HW2: Logistic regression

Marko Ivanovski

Data preparation

For the following homework, we received a dataset that contained data about basketball shots divided into five categories. All the shots are described with seven features. Features competition, player type, and movement were discrete string variables and needed to be encoded in numerical representation. I used one-hot-encoding in order not to suggest any order if a mapping to natural numbers would be have been used. Also, not to violate the assumptions of collinearity made by Linear Regression models, I removed one vector from each one-hot-encoded variable. Next, I plotted the correlation heatmap to inspect further correlations. Figure 1 shows that the one-hot-encoded variable 'movement_no' and 'TwoLegged' were highly correlated. I decided to remove the 'TwoLegged' variable since the drop in accuracy was smaller than 1% during the multiple train-test splits I performed. Finally, I have standardized the two continuous variables angle and distance.

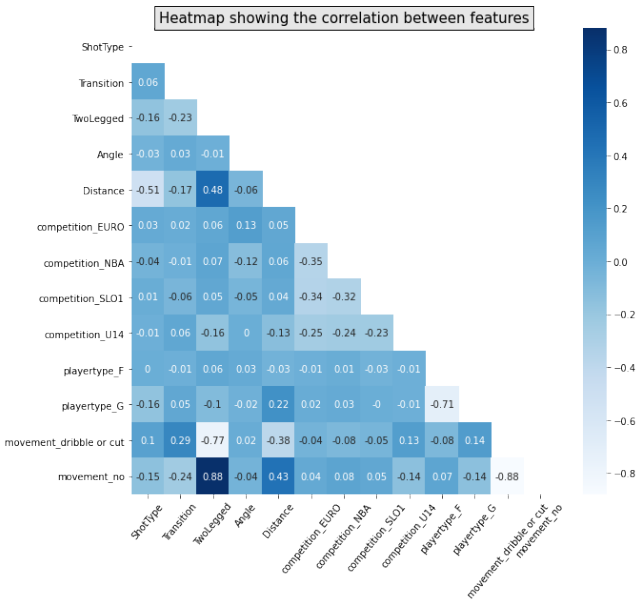

Figure 1. Correlation heatmap between variables.

Parameters interpretation

After I trained the categorical logistic regression model I obtained the matrix of parameters which can be seen in Figure 2. Since we fixed shot type other to be a reference point, all the analyses discussed below are relative to the shot being other.

The intercept

The intercepts give us a good idea of the category mean. If all variables were to be 0 we could say that the layup shot has the highest probability and above head has the lowest.

Numerical variables

For the numerical variables, the corresponding parameter value says how much does the log odds ratio increase (if $\beta > 0$) or decrease (if $\beta < 0$) for 1 unit change of the corresponding variable of that particular shot type. In our case, only the variables angle and distance are numerical. All the parameter values for the angle are close to zero, which easily suggests that this variable does not help in identifying the shot type. Finally, we have the distance which positively impacts only on the above head shot, which makes sense since above head shots are more likely made from a distance.

Binary variables(0,1)

Binary variables suggest how does the log odds change for the corresponding category if the variable is 1 relative to the variable being equal to 0. The transition variable being yes negatively impacts hook shot and above the head. If the shot were to be made in a transition, the log odds for it being a dunk are increased by 0.25. If the shot took place in NBA it increases the log odds of it being a dunk by 0.9 which is not surprising. Interestingly shots taken in the Slovenian national league increase the odds for all shot types, but mostly for hook shots. Finally, Slovenian U14 shots, decrease the odds for all shots, mostly for dunks and hook shots and the least for layouts. Both player types guard and forward do not dunk or make hook shots since they are further from the basket. Regarding the movement variable, it being dribble or cut decreases the log odds by 1.3 for the shot being hook, which makes sense since the defender has been outrun and increases the log only for the shot being above the head. Finally, movement no has a large magnitude for three categories which suggests it is a strong decider.

Probabilities

How do the log odds influence the probabilities? Following are some interesting examples with the expected probabilities for every category (the probabilities follow the order from the column of Figure 2).

	above head	layup	tip in	hook shot	dunk	other
intercept	-0.83	1.69	0.06	-0.17	0.23	0
transition	-0.15	0.14	0.05	-0.60	0.25	0
angle	0.00	0.02	-0.01	-0.01	-0.01	0
distance	0.24	-1.27	-1.66	-0.45	-1.37	0
competition_EURO	0.02	-0.18	0.56	0.60	0.43	0
competition_NBA	-0.04	-0.14	-0.05	0.03	0.90	0
competition_SLO1	0.31	0.40	0.56	0.87	0.35	0
competition_U14	-0.41	-0.37	-0.64	-0.84	-0.87	0
playertype_F	0.14	0.13	0.14	-0.09	-0.43	0
playertype_G	-0.12	0.11	-0.44	-0.77	-0.77	0
movement_dribble or cut	0.13	-0.03	-0.76	-1.41	-0.92	0
movement_no	2.01	-0.21	1.12	2.05	1.54	0

Figure 2. Matrix of parameters that form the linear predictor for every category.

