March Machine Learning Mania 2016

Predict the 2016 NCAA Basketball Tournament

Joe Tan Chin Yong Liu Zeyan Wei Yumou Xie Dai

CE/CZ4041 Machine Learning

Introduction

- Predict winners and losers of the men's 2016 NCAA basketball tournament
- NCAA men's basketball tournament: 68 teams (64 + 4 play-in teams)
- Seeding
  - 68 teams are ranked by the selection committee from 1 to 68 and divided to 4 regions
  - The top 4 teams will be distributed among the 4 regions and receive a No.1 seed within that region
  - The next 4 ranked teams will then be distributed among the 4 regions, as the No. 2 seed team and so on...

Introduction

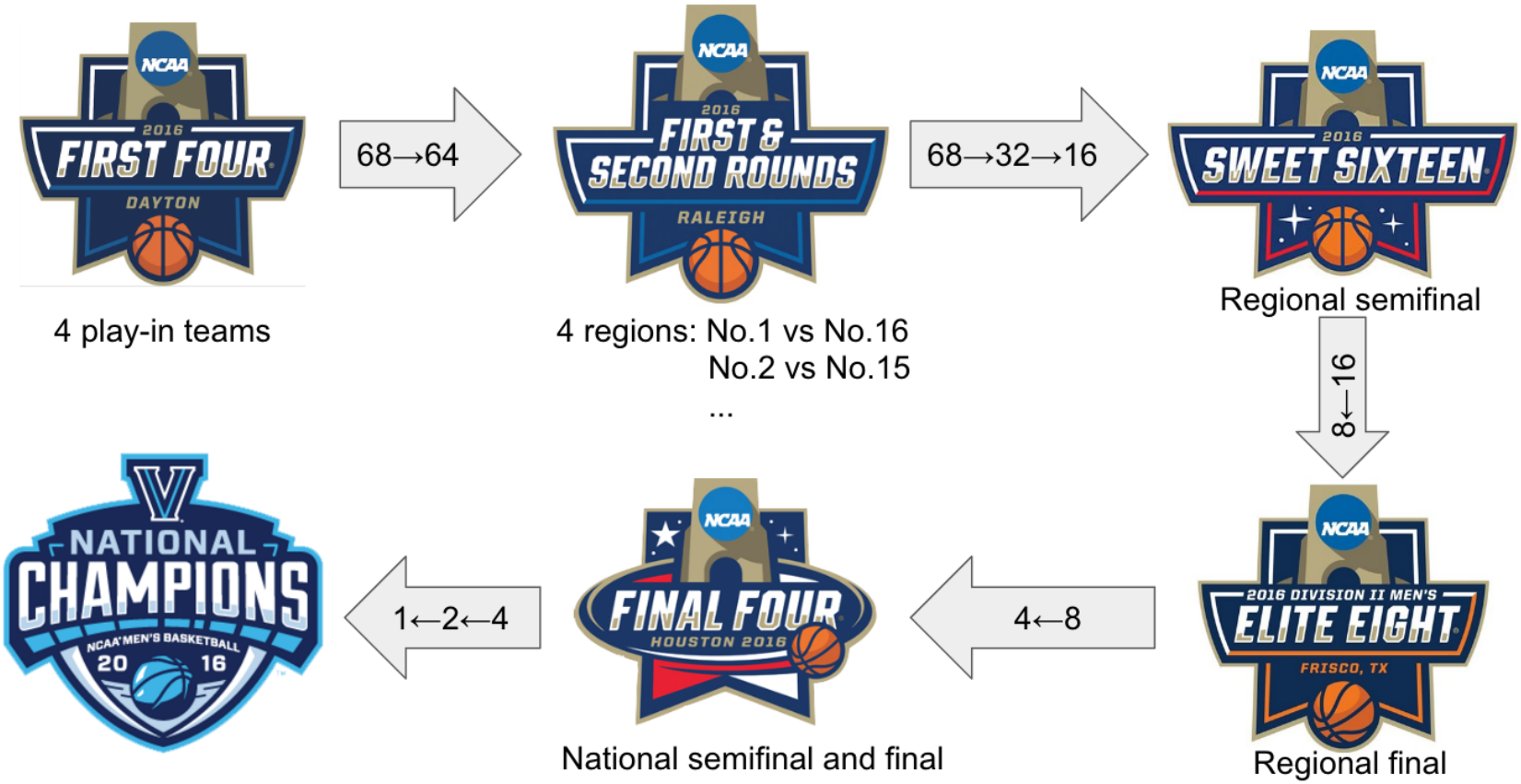


Figure 1: NCAA Basketball Tournament Schedule

Data Set

Field	Description	Example
Season	The year of the final tournament	2014
Daynum	The day the game was played on	20
Wteam	Id number of the winning team	1260
Wscore	No. of points scored by the winning team	77
Lteam	Id number of the losing team	1288
Lscore	No. of points scored by the losing team	71
Wloc	Location of the winning team: H (home), A (visiting), N (neutral)	H
Numot	No. of overtime periods in the game	0

