NBA Game Predictions based on Players' Chemistry

Prastuti Singh, Bai Yang Wang

{prastuti, bwang87}@stanford.edu

Introduction

- Does player chemistry affect a team's performance? Do players show up or choke when playing against certain other players? Such questions have constantly stirred up heated conversations but little efforts have been made to settle these issues.
- In this work, we will construct a model that captures the team chemistry with a correlation and a anti-correlation matrix for every player in the league from the 2014-2015 season to present.
- The model uses these matrices and the team roster for both teams in a game to made a prediction on each game.

Related Works

Work	Best Method	Accuracy
Uudmae [1]	Neural Network	65%
Avalon, Balci and Guzman [2]	GDA	65%
Beckler, Wang and Papamichael [3]	Weighted Linear Regression and k-means Clustering	73%

[1] J. Uudmae, "Predicting NBA Game Outcomes," Retrieved from http://cs229.stanford.edu/proj2017/final-reports/5231214.pdf.
[2] G. Avalon, B. Balci, J. Guzman, "Various Machine Learning Approaches to Predicting NBA Score Margins," Retrieved from http://cs229.stanford.edu/proj2016/report/Avalon_balci_guzman_various_ml_approaches_NBA_Scores_report.pdf
[3] M. Beckler, H. Wang, and M. Papamichael, "NBA Oracle," *Zuletzt besuchtam*, 17:2008–2009, 2013.

Dataset & Features

- Data for each game from the 2014-2015 season to the present 2019-2020 [4,5], from which we pick out the names of the team and the scores
  - Total number of data points: 6501
- Team roster for each team for the same seasons [5]
  - Total number of players: 963

Team 1 Team 2 Rosters

$$x = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} \longrightarrow t_i = \begin{bmatrix} 1 \\ 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

Vector length: Number of players in the league over the past five seasons; 1 if player is on the team and 0 otherwise

$$y = \begin{cases} 1, & \text{if } s_1 - s_2 > 0 \\ 0, & \text{otherwise} \end{cases} \quad \begin{matrix} s_1 - \text{Team 1's score} \\ s_2 - \text{Team 2's score} \end{matrix}$$

[4] https://www.kaggle.com/ionaskel/nba-games-stats-from-2014-to-2018
[5] https://www.basketball-reference.com

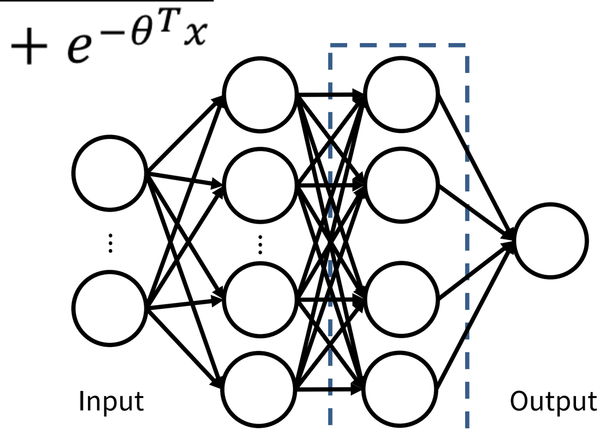
Methods

Linear Classifier

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

Neural Net

Multi-layer Perceptron (scikit)



Proposed Model: Quadratic Classifier

$$h(t_1, t_2) = g(\underbrace{t_1^T S t_1 - t_2^T S t_2}_{\text{team chemistry}} + \underbrace{2 t_1^T A t_2}_{\text{competitive performance}})$$

S - necessarily symmetric
A - necessarily anti-symmetric

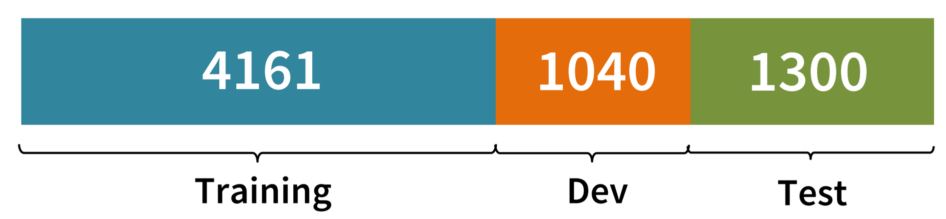
Rewrite as: $h_\theta(x) = g(x^T \theta x)$ where $x = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$, $\theta = \begin{bmatrix} S & A \\ -A & -S \end{bmatrix}$

In order to constrain S and A to be symmetric and anti-symmetric:

