Mid-term Report

Spatial Analysis of Basketball Match Using Machine Learning Method

Jicheng Lu

(525004048)

1. Introduction

NBA (National Basketball Association) is the most popular men's professional basketball league in the world. With the growing influence of the league, the basketball sports has become not only an athletic event but also a comprehensive research area which involves data mining, data analysis, etc. Thus, numerous studies have been carried out to analyze the game data and further obtain a better understanding of the sports.

Given the large quantity of the games played in the past a few decades, a very rich and detailed dataset has been established and several advanced indexes have been calculated (such as the Four Factors [001]) based on the raw data. However, it is still a great challenge to evaluate the basketball games with such a large dataset, and therefore, this is where the machine learning methods should be implemented over the data.

Many of the related researches lie in the game prediction. Shi et al. [1] predicted the result of each individual NCAA games using various machine learning algorithms, such as C4.5, Ripper, Multi-Layer Perceptron, Naive Bayes and Random Forest. Puranmalka [2] established a model with different feature selection algorithms for the prediction of a single match with the involvement of both team-level and player-level data. Torres [3] compared three methods (i.e. linear regression, maximum likelihood and multi-layer perceptron) in the prediction accuracy of NBA games. Another research interest lies in the unsupervised learning of the player behaviors. Miller et al. [4] used point process modelling and dimension reduction method to evaluate the shot habits of different basketball players. Chen et al. [5] applied and proved the effectiveness of a spatio-temporal model learning method for studying the group behaviors from basketball videos. Other studies focused on the improvement of the defensive effectiveness by identifying the different features [6] and prediction of the basketball trajectory using deep learning [7].

In this project, my plan is to apply an unsupervised protocol (i.e. a combination of point process modelling and dimensionality reduction method) to better evaluate the shot selection and efficiency of players in different scenarios. The expected results include the spatial assessment of the shot selection and efficiency of one player in one or multiple seasons based on the different player types (such as the 3D wings, offensive centers, floor generals, etc.) or different time periods (such as the last two minutes of the games)

Therefore, the results can provide a clue with the coach and manager to establish different offensive and defensive strategies or make decisions to sign free agent from the player market.

2. Methodology

The proposed approach is given follows:

1) Construct the count matrix $X_{n,v}$;
2) Fit an intensity surface λ_n;
3) Construct the data matrix Λ;
4) Find optimal matrices B and W such that $W \cdot B = \Lambda$, for some reduced dimension K;
5) Plot the NMF reconstructed surface to indicate the shot selection of a player or a team;

3. Resources

The data resources are given in the following links:

1) http://nbasavant.com/shot_search.php
2) https://www.basketball-reference.com/

The knowledge needed:

1) Gaussian process, Poisson process, log-Gaussian Cox process.
2) Elliptical slice sampling.
3) Non-negative matrix factorization.
4) Optimization of the matrices in non-negative matrix factorization.

4. Progress schedule

The expected progress schedule is given as follows:

Table 1. Expected progress schedule

Date	Progress
10.1~10.14	1) Literature reading; 2) Understanding of the point process method, including Gaussian process, Poisson process, log-Gaussian Cox process. 3) Understanding of different dimension reduction methods, including non-negative matrix factorization and PCA.
10.14~10.31	1) Implementation of the proposed approach in coding and possible bug fixed.
11.1~11.14	1) Generate some of the expected results. 2) Write mid-term report.
11.15 ~ 11.30	1) Generation of the entire results, according to the objective.
12.1 ~ 12.13	1) Write final report and slides for presentation.

4.1 Current results

First of all, two different sets of shooting data are obtained at nbasavant.com, containing the shooting points (both made shots and missed shots) of two players: Stephen Curry and LeBron James. As is shown in Figure 1, both of the two plots present the shot points at the half court. The blue circles indicate the made shots, while the red crosses indicate the missed shots. It can be seen that both of the two players shoot many times at the region close to the basket. However, Stephen Curry shoots more three points than LeBron James.

Next, discretized grids are generated in the rectangular region. In each tile v, the number of the data points is counted, X_v. Thus, a count matrix can be constructed for each player. Figure 2 shows the count matrix data for the two players. It can be seen that Stephen Curry has more colorful squares at the three-point line, while LeBron James has more colorful squares at the paint area.

