Machine Learning - Homework 3

Tim Kmecl, 10.4.2024

Implementation

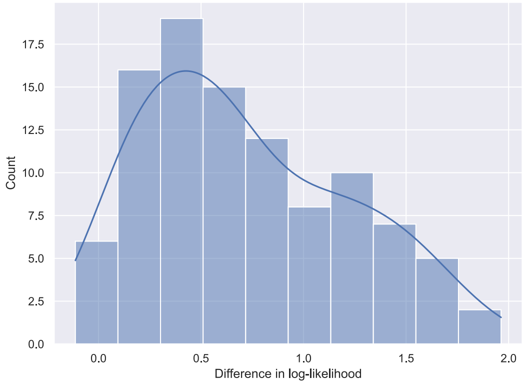
I implemented both logistic regressions as described. Both include intercept by default but can be disabled by passing an argument to the build method. Multinomial regression fixes preferences for the last class to 0 and ordinal fixes first threshold to 0, same as in the lectures. I compared the predictions and coefficients of my implementations of multinomial logistic regression to the one from sklearn on a toy dataset with 4 classes, and they return same results.

Data generating process

Multinomial regression is more flexible model and can model any data ordinal regression can, but also has more free parameters. Therefore on smaller datasets or those with large amount of classes ordinal regression should perform better as long as the DGP follows the assumptions of the ordinal regression.

My DGP generates data with 2 independent variables and 11 possible classes. Both two variables are gaussian, first with σ of 3 and the second of 1.5. 11 equidistant parallel lines are added to the plane. Classes are assigned to samples based on how many lines they are to the right of, when moved by a small gaussian noise (to make the classes not perfectly separable).

I chose the size of training data to be 200, and tested on 1000 samples. This was repeated 100 times on different random samples. Every time, difference between ordinal and multinomial regressions' log score was computed. Mean difference of log score is 0.72 with s.d. of 0.5. Ordinal regression performed worse only 3 times.



Data analysis

Preprocessing and building the model

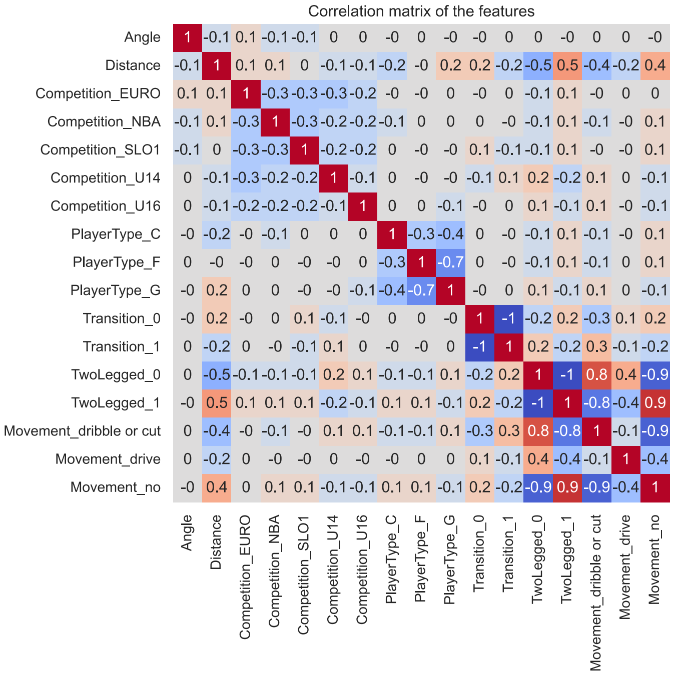
First, I one-hot encoded all the categorical features. The data was standardized to make coefficients for different features comparable among themselves. Additionally, this meant I didn't have to include the intercept. I trained the model on 100 different bootstrapped datasets, saved their coefficients and both their accuracy and log-score on OOB. Standardization was done based on only in-the-bag samples.

Means and 95% confidence intervals were calculated for coefficients (s.e. for accuracy and log-score). Mean accuracy of the model was 0.73±0.01, and log-score -0.74±0.02, so the model is clearly able to model the data to some degree, and the variability is not too high.

Features and their correlation

We can see from the correlation matrix on the next page that besides the negative correlations between one-hot encodings from the same original features, there are some other strong relationships. The higher the distance, the more likely a shot is two-legged and the less likely there is any movement. There is also strong relationship between two-leggedness and movement – if a shot is two legged, there most likely isn't any movement.

These relationships are important for interpretation, since coefficients of highly correlated features should be looked at and interpreted together.