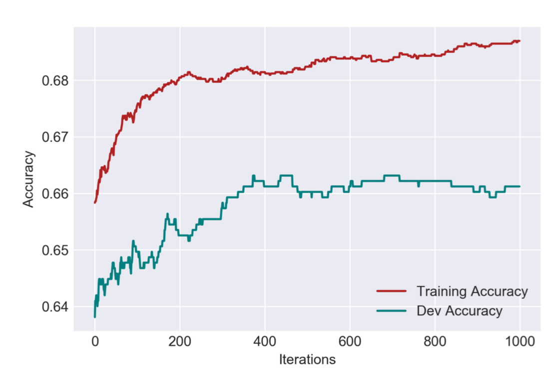
- Projected gradient descent [6] – project S and A to closest form that satisfies the constraints
- Regularization with additional cost terms $\|S - S^T\|_2$ and $\|A + A^T\|_2$

[6] N. He, "Lower bounds & Projected Gradient Descent." Retrieved from *http://niaohe.ise.illinois.edu/IE598_2016/pdf/IE598-lecture10-projected%20gradient%20descent.pdf*.

Results and Discussion



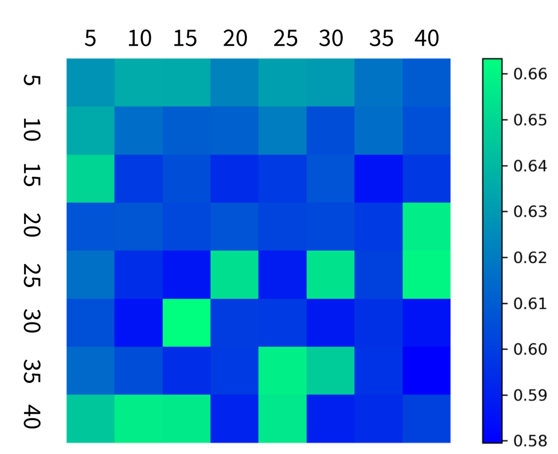
Linear Model



Dev accuracy for training on individual seasons is much worse than for entire dataset.

