Andrea Li and Tyson Chen

6.867 - Machine Learning

15 November 2016

Project proposal

For our final project, we are interested in creating a system to predict outcomes and margin of victories in sports. We will start by examining basketball in the National Basketball Association (NBA), and hope to eventually generalize the algorithm to be able to examine and analyze other sports including hockey in the National Hockey League and football in the National Football League.

Plan of action:

1. **(Andrea and Tyson, Sun. 11/20)** Conduct a preliminary review of the statistics/data methods used to create current NBA spreads. This will include studying what Vegas uses as well as studying the machine learning methods that other researchers have attempted to apply in the past to predict NBA outcomes.
2. **(Tyson, Sun. 11/27)** Scrape publicly available sources such as ESPN and basketball-reference.com in order to agglomerate together a dataset that contains NBA statistics that span the past 10 years.
3. **(Andrea and Tyson, Sun. 12/4)** Construct several systems that incorporate several machine learning techniques including regression, LASSO, and neural nets in order to predict the margin of victory of an upcoming NBA game.
    a. **(Andrea, Sun. 11/27)** Regression with various basis functions
    b. **(Tyson, Sun. 12/4)** Neural network approach
    c. **(Andrea and Tyson, Sun. 12/4)** Convolutional network
4. **(Andrea, Sun. 12/11)** Test the various machine learning techniques with historical data and compare accuracies against the accuracies reported by existing algorithms currently used by Vegas, other researchers, etc. (Intermediate checkpoint 1)
5. **(Tyson, Sun. 12/11)** Run our most effective algorithm(s) on real life data for one week to see if our algorithm is able to beat the Vegas spreads. (Intermediate checkpoint 2)
6. **(Andrea and Tyson)** Generalize the optimal machine learning architecture and system to other sports including hockey and football (if time allows and running ahead of schedule)

We envision the primary risk to the project to be the gathering of sufficient amounts of properly formatted data. In order to be most effective, we will need quite large training datasets, and this data will need to be mined from various reliable sports websites. Beyond this however, the data will additionally need to be formatted consistently in a way that the machine learning algorithm can process, and putting data from various sources into a uniform format can present unique challenges.

We will try to mitigate these risks by making use of existing open-source systems which may assist in the collection and cleaning of data. We have additionally chosen to analyze basketball, which is generally a relatively high scoring game (compared with soccer for example), in order to reduce the variance that may be caused by smaller amounts of data or fewer points per game. The other important mitigating factor in choosing basketball is that there are at least three different publicly available sources that track statistics.

We plan to use the Las Vegas Sports betting system as a baseline performance metric for our algorithm. This system is the main sports betting market-maker and uses a combination of predictive models in order to establish the various betting spreads. Our system, which will make use of complex models trained through machine learning, will be tested against the Vegas system in order to see whether or not we can achieve a greater accuracy in predicting spreads (margin of victory). Furthermore, if there exists enough of a discrepancy between our spread and the Vegas spread, then we have just unlocked an opportunity to make money. For example, the Vegas spread on the game between the Cleveland Cavaliers and the Indiana Pacers on Tuesday 11/15 is -4. This means that the Cavaliers are favored to win by 4 points. If our algorithm instead calculated the spread to be -7, then we could bet in favor of the Cavaliers by more than 4 points. If the Cavaliers then win by more than 4 points, then we would win money. The one thing to note is that the payoffs are not symmetrical; in other words, in order to win $100, we would have to first put in $110.