions of historic matches.

Mining Data

We wrote a Python script and downloaded 6,101 comments from all 64 games of the FIFA World Cup 2010 [3].

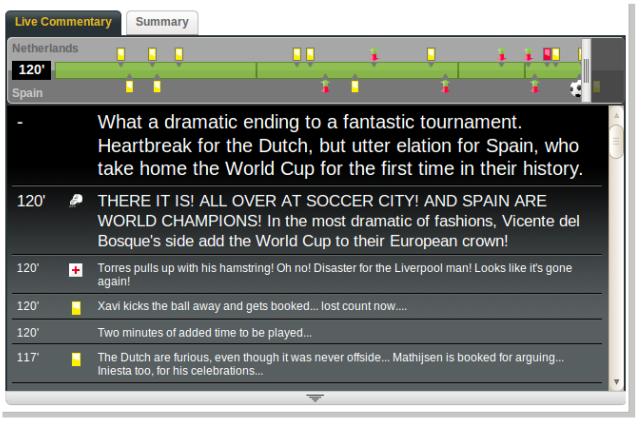

Illustration 1: Yahoo! World Cup 2010 Coverage

For the training set, we randomly chose 1,698 comments that we manually classified into 23 categories ("SoccerActions"): Injury, yellow card, added time, goal, goal kick, free kick, offside, red card, pass, kick off, substitute, game break, shot (on goal), clearance, corner kick, disallowed goal, own goal, foul, penalty goal, penalty miss, throw in, penalty awarded, and "other" (e.g. observations about the weather).

In addition, we "cleaned" tokens by removing punctuation marks from every comment; all in all, our global vocabulary consisted of 3,883 tokens.

Learning the Hypothesis

Data Representation

Our features were tokens from the vocabulary. We converted each comment into a binary vector of 3,883 elements, where the i-th element was 1, if the i-th token was present in the comment, and 0 otherwise. Our target variable was an integer between 1 and 23.

Balancing the Dataset

The number of SoccerActions in our original training set ranged from 1 (the number of "disallowed goals") to 462 (the number of "other" actions). We decided to simply copy the sets of comments from each category until each category had roughly the same number of comments and ended up with 287 to 462 comments per category.

Feature Selection

```
scores = [];
for x=x_mxn
    s = 0;
    for x_val = 0:1
        p_x = (sum(x == x_val) + 1)/(m+2);
        for y_val=y_vals
            p_y = (sum(y == y_val) + 1)/(m+k);
            indices = find(x == x_val);
            p_y_given_x = (sum(y(indices) == y_val)+1)/(size(indices,1)+k);
            p_x_and_y = p_y_given_x * p_x;
            s = s + p_x_and_y * log(p_x_and_y/(p_x*p_y));
        end
    end
    scores(end+1) = s;
end
X = X(:, [1 (find(scores > 0.01))+1]);
```

Illustration 2: Feature Selection via MI

In order to prevent overfitting and to speed up the gradient descent, we decided to calculate the mutual information (MI) between the values of every feature and the target variable. Empirically, we chose to keep only those features that had MI > 0.01. Removing features with higher MIs as well as keeping features with lower MIs resulted in higher test errors.

This process left us with just 382 tokens (among the removed tokens were the obviously useless “a” and “the”).

Learning Algorithm

Initially, we coded up 23 Naïve Bayes classifiers for each SoccerAction, but the training error stayed above 80%. Eventually, we decided to use the logistic regression for multi-class classification (aka the softmax regression). We decided to stop the gradient descent, if the training error stabilized and/or fell below 3%.

```
for counter=1:50
    old_theta = theta;
    for i=1:m
        theta(:, k) = 0;
        x = [1 x_mxn(i, :)]';
        exp_nu = exp(theta'* x);
        h = exp_nu/ sum(exp_nu);
        h(end) = 1 - sum(h(1:end-1));
        y = y_mxk(i, :)';
        theta = theta + x * alpha*(y-h)';
    end
    exp_nu = exp(theta' * [ones(m,1) x_mxn]');
    [v, i] = max(exp_nu ./ repmat(sum(exp_nu), k, 1));
    y = T(:, 1);
    accuracy = sum((i-1)' == y)/size(y, 1)
    if accuracy > 0.97 || accuracy == 1-alpha
        break;
    end
    alpha = 1-accuracy;
end
```

Illustration 3: Softmax Regression

In addition, we implemented a 4-fold cross validation on equally and randomly sliced subsets of the training data. We executed the algorithm for random training sets of 4937, 5924, 6911, 7898, 8885, and 9872 (i.e. all) elements.