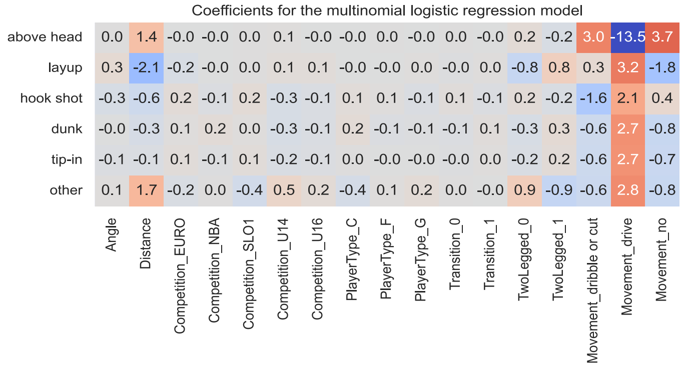


Correlation matrix of the features

	Angle	Distance	Competition_EURO	Competition_NBA	Competition_SLO1	Competition_U14	Competition_U16	PlayerType_C	PlayerType_F	PlayerType_G	Transition_0	Transition_1	TwoLegged_0	TwoLegged_1	Movement_dribble or cut	Movement_drive	Movement_no
Angle	1	-0.1	0.1	-0.1	-0.1	0	0	-0	0	-0	-0	0	0	-0	0	0	-0
Distance	-0.1	1	0.1	0.1	0	-0.1	-0.1	-0.2	-0	0.2	0.2	-0.2	-0.5	0.5	-0.4	-0.2	0.4
Competition_EURO	0.1	0.1	1	-0.3	-0.3	-0.3	-0.2	-0	-0	0	-0	0	-0.1	0.1	-0	0	0
Competition_NBA	-0.1	0.1	-0.3	1	-0.3	-0.2	-0.2	-0.1	0	0	0	-0	-0.1	0.1	-0.1	-0	0.1
Competition_SLO1	-0.1	0	-0.3	-0.3	1	-0.2	-0.2	0	-0	-0	0.1	-0.1	-0.1	0.1	-0	-0	0.1
Competition_U14	0	-0.1	-0.3	-0.2	-0.2	1	-0.1	0	-0	-0	-0.1	0.1	0.2	-0.2	0.1	0	-0.1
Competition_U16	0	-0.1	-0.2	-0.2	-0.2	-0.1	1	0	0	-0.1	-0	0	0.1	-0.1	0.1	-0	-0.1
PlayerType_C	-0	-0.2	-0	-0.1	0	0	0	1	-0.3	-0.4	0	-0	-0.1	0.1	-0.1	-0	0.1
PlayerType_F	0	-0	-0	0	-0	-0	0	-0.3	1	-0.7	0	-0	-0.1	0.1	-0.1	0	0.1
PlayerType_G	-0	0.2	0	0	-0	-0	-0.1	-0.4	-0.7	1	-0	0	0.1	-0.1	0.1	0	-0.1
Transition_0	-0	0.2	-0	0	0.1	-0.1	-0	0	0	-0	1	-1	-0.2	0.2	-0.3	0.1	0.2
Transition_1	0	-0.2	0	-0	-0.1	0.1	0	-0	-0	0	-1	1	0.2	-0.2	0.3	-0.1	-0.2
TwoLegged_0	0	-0.5	-0.1	-0.1	-0.1	0.2	0.1	-0.1	-0.1	0.1	-0.2	0.2	1	-1	0.8	0.4	-0.9
TwoLegged_1	-0	0.5	0.1	0.1	0.1	-0.2	-0.1	0.1	0.1	-0.1	0.2	-0.2	-1	1	-0.8	-0.4	0.9
Movement_dribble or cut	0	-0.4	-0	-0.1	-0	0.1	0.1	-0.1	-0.1	0.1	-0.3	0.3	0.8	-0.8	1	-0.1	-0.9
Movement_drive	0	-0.2	0	-0	-0	0	-0	-0	0	0	0.1	-0.1	0.4	-0.4	-0.1	1	-0.4
Movement_no	-0	0.4	0	0.1	0.1	-0.1	-0.1	0.1	0.1	-0.1	0.2	-0.2	-0.9	0.9	-0.9	-0.4	1

Interpreting the coefficients

Below, all the mean coefficients are shown for every category (type of shots). Positive ones (red) increase the likelihood of that type, whereas the negative ones (blue) decrease it.



Coefficients for the multinomial logistic regression model

	Angle	Distance	Competition_EURO	Competition_NBA	Competition_SLO1	Competition_U14	Competition_U16	PlayerType_C	PlayerType_F	PlayerType_G	Transition_0	Transition_1	TwoLegged_0	TwoLegged_1	Movement_dribble or cut	Movement_drive	Movement_no
above head	0.0	1.4	-0.0	-0.0	0.0	0.1	-0.0	-0.0	0.0	-0.0	0.0	-0.0	0.2	-0.2	3.0	-13.5	3.7
layup	0.3	-2.1	-0.2	-0.0	0.0	0.1	0.1	-0.0	-0.0	0.0	-0.0	0.0	-0.8	0.8	0.3	3.2	-1.8
hook shot	-0.3	-0.6	0.2	-0.1	0.2	-0.3	-0.1	0.1	0.1	-0.1	0.1	-0.1	0.2	-0.2	-1.6	2.1	0.4
dunk	-0.0	-0.3	0.1	0.2	0.0	-0.3	-0.1	0.2	-0.1	-0.1	-0.1	0.1	-0.3	0.3	-0.6	2.7	-0.8
tip-in	-0.1	-0.1	0.1	-0.1	0.1	-0.2	-0.1	0.0	-0.0	-0.0	-0.0	0.0	-0.2	0.2	-0.6	2.7	-0.7
other	0.1	1.7	-0.2	0.0	-0.4	0.5	0.2	-0.4	0.1	0.2	0.0	-0.0	0.9	-0.9	-0.6	2.8	-0.8

Below a plot of single features with the highest magnitude for each category with 95% CI.