Test accuracy: 64.8%

Neural Net



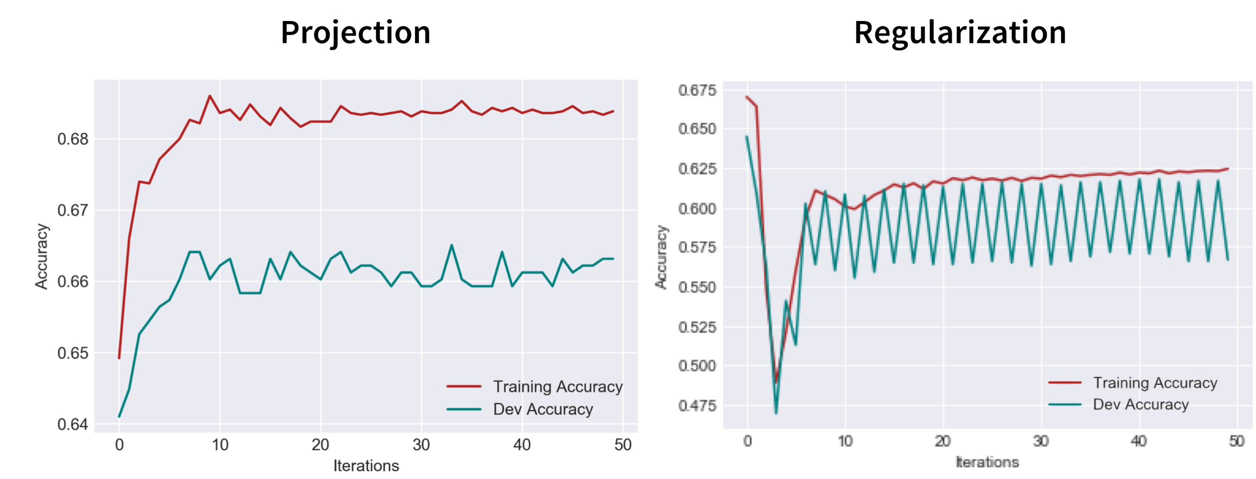
Dev Accuracy

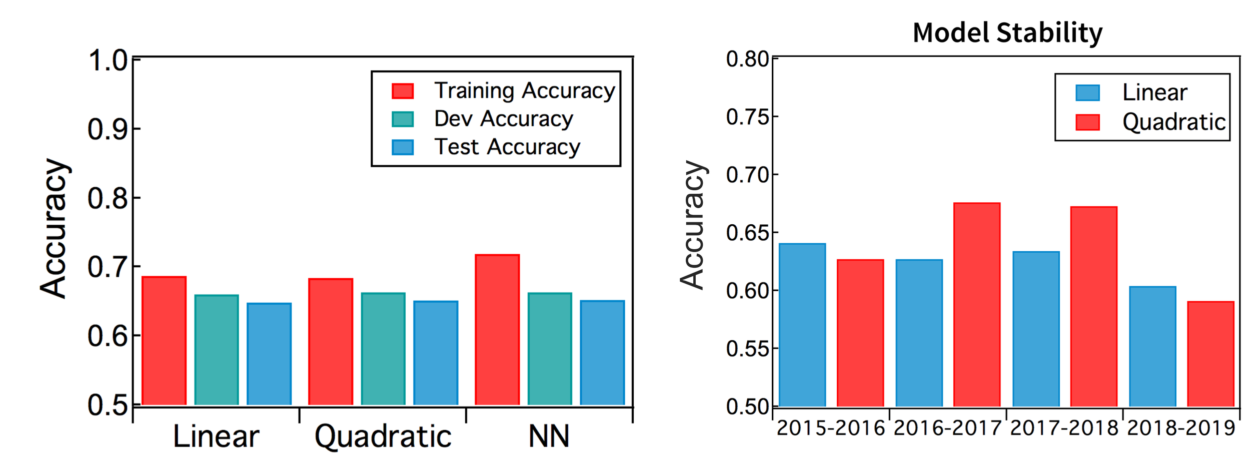
Parameters Tuned:
- Solver: [sgd, adam]
- Number of layers: 1, 2
- Number of nodes in each layer
- Hidden layer activation: [logistic, tahn]

Test accuracy: 65.2%

Quadratic Model

Projection vs Regularization



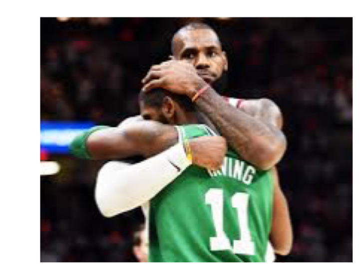


Training set: All seasons prior
Testing set: Current season

- All models perform about the same.
- Quadratic model is less likely to overfit than the other two. It seems to do better with predicting outcomes in the next season compared to the linear model (with the exception of the 2018-2019 season, which was plagued by injuries).
- Our models perform as well as most other models in the related works even though our features are minimal. There is considerable space for improvement.

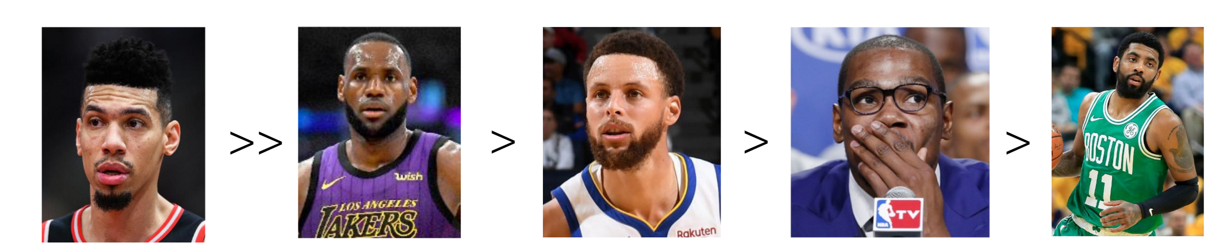
Parameter Inspection

Curry and Durant
- 'Synergy' = 0.246 (from S)
- 'Anti-synergy' [Curry over Durant] = 0.144 (from A)



LeBron and Irving
- 'Synergy' = 0.142
- 'Anti-synergy' [Lebron over Irving] = 0.0027

- LeBron and Irving are better teammates than Curry and Durant.
- When LeBron face Irving, he underperforms slightly.
- When Curry faces Durant, he steps up his game.
- Among many other observations: Paul & Griffin was a good pair and Westbrook & Durant was not.
- LeBron plays best when playing with Danny Green.



Future Work

- Instead of 0s and 1s, can we normalize the feature vectors by the number of minutes played by each player in each game. This would allow us to account for injuries and benching (player on team but not actively on court playing) and 1st string vs 2nd string. Unfortunately, this data has been difficult to obtain. We can also consider adding in player stats (e.g. 3-pt shooting percentage).
- Use k-means make predictions on new players