Multinomial and Regularization of Logistic Regression

CS109A Introduction to Data Science

Pavlos Protopapas, Natesh Pillai

Logistic Regression for predicting more than 2 Classes

There are several extensions to standard logistic regression when the response variable Y has more than 2 categories. The two most common are:

Ordinal logistic regression is used when the categories have a specific hierarchy (like class year: Freshman, Sophomore, Junior, Senior; or a 7-point rating scale from strongly disagree to strongly agree).

Multinomial logistic regression is used when the categories have no inherent order (like eye color: blue, green, brown, hazel, etc).

Logistic Regression for predicting more than 2 Classes

For example we could attempt to predict a student’s concentration:

$$y = \begin{cases} 1 & if\ \text{Computer Science (CS)} \\ 2 & if\ \text{Statistics} \\ 3 & \text{otherwise} \end{cases}$$

from predictors x_1 number of psets per week and x_2 how much time spent in library.

One vs. Rest (ovr) Logistic Regression

An option for nominal (not-ordinal) categorical logistic regression model is called the ’One vs. Rest’ approach.

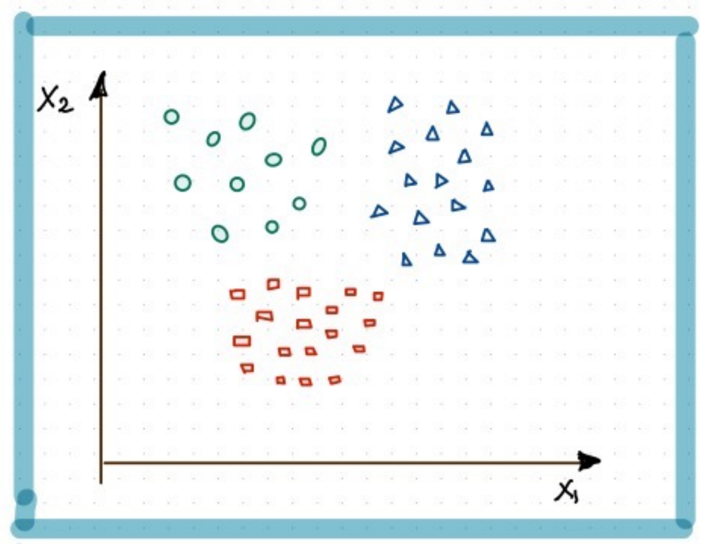
If there are 3 classes, then 3 separate logistic regressions are fit, where the probability of each category is predicted over the rest of the categories combined.

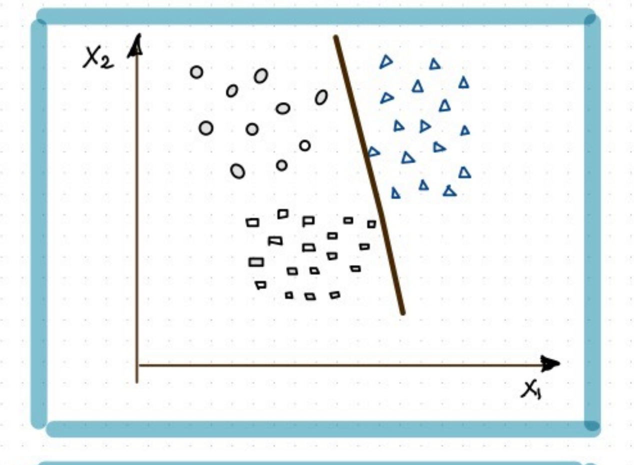
So for the concentration example, 3 models would be fit:

- a first model would be fit to predict CS from (Stat and Others) combined.
- a second model would be fit to predict Stat from (CS and Others) combined.
- a third model would be fit to predict Others from (CS and Stat) combined.