After constructing the count matrix for each player, we try to maximize the posterior of the log-Gaussian Cox process (LGCP). A log-Gaussian Cox process is a doubly-stochastic Poisson process with a spatially varying intensity function modeled as an exponential Gaussian process (GP):

$$\begin{aligned} &Z(\cdot) \sim GP(0,k) \\ &\lambda(\cdot) \sim \exp(Z(\cdot)) \\ &\mathrm{x}_1,\ldots,\mathrm{x}_N \sim PP(\lambda(\cdot)) \end{aligned}$$

where doubly-stochastic refers two levels of randomness: the random function $Z(\cdot)$ and the random point process with intensity $\lambda(\cdot)$.

For each player's set of data points, x_n, the likelihood of the point process is discretely approximated as follows:

$$p(x_n \mid \lambda_n(\cdot)) = \prod_{v=1}^{V} p(X_{n,v} \mid \lambda_{n,v})$$

where $\lambda_n(\cdot)$ is the exact intensity function, λ_n is the discretized intensity function. This approximation comes from the completely spatially random property of the Poisson process, allowing us to treat each tile independently. The probability of the count present in each tile is Poisson with uniform intensity ($\lambda_{n,v}$).

Explicitly representing the Gaussian random field (z_n), the posterior becomes

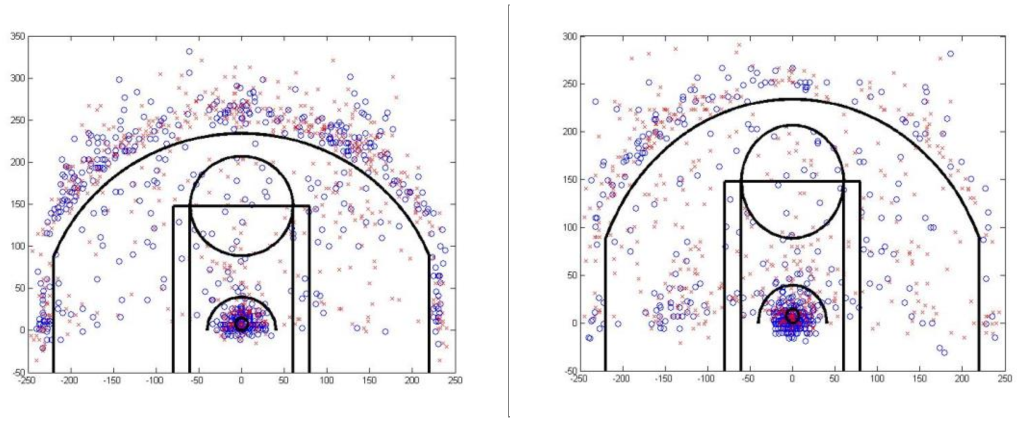
$$p\left(z_n \mid x_n\right) \propto p\left(x_n \mid \mathrm{z}_n\right) p\left(z_n\right) = \prod_{v=1}^{V} e^{-\lambda_{n,v}} \frac{\lambda_{n,v}^{\;X_{n,v}}}{X_{n,v}\,!} N\left(z_n \mid 0, K\right)$$

$$\lambda_n = \exp\left(z_n + z_0\right)$$

where z_0 is the bias term and the prior $p\left(z_n\right)$ is a mean zero normal with covariance K, determined by the squared exponential covariance function:

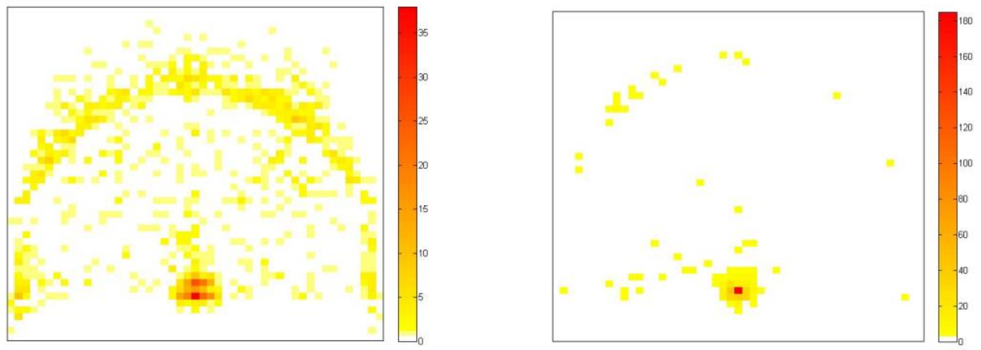
$$K_{v,u} = k\left(x_v, x_u\right) = \sigma^2 \exp\left(-\frac{\left\|x_v - x_u\right\|^2}{2\phi^2}\right)$$

