Machine Learning Engineer Nanodegree

Capstone Proposal

Miguel Ángel Martínez Fernández

April 1, 2018

Proposal

Domain Background

In this project, it will be analysed a dataset containing data on NBA Players game stats from 1980 to 2017. One goal of this project is to find the set of features that best describes a player position. A model will be trained to predict player positions based on their stats for that set of features. This model could be used by NBA trainers to rethink his players' position having into consideration their last year stats. Some players, when they get older, move their playing position to more interior roles, to compensate the loss of velocity. Machine learning has been previously used to make sports predictions. In the following link it can be found a critical survey of the literature on ML for sports result prediction, focusing on the use of neural networks (NN) for this problem.

Problem Statement

The problem to be solved is to predict which position should be playing a NBA player based on his stats. It is usual that players as they get older, they move slower and it is needed to change their playing position to more interior roles. Being able to predict when that change is needed can save both the trainer's and the player's season!

Our problem corresponds to what is known as a *Multi-Class Classification Problem*. The model has to predict a discrete number of labels, i.e.: point guard, small guard, center...

Datasets and Inputs

The original dataset for this project has been taken from this repository at Kaggle, which in turn has scrapped it from Basketball-reference.

Nevertheless, the original dataset has been preprocessed as follows: - The number of features in the original dataset has been reduced from 52 to 22, to trying to reduce the impact of the curse of dimensionality. - The rows with an empty value in any of the features have been removed, to have a consistent set of features.

An unsupervised learning algorithm will be applied to the preprocessed version of the original dataset to select which features best describe a player position.

The dataset consists on 18609 samples. Each sample contains the following information:

Field	Description	Type
Year	Season	Numeric
Age	Age	Numeric
G	Games	Numeric
MP	Minutes played	Numeric
FG	Field goals	Numeric
FGA	Field goal attempts	Numeric
3P	3-point field goals	Numeric
3PA	3-point field goal attemps	Numeric
2P	2-point field goals	Numeric
2PA	2-point field goal attemps	Numeric
FT	Free throws	Numeric
FTA	Free throw attempts	Numeric
ORB	Offensive rebounds	Numeric
DRB	Defensive rebounds	Numeric
TRB	Total rebounds	Numeric
AST	Assists	Numeric
STL	Steals	Numeric
BLK	Blocks	Numeric
TOV	Turnovers	Numeric
PF	Personal fouls	Numeric
PTS	Points	Numeric
Pos	Position	String

Please notice that the last column (`Pos`, the player's position) corresponds to the feature we would like our model to predict based on the values of the other features.

In the table below it can be seen the different values that the `Pos` feature can contain, its description, and the number of samples per value.

Value	Description	# samples
C	Center	3737
PF	Power forward	3919
PG	Poing guard	3737
SF	Small forward	3547
SG	Shooting guard	3669

For your reference I have copied below the first eleven records of the dataset in both plain text and table formats:

Sample data in plain text:

```
Year,Age,G,MP,FG,FGA,3P,3PA,2P,2PA,FT,FTA,ORB,DRB,TRB,AST,STL,BLK,TOV,PF,PTS,Pos
1980,32,82,3143,835,1383,0,1,835,1382,364,476,190,696,886,371,81,280,297,216,2034,C
```

1980,25,67,1222,153,318,0,1,153,317,56,82,62,129,191,87,35,12,39,118,362,PF
1980,25,75,2168,465,875,0,2,465,873,188,236,158,451,609,322,108,55,218,237,1118,C
1980,31,80,2864,383,794,4,18,379,776,361,435,59,138,197,671,106,10,242,218,1131,PG
1980,31,26,560,27,60,0,0,27,60,32,50,29,86,115,40,12,15,27,66,86,C
1980,28,20,180,16,35,1,1,15,34,5,13,6,22,28,26,7,4,11,18,38,SG
1980,22,67,726,122,271,0,0,122,271,68,101,71,126,197,28,21,54,79,116,312,PF
1980,25,82,2438,545,1101,16,47,529,1054,171,227,240,398,638,159,90,36,133,197,1277,
SF 1980,28,77,2330,384,760,1,3,383,757,139,209,192,264,456,279,85,49,189,268,908,SF
1980,27,20,287,24,60,0,0,24,60,16,32,34,43,77,18,5,12,18,52,64,PF

Same sample data in table format:

Year	Age	G	MP	FG	FGA	3P	3PA	2P	2PA	FT	FTA	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	Pos
1980	32	82	3143	835	1383	0	1	835	1382	364	476	190	696	886	371	81	280	297	216	2034	C
1980	25	67	1222	153	318	0	1	153	317	56	82	62	129	191	87	35	12	39	118	362	PF
1980	25	75	2168	465	875	0	2	465	873	188	236	158	451	609	322	108	55	218	237	1118	C
1980	31	80	2864	383	794	4	18	379	776	361	435	59	138	197	671	106	10	242	218	1131	PG
1980	31	26	560	27	60	0	0	27	60	32	50	29	86	115	40	12	15	27	66	86	C
1980	28	20	180	16	35	1	1	15	34	5	13	6	22	28	26	7	4	11	18	38	SG
1980	22	67	726	122	271	0	0	122	271	68	101	71	126	197	28	21	54	79	116	312	PF
1980	25	82	2438	545	1101	16	47	529	1054	171	227	240	398	638	159	90	36	133	197	1277	SF
1980	28	77	2330	384	760	1	3	383	757	139	209	192	264	456	279	85	49	189	268	908	SF
1980	27	20	287	24	60	0	0	24	60	16	32	34	43	77	18	5	12	18	52	64	PF

Solution Statement

The dataset described in the previous section will be used as input of an unsupervised learning algorithm which, making use of principal component analysis, will return which features best describe a player position.

The features selected by the previous algorithm will be used as input features, using the player game position as the label to train a supervised learning algorithm.

The supervised learning algorithm will be trained with only the 80% of the dataset. The remaining 20% will be used to test the model and ensure that it can successfully predict a player's position based on his benchmark. The predictions and the labels of the testing set will be compared, and the model accuracy will be calculated.

Whenever an NBA trainer would like to reconsider the position of any of his team players, he will only need to enter the player stats corresponding to the features previously mentioned as input to the model, and it will return the players' predicted position.

Benchmark Model

Our soulution will be benchmarked against a neural network consisting in a SimpleRNN layer will be implemented.

The accuracy of the benchmark model and the new model will be compared. The objective is to discern if the combination of feature selection and hiperparameter tunning on decision trees can outperform a basic neural network approach.

Evaluation Metrics

In the *datasets and inputs* section it can be seen that the classes are balanced. That characteristic of our dataset will let us use the accuracy of the predictions to evaluate the performance of this project solution.

Project Design

The following techniques will be applied to complete the project goal: 1) Unsupervised learning will be used to find the set of features that best describes a player position. Principal component analysis (PCA) will help conclude the underlying structure of the dataset. The features that best describe a player position will be selected as the input to the supervised learning algorithm. The label will be the player position.

2) Supervised learning will be used to train a model that predicts player positions based on their stats. The model will be trained using the decision tree algorithm. The grid search technique will be used to optimise the 'max_depth' parameter for the decision tree. ShuffleSplit cross-validation technique will be used when training the model. The model performance will be validated against the testing set. The higher the accuracy is the more likely the model will be to be used by NBA trainers.

To select the best model and the solution for this exercise, it will be applied following rules: - Between models which perform similarly, the model with a smaller max_depth will be selected. - If there is a model which performs extremely well compared to the others, it will be selected regardless the max_depth.