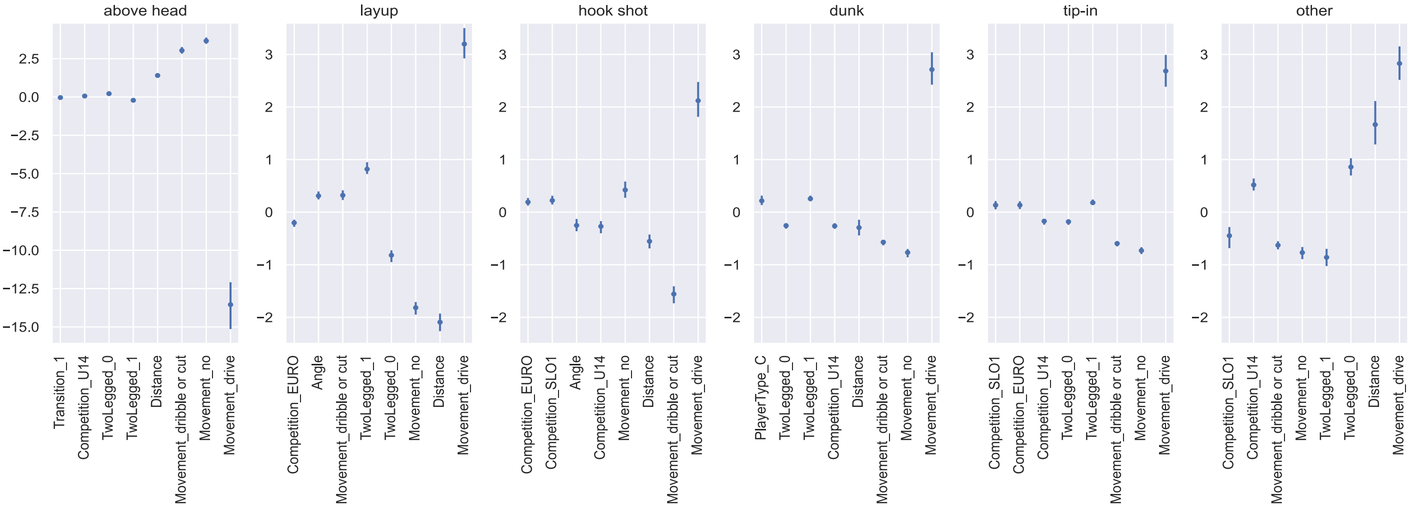
We can see the most important variable is movement, with drive massively decreasing probability of above head shot, and dribble or no movement increasing it strongly for it. Dribble decreases probability for hook shot and increases a bit for layup relative to others, but no movement decreases it for layup and increases for hook shot. This is in line with how these shots are performed in the game.

Distance also has a large effect. Large distance increases probability of a shot being above head or other, and decreases it for dun, hook shot and especially layup. These are the kind of shots performed relatively near the basket.

Next, we can see that transition and angle have little effect on the categorization. Similar can be said for player type, with the exception that centers decrease probability for other shots and increase a bit for dunks. We can also see that shots that are not two legged are more likely to be of type "other".

Finally let's look at the effect of different competitions. The most obvious is the difference between coefficients for U14 and the others. Competition being U14 decreases the probability for hook shots, dunks and tip-ins. This is probably because due to their lower height, these kind of shots are more difficult and therefore performed less.

As we can see from this analysis, most important variables predicting the kind of shot are distance and movement, with some of the others also affecting the prediction to a lesser degree. Their relative effects on different categories are in general similar to what I would expect based on the game itself, since they mostly have practical explanations. Interpreting the coefficients of multinomial regression therefore gives us a degree of insight into the game itself purely from the data.


above head
Transition_1
Competition_U14
TwoLegged_0
TwoLegged_1
Distance
Movement_dribble or cut
Movement_no
Movement_drive
layup
Competition_EURO
Angle
Movement_dribble or cut
TwoLegged_1
TwoLegged_0
Movement_no
Distance
Movement_drive
hook shot
Competition_EURO
Competition_SLO1
Angle
Competition_U14
Movement_no
Distance
Movement_dribble or cut
Movement_drive
dunk
PlayerType_C
TwoLegged_0
TwoLegged_1
Competition_U14
Distance
Movement_dribble or cut
Movement_no
Movement_drive
tip-in
Competition_SLO1
Competition_EURO
Competition_U14
TwoLegged_0
TwoLegged_1
Movement_dribble or cut
Movement_no
Movement_drive
other
Competition_SLO1
Competition_U14
Movement_dribble or cut
Movement_no
TwoLegged_1
TwoLegged_0
Distance
Movement_drive