EECS 498-010 Project Proposal: Machine learning on NBA data

Kai-Cheng Ku (kyleku)

1. Motivation

Sports analytics are increasingly prevalent these days. People use these methods to figure out how to play the game or build the team more efficiently. Since I'm fascinated by basketball, I'd like to use the machine-learning techniques learned in class (perhaps also some not covered in the lecture) to deal with real-world data from the NBA.

I expect to deal with two kinds of problems. The first one is the shot prediction. One of the most important abilities (at least what most audience cares about) of a player is his ability to score. I believe building a model to predict whether a shot will be made or not helps form the best strategy to make a shot.

The second one is to classify the type of players. Traditionally, people classify basketball players as center, power forward, small forward, point guard, and shooting guard by their height and weight. However, in recent days, the distinction between different positions is getting vague. For example, a big can go outside and shoot the 3-pointer instead of only fighting in the paint. I wonder if it is possible to show a new way to identify the player's types by unsupervised machine learning techniques.

2. Supporting Materials

The datasets I like to use for the first problem are the NBA 2014~2015 shot logs [1] and Kobe Bryant's Shot Selection [2]. I chose these two datasets because NBA stopped exposing its API publicly a few years ago. To get the most recent data, I have to either spend a lot of time scraping some websites or pay for it. I believe the essence of the project is to let us get more familiar with machine learning methods instead of web scraping. Moreover, methods developed in the project should be easily extended to the newer dataset. For the second problem, I can scrape the player stats from [3]. Since different from the shot log (play-by-play data), player stats are well-organized.

For the first problem, I'd like to use classic classification methods such as logistic regression and SVM. Lecture slides already explain them in detail. However, data in 2014~2015 shot logs also contain some categorical variables such as "defender ID." To fully utilize the information, I think it is a good idea to use methods that can handle categorical variables inherently, such as the random forest, which we didn't learn in class. There are many relevant online resources [4]. I would also try to handle the problem with regression methods and log loss if time permitted [5]. The output will then be the probability of making the shot or not. For the second problem, I would like to implement the methods covered in class.

3. Plan

First, real-world data is a lot messier than the datasets we tackled in the homework. Thus, I'll spend some time preprocessing the data, and let it fit each model that I would like to implement.

An outline of the project plan is listed below. Since there are only roughly 3.5 weeks for the project, I labeled some of the bullet points as "if time permitted" and perhaps finish them after the presentation...

(1) shot prediction (on 2014~2015 shot logs)

- logistic regression
- SVM with kernels
- decision tree/random forest
- boosting (if time permitted)
- regression methods, the output will be the probability of making the shot (if time permitted)
- Kobe Bryant Shot Selection dataset (if time permitted)

(2) clustering: k-means, and maybe other methods that will be taught in further lectures (if time permitted)

Reference:
[1] nba 2014~2015 shot logs
https://www.kaggle.com/datasets/dansbecker/nba-shot-logs
[2] Kobe Bryant Shot Selection
https://www.kaggle.com/competitions/kobe-bryant-shot-selection/overview
[3] https://www.basketball-reference.com/leagues/NBA_2022_totals.html
[4] https://www.freecodecamp.org/news/how-to-use-the-tree-based-algorithm-for-machine-learning/
[5] NBA Shot Prediction and Analysis by hwchase17
[6] Z Terner and A Franks. 2021. Modeling Player and Team Performance in Basketball
https://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-040720-015536
[7] Applications of Machine Learning: Basketball Strategy
https://dspace.mit.edu/bitstream/handle/1721.1/123043/1127911338-MIT.pdf?sequence=1&isAllowed=y