One vs. Rest (ovr)

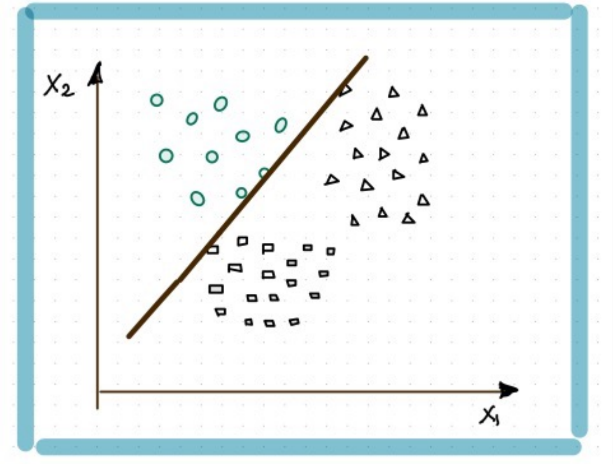
Classifying three classes, Red, Blue and Green can be turn into three binary Logistic Regressions





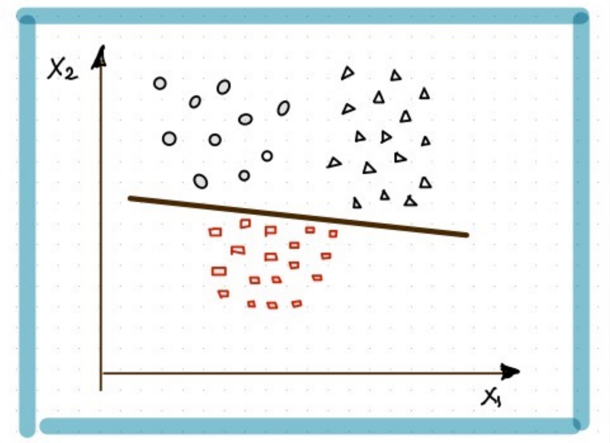
Blue vs others

$$\log\frac{P(b)}{P(o)} = \beta_b X$$



Green vs others

$$\log\frac{P(g)}{P(o)} = \beta_g X$$



Red vs others

$$\log\frac{P(r)}{P(o)} = \beta_r X$$

`sklearn` normalizes the output of each of the three models when predicting probabilities:

$$\tilde{P}(b) = \frac{P(b)}{P(g) + P(b) + P(r)}$$

$$\tilde{P}(g) = \frac{P(g)}{P(g) + P(b) + P(r)}$$

$$\tilde{P}(r) = \frac{P(r)}{P(g) + P(b) + P(r)}$$

‘True’ Multinomial Logistic Regression

Another option for multiclass a logistic regression model is the *true* ‘multinomial’ logistic regression model.

One of the classes is chosen as the *baseline group* (think $Y = 0$ group in typical logistic regression), and the other *K* – 1 classes are compared to it. Thus, a sequence or binary models (K-1) models are built to predict being in class *k* from class *K*:

$$\ln\left(\frac{P(Y = k)}{P(Y = K)}\right) = \beta_{0,k} + \beta_{1,k}X_1 + \cdots + \beta_{p,k}X_p$$

‘True’ Multinomial Logistic Regression

This is mathematically equivalent to using the softmax function:

$$P(y = k) = \frac{e^{X\beta_k}}{\sum_{k=1}^{K} e^{X\beta_k}}$$

And the cross-entropy as the loss function:

$$L = -\sum_{i}\sum_{k} \mathbb{I}(y_i = k) P\,(\mathrm{y_i = k})$$

‘multinomial’ vs. ‘ovr’

multinomial is slightly more efficient in estimation since there technically are fewer parameters (though `sklearn` reports extra ones to normalize the calculations to 1) and is more suitable for inferences/group comparisons.

ovr is often preferred for determining classification: you simply just predict from all 3 separate models (for each individual) and choose the highest probability.

They give VERY similar results in estimated probabilities and classifications.

Classification for more than 2 Categories

When there are more than 2 categories in the response variable, then there is no guarantee that $P(Y = k) \geq 0.5$ for any one category.

So any classifier based on logistic regression will instead have to select the group with the largest estimated probability.

The classification boundaries are then much more difficult to determine. We will not get into the algorithm for drawing these in this class.