HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 1 (2024)

Machine Learning as A Tool to Predict NBA Playoff Outcomes

Benjamin Wang

Marriotts Ridge High School, USA

ABSTRACT

Machine learning is the field of computer science that uses data to make predictions and decisions. The problem we consider in this article belongs to the class known as supervised learning and the technique we use is logistic regression. After explaining supervised learning and logistic regression, we use a data set to develop a computational model able to predict the outcome of an NBA playoff game.

Introduction

Machine learning is a branch of computer science that uses data to develop models that can be used to make predictions or make decisions. This is the subject of this article.

The data set that we consider is a collection of regular season statistics of all NBA teams and whether the team won or not against an opponent team in an NBA final game. Part of this data set is shown in Table 1. Each line corresponds to an NBA final game with statistics of the main team and the opponent team. For example, the first line tells us the total rebound per game, assist per game, and points per game for two teams in one game at the 1981 NBA final. In the game, the Celtics with 43 total rebounds per game, 26.9 assists per game, and 109.9 points per game in the regular season won the game against the Rockets with 43.5 total rebounds per game, 25.6 assists per game, and 108.3 points per game in the regular season. Similarly, the same Celtics team lost to the same Rockets team with the same corresponding regular season statistics for both teams. While only 5 games' statistics are shown in that table, our dataset contains information from 191 NBA final games. The goal of this article is to develop a model using this data so it can later be used to decide whether an NBA team can win against the opponent team.

The problem just described belongs to a class of problems known as supervised learning. The data of supervised learning problems consists of information about a collection of examples. The example in our case is the NBA final games. The information from each example consists of two types, features and labels. In our case, each example has seven features, the year, the main team's total rebounds per game, the main team's total assists per game, the main team's total points per game, the opponent team's total rebound per game, the opponent team's total assists per game, opponent team's total points per game, and one label, whether the main team won or not. In general, the label is what we want the model to predict when the features of new examples are fed into the model as input.

The goal is to develop a model that can predict the label if we are given the features of a new example. The strategy of machine learning is to start with a dataset where both features and the labels of the examples are known. In our case, this means the main team's total rebounds per game, the main team's total assists per game, the main team's total points per game, the opponent team's total rebound per game, the opponent team's total assists per game, opponent team's total points per game, and the label, whether the main team won or not, are known for all games in the dataset. This dataset is called the training set, and the examples in this set are called training examples. We will go more into detail later in this article. Machine learning techniques are then used on this dataset to develop a model. Once the model is developed, it can be used to predict the labels of

new examples from which we know only its features. In our case, once the model is developed, we can use the model to decide if the main team should win or not based the regular season statistics of the main team and the opponent team.

In the following sections, we will explain the steps that we took to build the model that helped us make predictions. The second section explains the modifications done to the datasets; The third section defines the binary classification problem. The fourth section explains logistic regression, the method to use for the binary classification problem. The fifth section defines and explains the role of binary cross-entropy error. The sixth section defines the training set and validation set. The seventh section ends the article with some conclusions.

Table 1. Dataset with regular season statistics and finals game outcome

	Win	Year	Team	Opponent Team	trb_per_game_x	ast_per_game_x	pts_per_game_x	trb_per_game_y	ast_per_game_y	pts_per_game_y
0	1	1981	Celtics	Rockets	43.6	26.9	109.9	43.5	25.6	108.3
1	0	1981	Celtics	Rockets	43.6	26.9	109.9	43.5	25.6	108.3
2	1	1981	Celtics	Rockets	43.6	26.9	109.9	43.5	25.6	108.3
3	0	1981	Celtics	Rockets	43.6	26.9	109.9	43.5	25.6	108.3
4	1	1981	Celtics	Rockets	43.6	26.9	109.9	43.5	25.6	108.3

The Dataset

The Dataset that we used to create the model is the result of combining two separate datasets. One includes the past NBA teams' regular season statistics, shown in Table 2. The other dataset includes the results of NBA Finals for each individual team, shown in Table 3. The regular season statistics dataset was simplified to only include three team statistics: total rebound per game, assist per game, and points per game. The NBA Finals results dataset is simplified to only contain whether the team won or not. The combination occurred by using Python methods to match the two datasets with their team's name and the year of the statistics.

Table 2. NBA team's regular season statistics

season	lg	team	fg_per_game	trb_per_game	ast_per_game	stl_per_game	blk_per_game
2023	NBA	Atlanta Hawks	44.6	44.4	25	7.1	4.9
2023	NBA	Boston Celtics	42.2	45.3	26.7	6.4	5.2
2023	NBA	Brooklyn Nets	41.5	40.5	25.5	7.1	6.2
2022	NBA	Toronto Raptors	40.6	45.3	22.1	9	4.6
2022	NBA	Utah Jazz	40.6	46.3	22.4	7.2	4.9
2022	NBA	Washington Wizards	40.6	43.1	25	6.4	5
2022	NBA	League Average	40.6	44.5	24.6	7.6	4.7

2021	NBA	Cleveland Cavaliers	38.6	42.8	23.8	7.8	4.5
2021	NBA	Dallas Mavericks	41.1	43.3	22.9	6.3	4.3
2021	NBA	Denver Nuggets	43.3	44.4	26.8	8.1	4.5

Table 3. NBA Finals statistics

Year	Team	Game	Win	Home	FG	TP	FT
1980	Lakers	1	1	1	48	0	13
1980	Lakers	2	0	1	48	0	8
1980	Lakers	3	1	0	44	0	23
1980	Lakers	4	0	0	44	0	14
1980	Lakers	5	1	1	41	0	26
1980	Lakers	6	1	0	45	0	33
1981	Celtics	1	1	1	41	0	16
1981	Celtics	2	0	1	41	0	8
1981	Celtics	3	1	0	40	2	12
1981	Celtics	4	0	0	35	0	16
1981	Celtics	5	1	1	41	0	27
1981	Celtics	6	1	0	43	1	15
1982	Lakers	1	1	0	49	0	26
1982	Lakers	2	0	0	35	0	24

Binary Classification Problem

In our example, the prediction or label that the model generates is whether the main team will win or not against the opponent team with the features. The label is categorical and is categorized into two categories, win or lose. This type of categorical problem with two categories as the target variable is called a binary classification problem.