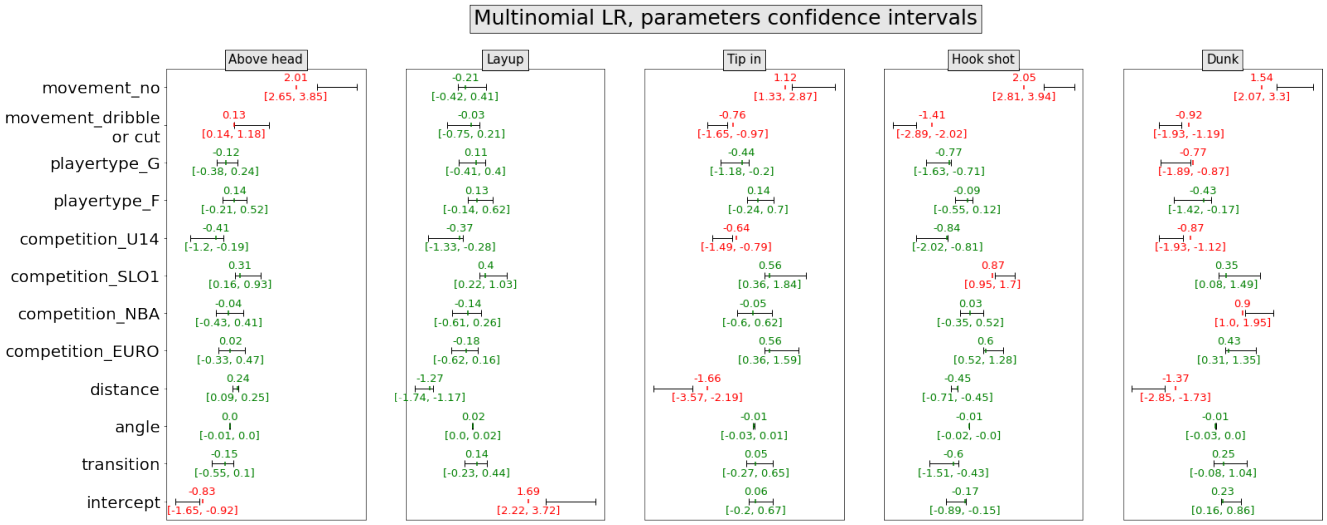

Figure 3. Multionmial LR parameters confidence intervals with the final parameter values when trained on the whole dataset.

- A shot that has been made in transition by the forward from a stationary point, at whichever angle and distance 9 in the Slovenian's first league have 50% chance of being above shot or hook shot.
- A shot made in transition by the forward using dribble or cut movement, at whatever angle and from distance 0.2 in Slovenian's first league is layup with 78% and tip-in 14%. Interestingly, changing only the league to NBA increases the shot being dunk to 15% and layup is decreased to 70%.
- A shot made in transition, at whatever angle, from distance 0.2 in the EURO league by the forward within stationary movement is 42% tip-in, 26% dunk, and 16% layup. Changing the movement to dribble or cut immediately increases the chance of it being layup to 47% and tip-in 33%.
-

Parameters uncertainty

Figure 3 illustrates the parameters uncertainty. I calculated the 95% confidence intervals using bootstrap (using 100 repetitions and n instances). Under the assumption the 100 values for each parameter forms a distribution, I calculated the 2.5% and 97% percentiles. It also happened that some final parameter values were outside their confidence intervals. This might be a consequence of not enough bootstrap repetitions or repetitions in the bootstrap samples.

Data generating process

Since multinomial logistic regression has more parameters, ordinal would work in cases when we have fewer data to fit these parameters. As such, my data generating process is quite simple, the vector X is the instance number with an added randomly generated number from a standard normal distribution. The vector Y is the reminder we get when dividing the integer part of X by 2. For training, I only used 10 instances and the log loss errors I got on 1000 instances for both models were:

- ordinal logistic regression: 16.11
- multinomial logistic regression: 786275.7