Table 1: RegularSeasonCompactResults.csv

Data Set

Field	Description	Example
Id	season_team1id_team2id	2013_1104_1129
Pred	The predicted probability that the first team will win	0.5

Table 2: SampleSubmission.csv

Approach

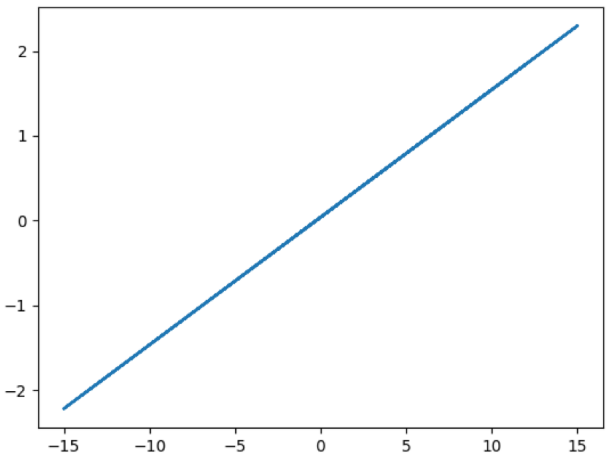
- Quantify the skill level of all teams
- Measure the difference between the skills of each team
- Predict outcomes of matches between teams
- Logistic regression (TensorFlow)
- TrueSkill™

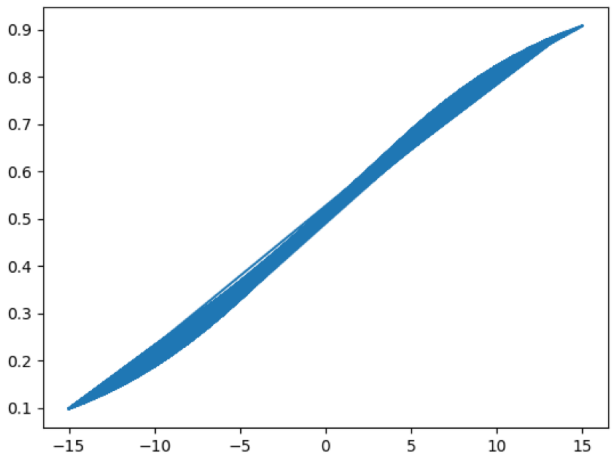
Logistic Regression

- Extract seeds for all teams
  - Seed 1 $>$ Seed 15
- Match seeds to teams
- Get difference between the seeds
  - Team 2 $-$ Team 1
- Train model with seed difference and winning team
- $y = wx + b$
  - w: 0.150500
  - b: 0.039660
- Sigmoid activation function
- Result of 0.588910

Logistic Regression

- x-axis: Difference in seed
- y-axis: Probability of Team 1 winning





TrueSkill™: Introduction

- A player-rating system developed by Microsoft Research to rank players on Xbox LIVE for matchmaking.
- A form of Bayesian Inference
- Key idea to model each player's "true skill" as a Gaussian distribution $N \sim (s_i|\mu_i, \sigma_i^2)$ (prior)
- The outcome denoted as r (evidence) $\rightarrow P(r|s_i)$ (likelihood)

