Regresión logística

Ángel Mora Bonilla – Manuel Ojeda Aciego

Universidad de Málaga
Dpto. de Matemática Aplicada

Curso 2019–2020

What is Logistic Regression?

- Logistic regression is yet another technique borrowed by machine learning from the field of statistics.
- Linear regression is not capable of predicting probability.
  If you use linear regression to model a binary response variable, for example, the resulting model may not restrict the predicted Y values within 0 and 1.
- Here's where logistic regression comes into play, where you get a probability score that reflects the probability of the occurrence at the event.
- Predicting a qualitative response for an observation can be referred to as **classifying** that observation, since it involves assigning the observation to a category, or class.

What is Logistic Regression?

- The methods that are often used for classification first predict the probability of each of the categories of a qualitative variable, as the basis for making the classification.
- Logistic regression is an instance of **classification technique** that you can use to predict a qualitative response. For example, logistic regression models the probability that gender belongs to a particular category.
- Logistic regression belongs to a family, named Generalized Linear Model (GLM), developed for extending the linear regression model.
- For instance, for gender classification, where the response gender falls into one of the two categories, male or female, you'll use logistic regression models to estimate the probability that gender belongs to a particular category.

What is Logistic Regression?

- Logistic Regression is part of a larger class of algorithms known as **Generalized Linear Model** (glm).
- In 1972, Nelder and Wedderburn proposed this model with an effort to provide a means of using linear regression to the problems which were not directly suited for application of linear regression.
- They proposed a class of different models (linear regression, ANOVA, Poisson Regression etc) which included logistic regression as a special case.
- Logistic regression does not return directly the class of observations. It allows us to estimate the probability p of class membership.
- You need to decide the threshold probability at which the category flips from one to the other. By default, this is set to $p = 0{,}5$, but in reality it should be settled based on the analysis purpose.

Derivation of Logistic Regression Equation

- The fundamental equation of generalized linear model is:

$$g(E(y)) = \alpha + \beta x_1 + \gamma x_2$$

  Here, $g()$ is the link function, $E(y)$ is the expectation of target variable and $\alpha + \beta x_1 + \gamma x_2$ is the linear predictor (α, β, γ to be predicted).

- The role of link function is to *link* the expectation of y to linear predictor.

Derivation of Logistic Regression Equation

- For example, the probability $Pr(gender = female \mid longhair)$ of being gender female given longhair (abbreviated as $p(longhair)$) ranges between 0 and 1.
- Then, for any given value of longhair, a prediction can be made for gender.
- Given X as the explanatory variable and Y as the response variable, how should you model the relationship between $p(X) = Pr(Y = 1 \mid X)$ and X?
- The linear regression model represents these probabilities as:

$$p(X) = \alpha + \beta X$$

- The problem with this approach is that, any time a straight line is fit to a binary response that is coded as 0 or 1, in principle we can always predict $p(X) < 0$ for some values of X and $p(X) > 1$ for others.

Derivation of Logistic Regression Equation

- To avoid the previous problem, we can use the logistic function to model $p(X)$ that gives outputs between 0 and 1 for all values of X.
- The standard logistic function, for predicting the outcome of an observation given a predictor variable x, is the sigmoid, an s-shaped curve defined as

$$p = \frac{e^y}{1 + e^y} = \frac{1}{1 + e^{-y}}$$

where $y = \alpha + \beta x$, and p is the probability of event to occur given x.

- Then, we have

$$p(x) = p(event = 1 \mid x) = \frac{1}{1 + e^{-(\alpha + \beta x)}}$$

Derivation of Logistic Regression Equation

- Equivalently,

$$\frac{p}{1-p} = e^{\alpha+\beta x}$$

- By taking the logarithm of both sides, the formula becomes a linear combination of predictors:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta x.$$

- When we have multiple predictor variables, the logistic function looks like:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

- Note that a positive β_i indicates that increasing x_i will be associated with increasing p and vice versa.

Terminology

- The quantity $\log(\frac{p}{1-p})$ is the log-odd or logit.
- The odds reflect the likelihood that the event will occur. It can be seen as the ratio of "successes" to "non-successes".
- Technically, odds are the probability of an event divided by the probability that the event will not take place.
- For example, if the probability of being diabetes-positive is 0.3, the probability of "won't be" is $1 - 0{,}3 = 0{,}7$, and the odds are 1.0.
- The probability can be calculated from the odds as

$$p = \frac{Odds}{1 + Odds}$$

Evaluation of the model

- Akaike Information Criteria (AIC). The less the better.
- Null Deviance and Residual Deviance. The less the better, again.
- Confusion matrix.
- Receiver Operator Characteristic. It determines the model's accuracy using Area Under Curve (AUC). The bigger the better.

Important Points

- GLM does not assume a linear relationship between dependent and independent variables.
- However, it assumes a linear relationship between link function and independent variables in logit model.
- The dependent variable need not be normally distributed.
- It does not use OLS (Ordinary Least Square) for parameter estimation. Instead, it uses maximum likelihood estimation (MLE).
- Errors need to be independent but not normally distributed.

Things to consider

- Before proceeding to the fitting process, it is very important cleaning and formatting the data. This preprocessing step often is crucial for obtaining a good fit of the model and better predictive ability.
- Check for empty or small cells by doing a cross-tab between categorical predictors and the outcome variable. If a cell has very few cases (a small cell), the model may become unstable or it might not run at all.
- Separation or quasi-separation (also called perfect prediction), a condition in which the outcome does not vary at some levels of the independent variables.
- Logit models require more cases than OLS regression because they use maximum likelihood estimation techniques.

Things to consider

- It is sometimes possible to estimate models for binary outcomes in datasets with only a small number of cases using exact logistic regression.
- It is also important to keep in mind that when the outcome is rare, even if the overall dataset is large, it can be difficult to estimate a logit model.
- Many different measures of pseudo-R-squared exist. They all attempt to provide information similar to that provided by R-squared in OLS regression.
- Diagnostics: The diagnostics for logistic regression are different from those for OLS regression.