Logistic Regression

Vito Založnik[1]

[1] *vz9592@student.uni-lj.si, 63210496*

1. Basketball shots analysis

1.1 Introduction

We were given a dataset with information about 5024 basketball shots. The dataset contained 3 categorical features (Competition, PlayerType, Movement) and 4 numerical (Transition, TwoLegged, Angle, and Distance). Our task was to use Multinomial logistic regression to provide insights into the relationship between shot type and the other features.

1.2 Preproccessing

All categorical features were encoded using One-Hot encoding. Since we did not use a closed-form solution for our model, for the sake of interpretability we did not need to drop the first feature while performing One-Hot encoding. Numerical features' Movement and Angle were normalized to $[0,1]$ using MinMax Scaler.

1.3 Shot analysis

After data preprocessing, adding an intercept and fitting the model, we obtained a weight matrix of dimensions 16×5. There were 6 different prediction classes. Class *other* was used as the reference value since it was the least interesting among all the different shot types. We trained the model 30 times on different bootstrap samples and calculated the mean of each weight and estimated its uncertainty using standard error (SE).

In table 1 are stored all regression weights. Weight w, from feature f and shot type s, can be interpreted as e^w is the odds ratio that associates the action with feature f to take shot s. From the analysis we could take next conclusions:

- Dunks are more likely performed by higher players (position C) and from smaller distances. They are also mostly two-legged and without dribbling or cutting. Dunks are more common in NBA than SLO or EURO. There are also almost no dunks in the U14 category.
- Tip-ins are, similar to dunks, mostly performed close to the rim, two-legged, and without dribbling or cutting. Angle does not seem to have any effect on it. Tip-ins are also not seen in U16 or even in the U14 category. They are more common in Europe than NBA.
- Layups are, opposite to dunks and tip-ins, preferable to perform when moving. There don't seem to be any differences between the age and league of players. Layups are still mostly performed two-legged and close to the rim.
- Hook-shot is, with some domain knowledge, known as one of the technically more difficult shots in basketball. This could also be seen when looking at weights at U14 and U16 features. There seems to be a high correlation between moving and not making the shot. Dribbling also has a negative impact on the odds of taking the shot. Distance also degrades the likelihood, but not as much as at dunk, tip-in, or layup.
- Above headshots are interestingly the only ones that are not negatively impacted by the distance and age of the players. They are typically performed when moving or dribbling or cutting and without momentum.
- Angle and transition are the least influential features over all shot types.

These seem to be pretty realistic results, backed up by practical explanations.

2. Ordinal regression

2.1 Introduction

We had to come up with a data-generating process, where ordinal logistic regression has a better log score than multinomial logistic regression. The size of the test set was set to 1000, the size of the train set was up to us.

2.2 Data generating procces

We know, that multinomial logistic regression has more parameters than ordinal logistic regression. While this means that multinomial regression can cover more complex problems this also means that it needs more training data, takes longer to build,... Ordinal logistic regression is, because of its simplicity, a more robust model that is harder to underfit. This is why we decided to take only 150 samples for our train dataset. We also unevenly distributed classes and made them ordered. Our DGP consisted of prediction classes $[0,1,2]$ with probabilities of occurring $[0.05, 0.3, 0.65]$. There were also three features: gender, income, and age.

$$\begin{aligned} gender &\sim Bernoulli(0.5) \\ income &\sim N(0,1) \\ age &\sim N(0,1) \end{aligned}$$

All features were generated as already normalized. Then we evaluated our models with the $log_score = \sum_i^N -ln(p_i)$ over all test samples where p_i is the probability of the model predicting the ground truth class of i^{th} test sample. From our DGP,

Feature\ShotType	dunk	tip-in	layup	hook shot	above head
Transifition	0.817 ± 0.085	0.457 ± 0.076	0.381 ± 0.027	-0.539 ± 0.049	-0.176 ± 0.020
TwoLegged	8.956 ± 0.165	8.428 ± 0.144	11.800 ± 0.212	1.530 ± 0.054	1.277 ± 0.039
Angle	-0.481 ± 0.080	-1.42 ± 0.099	0.940 ± 0.040	-1.092 ± 0.047	-0.126 ± 0.036
Distance	-24.906 ± 0.472	-35.245 ± 0.655	-14.185 ± 0.241	-6.600 ± 0.088	1.057 ± 0.069
EURO	1.244 ± 0.062	1.412 ± 0.064	-0.263 ± 0.025	0.186 ± 0.035	-0.197 ± 0.022
NBA	1.726 ± 0.070	0.691 ± 0.076	-0.092 ± 0.029	-0.690 ± 0.034	-0.418 ± 0.019
SLO	1.481 ± 0.076	1.803 ± 0.068	0.661 ± 0.041	0.513 ± 0.049	0.198 ± 0.038
U14	-9.086 ± 0.201	-8.065 ± 0.163	-0.862 ± 0.024	-2.375 ± 0.045	-1.156 ± 0.029
U16	-1.486 ± 0.212	-1.310 ± 0.178	-0.111 ± 0.037	-1.423 ± 0.037	-0.800 ± 0.025
PlayerType C	-0.832 ± 0.064	-1.273 ± 0.055	0.044 ± 0.026	-0.624 ± 0.034	-0.494 ± 0.032
PlayerType F	-2.691 ± 0.067	-2.125 ± 0.068	-0.346 ± 0.026	-1.334 ± 0.041	-0.833 ± 0.030
PlayerType G	-2.596 ± 0.068	-2.069 ± 0.051	-0.367 ± 0.023	-1.830 ± 0.033	-1.045 ± 0.027
Movement_dribble or cut	-5.298 ± 0.114	-5.269 ± 0.106	3.504 ± 0.057	-6.577 ± 0.122	3.510 ± 0.100
Movement_drive	-2.713 ± 0.069	-2.783 ± 0.074	3.184 ± 0.069	-4.654 ± 0.100	-10.601 ± 0.253
Movement_no	1.891 ± 0.055	2.585 ± 0.068	-7.359 ± 0.167	7.441 ± 0.137	4.717 ± 0.099
Intercept	-6.120 ± 0.144	-5.468 ± 0.119	-0.669 ± 0.051	-3.789 ± 0.080	-2.373 ± 0.062

Table 1. Matrix of regression weights with its standard error

We drew a training sample of size 150, 100 times, trained the models, and evaluated them on a test dataset of size 1000, again drawn from our DGP. We evaluated the uncertainty of our models with standard error. Multinomial logistic regression came with a log score of 836.415 ± 8.213, while ordinal logistic regression obtained a log score of 812.175 ± 2.395. We can see, that not only that the ordinal regression has a better log score, but it is also much more certain.