$$P(s_i|r) = \frac{P(r|s_i)P(s_i)}{P(r)} \tag{1}$$

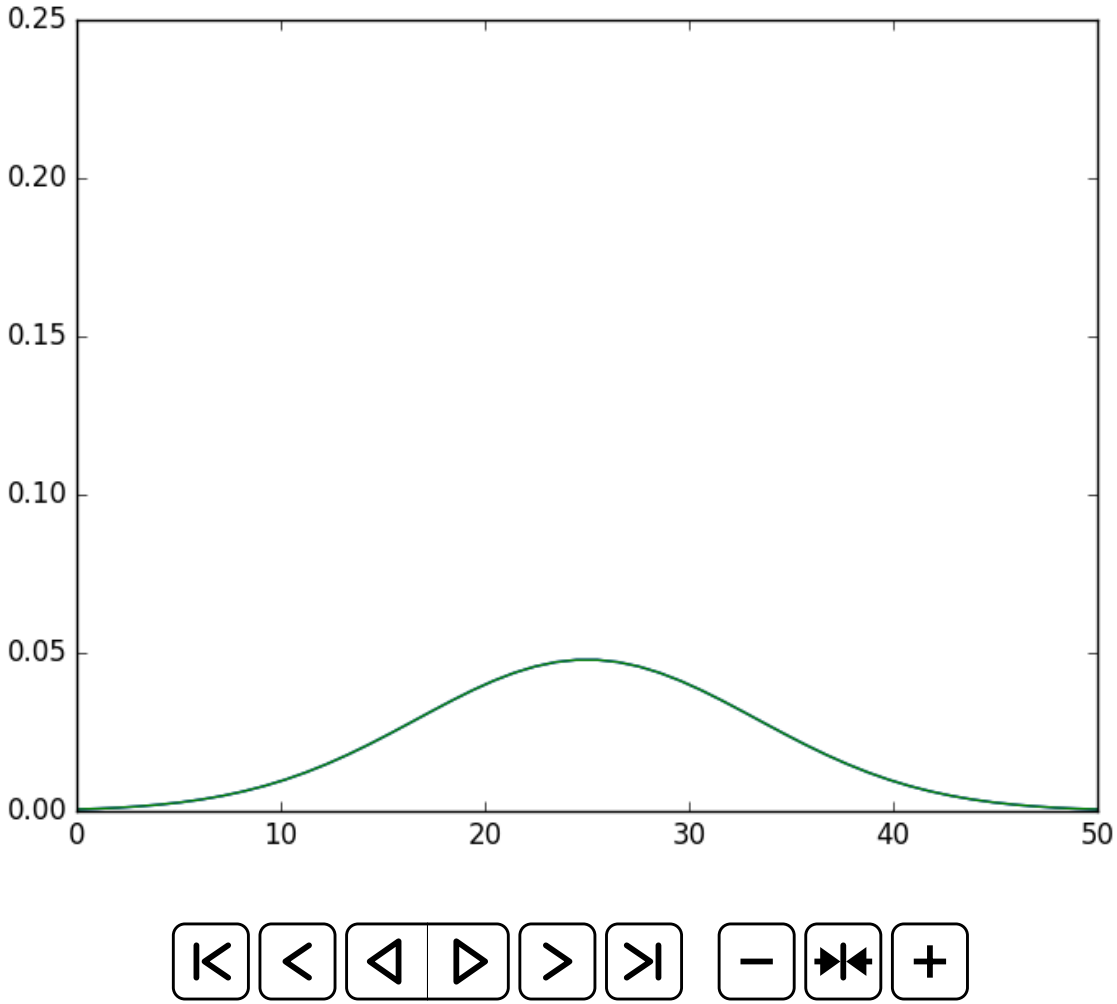
TrueSkill™: How It Works

- Start with $N_1 \sim (\mu_1 = 25, \sigma_1 = \frac{25}{3})$ and $N_2 \sim (\mu_2 = 25, \sigma_2 = \frac{25}{3})$
- As the tournament goes, TrueSkill™ updates each player's skill based on match outcomes ($Y = 1$ when Player 1 win)
- The wining probability of **an upcoming match** can be calculated based on each player's true skill distribution

$$P(Y = 1) = \Phi\left(\frac{\mu_1 - \mu_2}{\sqrt{2\beta^2 + \sigma_1^2 + \sigma_2^2}}\right) \tag{2}$$

where Φ is the standard Gaussian CDF and β represents performance variations around skill (default value $\sigma/2$)

TrueSkill™: Demo



Models

- MOV = Margin of Victory
- HCA = Home Court Advantage
- OAD = Overtime as Draw

Model	**Description**	**Parameters**
Naïve TrueSkill	Bare TrueSkill™	$\mu = 25, \sigma = 25/3$
MOV Unfolding	If MOV $> m$, recursively treat it as multiple victories	$m = 11$
HCA	Deduct h scores if victory at home court	$h = 1.5$
OAD	Treat overtime as having equal scores	-

Table 3: Models

Evaluation: Log loss

$$LogLoss = -\frac{1}{n}\sum_{i=1}^{n}[y_i \ln(\hat{y}_i) + (1 - y_i)\ln(1 - \hat{y}_i)] \tag{3}$$

where

- n is the number of games played
- $\hat{y}_i$ is the predicted probability of team 1 beating team 2
- y_i is 1 if team 1 wins, 0 if team 2 wins

Fine-tuning

- Idea: if we are very confident about a prediction, why not "all-in"?
- A tradeoff between a small reward with large probability and a huge penalty with (hopefully) small probability
- How high is the penalty? For example, we turn a 0.9 prediction into 1
  - If the outcome is 1, the score will improve by $\frac{-\ln(0.9)}{2279} = 4.623 \times 10^{-5}$
  - However if the outcome is 0, the penalty will be infinity
- Our adjustments: make all predictions > 0.9 be 1 and all predictions < 0.1 be 0

Results

- MOV = Margin of Victory
- HCA = Home Court Advantage
- OAD = Overtime as Draw

Model	**Log loss**	**Rank**	**Top**
Logistic Regression	0.588910	283^{rd}	47.08%
Naïve TrueSkill	0.527775	11^{th}	1.67%
MOV Unfolding	0.527510	11^{th}	1.67%
HCA	0.524036	6^{th}	0.83%
OAD	0.520875	6^{th}	0.83%
Fine-tuning	0.513572	5^{th}	0.67%

Table 4: Final Results

Conclusion

- TrueSkill™ algorithm can give a reasonable estimate of the "true skill" of a player (team) and calculate the wining probability of a future match
- We selected three more features to give a better prediction

Questions & Answers

Any questions?