Logistic regression

Blaž Erzar

Machine Learning For Data Science 1 2023/24
Homework 3 report

Abstract

This report presents the implementation of the multinomial and ordinal logistic regression models. The coefficients of the multinomial logistic regression can be used to gain insights into the target variable relationships, so we take advantage of this and explain the properties of target classes in the basketball shot dataset. Because this dataset does not contain ordinal classes, we generate a dataset from an artificial data-generating process, where ordinal logistic regression performs better than the multinomial one, with a 25% *higher accuracy and log score twice as small.*

1 Introduction

In this homework, we implemented two flavours of the logistic regression model, which are used for classification problems containing more than two classes. Such problems are multinomial. The first model, **multinomial** logistic regression trains a linear predictor for each class except one. These predictors predict latent strengths and are used to compute probabilities. On the other hand, **ordinal** logistic regression trains a single linear predictor that is used for classifying instances with classes that have some natural ordering.

We use the first model to gain insights into relations between shot type and features of the basketball shot dataset from [1]. In the end, we also present a data-generating process where ordinal logistic regression performs better than the multinomial one.

2 Methods

2.1 Multinomial regression

This model contains $m-1$ linear predictors with parameters β_j, where m is the number of different classes. Each predictor is used to compute the latent strength

$$u_{i,j} = \beta_j^\mathsf{T} \mathbf{x}^{(i)},$$

from which we compute class probabilities $P_i(j)$ using *softmax*. We optimise these parameters by maximising the log-likelihood

$$\ell(\beta) = \sum_{i=1}^{n} \log P_i\left(y^{(i)}\right)$$

using the L-BFGS-B optimisation method.

2.2 Data preprocessing

The dataset we are using contains both categorical and numerical features. To be able to use them, some preprocessing is needed. The class variable is first encoded using numbers $1, \ldots, m-1$ instead of names of classes. Numerical features *Angle* and *Distance* are normalised to the interval $[0, 1]$. We choose normalisation instead of standardisation because it makes sense that these values are positive. Finally, we encode categorical features using one-hot-encoding and remove one column for each feature because it is redundant.

2.3 Ordinal regression

In ordinal regression, we train a single linear predictor with parameters β. It is used to compute the latent parameter

$$u_i = \beta^\mathsf{T} \mathbf{x}^{(i)}$$

from which we compute the probabilities

$$P_i(j) = F(b_j - u_i) - F(b_{j-1} - u_i),$$

where F is the CDF of the standard logistic distribution and b_j are trained splits. For m classes we train $m-2$ splits together with the parameters. We use the same likelihood and optimisation as in the multinomial case.

3 Results

3.1 Shot type relationships

To measure the relationships between the target variable and other features we use coefficients of a trained multinomial logistic regression model. We obtain plots similar to feature importance for all except one class. The missing class in our case is *other*. The results are shown in Figure 1. We used bootstrap to estimate the uncertainty of these values. A high value indicates that a high feature value positively correlates with the specific class.

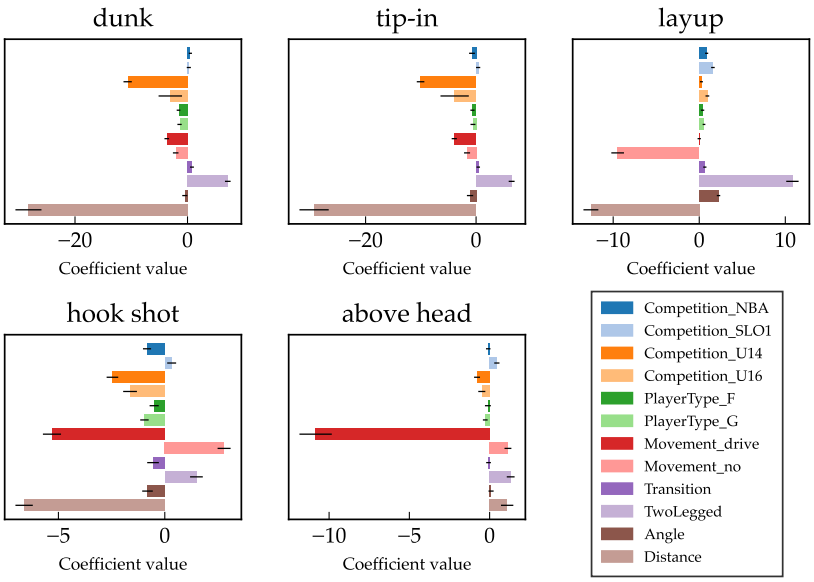

Figure 1: *Multinomial logistic regression coefficients for* 5 *classes of the basketball shot dataset.*

3.2 Ordinal regression better

For the basketball shot dataset, the ordinal logistic regression does not make sense because the classes have no natural ordering. We can create a dataset set where this is not the case. Let the target variable be

$$y = -0.2 \cdot \mathcal{N}(0,1) + 0.6 \cdot \mathcal{N}(-1,1) + 0.1 \cdot \mathcal{N}(1,1) + 0.5 \cdot \mathcal{U}[-1,1] + \mathcal{U}\{0,5\} + \varepsilon,$$

where ε is an error term distributed normally. We clip the values of y to integer values between 0 and 5. The comparison of both models for this data-generating process is shown in Table 1. The models were trained with 100 instances and trained with 1000.

Model	Accuracy [%]	Log score
Multinomial	48.8	−1317.39
Ordinal	73.7	−625.74

Table 1: *Test accuracies and log scores (higher is better) for the multinomial and ordinal logistic regression using the example data-generating process.*

4 Discussion

From Figure 1 we can conclude discriminative properties between different shot types. We can see that dunks are less common in the U14 league, which is expected due to younger and shorter players. Also, dunks are done close to the rim, so a negative coefficient for *Distance* is expected. The positive value for *TwoLegged* also makes sense because most dunks from under the basket are done with a two-legged jump.

Tip-in has similar coefficients to dunks. They are made close to the rim and require a high vertical jump from the player, which is usually two-legged.

Layups and hook shots are also shots made closer to the basket because the *Distance* coefficient is negative. Layups are done while running towards the basket so a negative *Movement_no* coefficient is expected. They are also mostly done two-legged. Hook shot, on the other hand, is done without movement, so the *Movement_no* coefficient is high and *Movement_drive* is low. These shots are less common in the younger leagues and they are usually not performed with a high angle towards the basket, since *Angle* value is negative.

The last shot type is an above-head shot. It is the only shot done from a distance and the *Distance* coefficient is positive. Similarly to a hook shot, these shots are done stationary because *Movement_drive* is very negative and *Movement_no* is slightly positive.

Finally, we can also observe in Table 1 that the created data-generating process does work better with the ordinal logistic regression. Comparing it with the multinomial version, the accuracy increases by 25% and the log score is twice as small.

References

[1] Frane Erčulj and Erik Štrumbelj. "Basketball Shot Types and Shot Success in Different Levels of Competitive Basketball". In: *PLOS ONE* 10.6 (June 2015), pp. 1–14. DOI: 10.1371/journal.pone.0128885. URL: https://doi.org/10.1371/journal.pone.0128885.