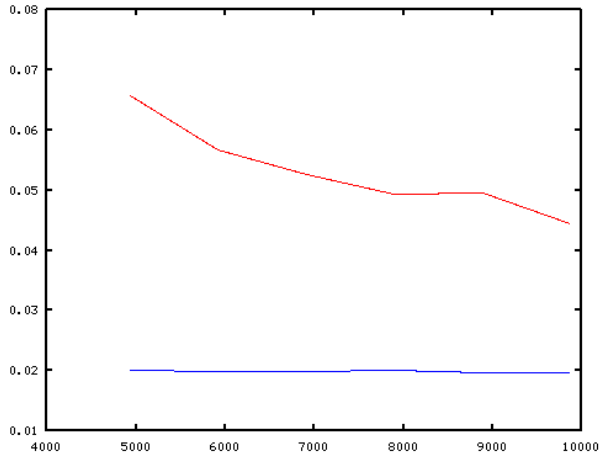

Illustration 4: Training (below) and Test Errors

Then we plotted the average training error (blue) and the average test error (red) against the size of each training set (m).
Finally, we trained a 383-by-23 matrix (“theta”) using the entire training set and obtained the following training errors per category:

Category	Training Error (in %)
Injury	0
Yellow card	0
Added time	0
Goal	0
Goal kick	0
Free kick	2.34
Offside	0
Red card	0
Pass	14.71
Kick-off	0
Substitute	1.05
Game break	0
Shot	19.87
Clearance	24.28

Corner kick	0
Disallowed goal	0
Own goal	0
Foul	0
Penalty goal	0
Penalty miss	0
Throw in	0
Penalty awarded	0
Other	9.31

The overall training error was 2.61%.

Proof of Concept

We used the obtained theta to classify each of the 6,101 comments from every World Cup game via the softmax function. Then for each game, we counted every occurred SoccerAction and put the results into a 23-element vector. Here a red flag came up, because unfortunately, we were not able to classify “goal”- the most important SoccerAction- correctly, counting much fewer goals (6) than were actually scored in the tournament (145). Nevertheless, we believed that we were at least somewhat on the right track and decided to use the Principal Component Analysis to plot each “game”-vector in 3D to possibly gain some insights.

```
%normalize
mu = mean(actions);
actions = actions - repmat(mu, m, 1);
for j=1:n
    var = sum(actions(:,j) .^2)/m;
    if var > 0
        actions(:,j) = actions(:,j) / sqrt(var);
    end
end
X=cov(actions);
[v,V] = eig(X); [v,S] = sort(-diag(V)); V = V(:,S); v = -v;
y = V(:, 1:3)' * original_actions';
```

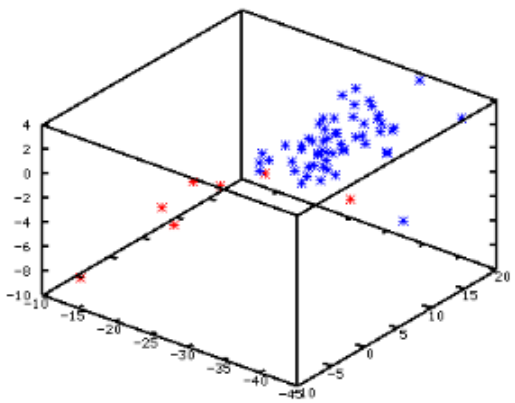

Illustration 5: Principal Component Analysis (lighter dots = Spain's games)

As seen below, PCA was able to separate most games of the eventual World Champion Spain (red points) from other games in the tournament (blue points). Indeed, 5 of Spain's 7 games ended with a score 1-0 [4], which many observers attributed to the unusually destructive strategies of their tactically inferior opponents [5].
Therefore, we have a reason to believe that our approach has some potential after all.

Discussion

So where did we go wrong?
In retrospect we probably were not correct to apply the logistic regression to classes that are subsets of each other (e.g. a goal is also a shot). A better approach would be to create a top-down decision tree of several softmax classifiers. In fact, we already identified the 5 most general classes for the initial softmax regression: foul, shot, pass, substitute, and “other”.
Moreover, we should balance the training dataset with an algorithm that is better grounded in statistical theory (e.g. bootstrapping).

Future Research

Once we have a reliable classifier for soccer comments, we will model each soccer game as a Markov decision process. SoccerActions will obviously form the set of actions. Our states will likely have three parameters: the win probability of each team as well as an indicator of who has the ball (we will derive the initial state from the ELO ratings).

Furthermore, we believe that we can estimate the probability of a team that commits a SoccerAction keeping the ball as well as the transition probabilities between the states from the hundreds of thousands of comments that were created during actual soccer games.

Finally, we would like to develop an algorithm that adapts a comment from a real game to a simulated situation in a virtual soccer match.

You are most welcome to follow our progress on scikick.com.

References

[1] “FIFA World Cup Simulator” http://www.scikick.com/

[2] “World Football Elo Ratings” http://www.eloratings.net/

[3] “Yahoo! Sport World Cup 2010” http://uk.eurosport.yahoo.com/football/world-cup/netherlands-spain-361809.html

[4] “Wikipedia 2010 FIFA World Cup” http://en.wikipedia.org/wiki/2010_FIFA_World_Cup

[5] “Wikipedia Tiki-taka” http://en.wikipedia.org/wiki/Tiki-taka

… and, of course, Prof. Ng's lecture notes.