where σ^2 is the marginal variation and ϕ is the length scale, which determines the smoothness of the surface. To overcome the high correlation induced by the court's spatial structure, we use the elliptical slice sampling method [8] to approximate the posterior of λ_n for each player. (The elliptical slice sampling code can be obtained from: http://homepages.inf.ed.ac.uk/imurray2/pub/10ess/#code). Figure 3 presents the LGCP surface for the two players.



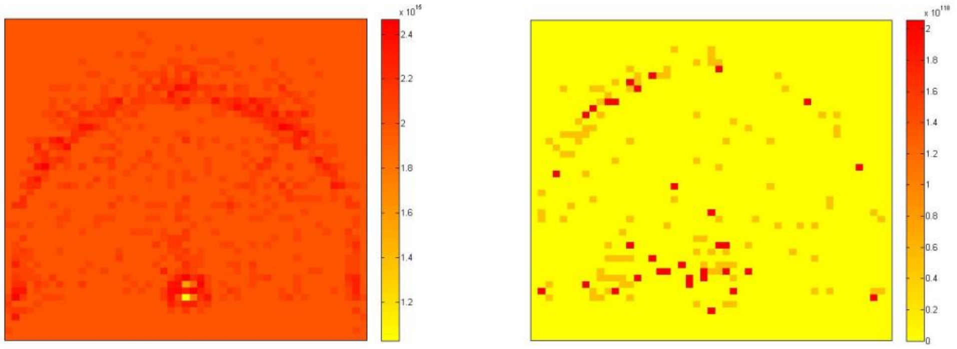
(a) Stephen Curry (b) LeBron James

Figure 1. Raw data points



(a) Stephen Curry (b) LeBron James

Figure 2. Count matrix.



(a) Stephen Curry

(b) LeBron James

Figure 3. LGCP surface.

4.2 Challenges

The objective of this project is to use an unsupervised dimensionality reduction technique (non-negative matrix factorization, NMF) to summarize the shooting habits of NBA players.

Based on the results obtained, the main challenge lies in the access of the intensity parameter in the Poisson process. Currently, we use the proposed the elliptical slice sampling method, but the results (LGCP surface in Figure 3) are not satisfactory. More efforts need to be put on the understanding of the algorithm and data property.

In the next step, the non-negative matrix factorization method will be applied to decompose the intensity matrix (Λ). The KL loss function will be used as an objective function to conduct the NMF optimization. Thus, the challenges lie in the optimization strategy selection based on the problems.

If there is still some time left, I plan to use latent variable model to convert shooting habits to efficiency, which can indicate the performance of a player more clearly.

5. References

[1] Shi, Z., Moorthy, S. and Zimmermann, A., 2013, October. Predicting NCAAB match outcomes using ML techniques–some results and lessons learned. In *ECML/PKDD 2013 Workshop on Machine Learning and Data Mining for Sports Analytics*.

[2] Puranmalka, K., 2013. *Modelling the NBA to make better predictions* (Doctoral dissertation, Massachusetts Institute of Technology).

[3] Torres, R.A., Prediction of NBA games based on Machine Learning Methods. University of Wisconsin Madison.

[4] Miller, A., Bornn, L., Adams, R. and Goldsberry, K., 2014, January. Factorized point process intensities: A spatial analysis of professional basketball. In *International Conference on Machine Learning* (pp. 235-243).

[5] Chen, C.H., Liu, T.L., Wang, Y.S., Chu, H.K., Tang, N.C. and Liao, H.Y.M., 2015, October. Spatio-Temporal Learning of Basketball Offensive Strategies. In *Proceedings of the 23rd ACM international conference on Multimedia* (pp. 1123-1126). ACM.

[6] Block, A., Machine Learning Examination of NBA Defense.

[7] Shah, R. and Romijnders, R., 2016. Applying deep learning to basketball trajectories. *arXiv preprint arXiv:1608.03793*.

[8] Murray, Iain, Adams, Ryan P., and MacKay, David J.C. Elliptical slice sampling. Journal of Machine Learning Research: Workshop and Conference Proceedings (AISTATS), 9:541–548, 2010.