In this example of the NBA Finals outcome problem, 1 represents the main team won, 0 represents the main team lost. For the model that we created based on the dataset, the function of the model is to take the features as inputs and output either 0 or 1 to represent the label.

The label which is the output of the model can be represented by $\hat{y}$. There are 7 features that are used as inputs to generate the label which can be represented as $x_1, x_2, x_3, x_4, x_5, x_6, x_7$. The seven features are the year of the game, the main team's total rebounds per game, the main team's total assists per game, the main team's total points per game, the opponent team's total rebound per game, the opponent team's total assists per game, opponent team's total points per game. For this example, the output $\hat{y}$ is a function of $x_1, x_2, x_3, x_4, x_5, x_6, x_7$, $\hat{y} = \hat{y}\,(x_1, x_2, x_3, x_4, x_5, x_6, x_7)$. If $\hat{y} < 0.5$, it indicates the label is 0 and the main team will lose. If $\hat{y} > 0.5$, it indicates the label is 1 and the main team will win.

Logistic Regression

Logistic Regression is a machine learning technique for binary classification problems. The function of logistic regression can be represented as $\hat{y}_i = \sigma(w_1x_1 + w_2x_2 + \cdots + w_kx_k + b)$. The sigmoid function σ can be defined mathematically by $\sigma = \frac{1}{1+e^{-x}}$. The graph of the sigmoid function is shown in Figure 1:

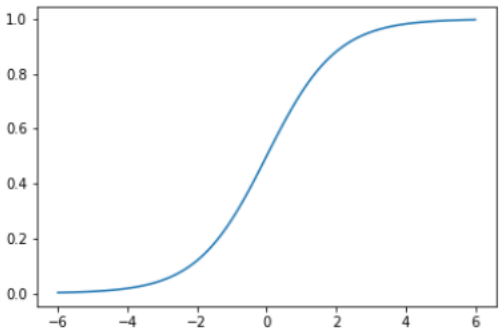

Figure 1. Plot the graph of the sigmoid function

The important properties of the sigmoid function are:

1. $0 < \sigma(x) < 1$ for all x
2. $\sigma(x)$ is an increasing function of x
3. $\sigma(x)$ becomes arbitrarily close to 0 as x becomes large in absolute value but negative
4. $\sigma(x)$ = becomes arbitrarily close to 1 as x increases
5. $\sigma(0) = 0.5$

The letter k in the function is the number of features in the problem, which in our example is 7. The $w_1, w_2, w_3, w_4, w_5, w_6, w_7, b$ are called parameters. These parameters are modified to minimize the binary cross entropy error.

With the process of choosing parameters based on the binary cross entropy error discussed in the next section, the function of logistic regression with certain parameters is sufficient to generate the predicted label $\hat{y}$.

Binary Cross Entropy Error

Binary cross entropy error is the measure of the difference between the actual label y and the label generated by the model that we created $\hat{y}$. Note y is either 0 or 1 and $0 < \hat{y} < 1$. The binary cross entropy error is defined by

$$BCE(y, \hat{y}) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}))$$

Certain features $BCE(y, \hat{y})$ of this equation are

1. $BCE(y, \hat{y}) \geq 0$
2. The closer $\hat{y}$ is to y, the smaller the binary cross entropy error is
3. If $\hat{y} = y$, then $BCE(y, \hat{y}) = 0$

Because the parameters used affect the predicted label, which makes difference in the binary cross entropy error, we can change the parameter to see the difference between the actual label and the predicted label. The parameters with the minimum mean binary cross entropy for the training set in the repetition of the previous

process will be used to create the model. The mean binary cross entropy error is the average binary cross entropy error on the examples in the set.

Training Set and Validation Set

While creating the model with the dataset, we must test the accuracy of the model. The method of splitting the dataset into training set and validation set helps to test the accuracy. In common practice, 75% of the dataset will be randomly selected to develop the model with the machine learning techniques that we mentioned above. Because we are using the training set to develop the model, the features and the labels of the training set will be used. The rest 25% of the dataset will be used to test accuracy. The model created with the training set will solely access the features of the validation set and use that to generate the predicted label $\hat{y}$. At the end, the accuracy of the model will be found by comparing the original labels of the validations set y and the predicted labels $\hat{y}$.

The accuracy on the validation set: 71%

Discussion

In this article, we provided an overview of supervised learning and logistic regression. The process of using NBA teams' statistics and machines learning techniques to predict the outcome of NBA playoff games is thoroughly explained. With the accuracy of 71%, the result showed the predictive power of the model. To improve the accuracy of the dataset, more features, like steals per game and turnovers per game, can be added to predict the outcome more accurately.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

[1] Andriy Burkov. *The hundred-page machine learning book*, volume 1. Andriy Burkov Canada, 2019.
[2] Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The elements of statistical learning,* volume 1. Springer series in statistics New York, 2001.
[3] Tom M Mitchell et al. Machine learning. 1997.
[4] Toby Segaran. *Programming collective intelligence: building smart web 2.0 applications.* " O'Reilly Media, Inc.", 2007.