University : Lebanese university – Faculty of engineering – 3

Mini-Project

Prepared By:

Ali Yaghi - 5462

Hassan Awad - 5519

Directed By:

Dr. Mohammad Awdeh

Table Of Index

Logistic Regression

Introduction:

Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name logistic regression is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. The name multinomial logistic regression is usually reserved for the case when the dependent variable has three or more unique values, such as Married, Single, Divorced, or Widowed. Although the type of data used for the dependent variable is different from that of multiple regression, the practical use of the procedure is similar. Logistic regression competes with discriminant analysis as a method for analyzing categorical-response variables. Many statisticians feel that logistic regression is more versatile and better suited for modelling most situations than is discriminant analysis. This is because logistic regression does not assume that the independent variables are normally distributed, as discriminant analysis does. This program computes binary logistic regression and multinomial logistic regression on both numeric and categorical independent variables. It reports on the regression equation as well as the goodness of fit, odds ratios, confidence limits, likelihood, and deviance. It performs a comprehensive residual analysis including diagnostic residual reports and plots. It can perform an independent variable subset selection search, looking for the best regression model with the fewest independent variables. It provides confidence intervals on predicted values and provides ROC curves to help determine the best cutoff point for classification. It allows you to validate your results by automatically classifying rows that are not used during the analysis

The Logit and Logistic Transformations:

In multiple regression, a mathematical model of a set of explanatory variables is used to predict the mean of a continuous dependent variable. In logistic regression, a mathematical model of a set of explanatory variables is used to predict a logit transformation of the dependent variable. Suppose the numerical values of 0 and 1 are assigned to the two outcomes of a binary variable. Often,

the 0 represents a negative response and the 1 represents a positive response. The mean of this variable will be the proportion of positive responses. If p is the proportion of observations with an outcome of 1, then 1-p is the probability of a outcome of 0. The ratio p/(1-p) is called the odds and the logit is the logarithm of the odds, or just log odds. Mathematically, the logit transformation is written:

$$l = \text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$$

The Log Odds Ratio Transformation:

The difference between two log odds can be used to compare two proportions, such as that of males versus females. Mathematically, this difference is written:

$$\begin{aligned} l_1 - l_2 &= \text{logit}(p_1) - \text{logit}(p_2) \\ &= \ln\left(\frac{p_1}{1-p_1}\right) - \ln\left(\frac{p_2}{1-p_2}\right) \\ &= \ln\left(\frac{\left(\frac{p_1}{1-p_1}\right)}{\left(\frac{p_2}{1-p_2}\right)}\right) \\ &= \ln\left(\frac{p_1(1-p_2)}{p_2(1-p_1)}\right) \\ &= \ln(OR_{1,2}) \end{aligned}$$

This difference is often referred to as the log odds ratio. The odds ratio is often used to compare proportions across groups. Note that the logistic transformation is closely related to the odds ratio. The reverse relationship is:

$$OR_{1,2} = e^{(l_1 - l_2)}$$

The Logistic Regression and Logit Models:

In logistic regression, a categorical dependent variable Y having G (usually G = 2) unique values is regressed on a set of p independent variables X1 X2 Xp , ,..., . For example, Y may be presence or absence of a disease, condition after surgery, or marital status. Since the names of these partitions are arbitrary, we often refer to them by consecutive numbers. That is, in the discussion below, Y will take on the values 1, 2, … G. In fact, NCSS allows Y to have both numeric and text values, but the notation is much simpler if integers are used. Let:

$$\mathrm{X} = \left(X_1, X_2, \cdots, X_p\right)$$

$$B_g = \begin{pmatrix} \beta_{g1} \\ \vdots \\ \beta_{gp} \end{pmatrix}$$

The logistic regression model is given by the G equations

$$\ln\left(\frac{p_g}{p_1}\right) = \ln\left(\frac{P_g}{P_1}\right) + \beta_{g1}X_1 + \beta_{g2}X_2 + \cdots + \beta_{gp}X_p$$

$$= \ln\left(\frac{P_g}{P_1}\right) + \mathrm{XB}_g$$

Here, pg is the probability that an individual with values X1 X2 Xp , ,..., is in outcome g. That is,

$$p_g = \Pr\left(Y = g \mid \mathrm{X}\right)$$

Usually X1 ≡ 1 (that is, an intercept is included), but this is not necessary. The quantities P P PG , ,..., 1 2 represent the prior probabilities of outcome membership. If these prior probabilities are assumed equal, then the term ( ) 1 ln Pg / P becomes zero and drops out. If the priors are not assumed equal, they change the values of the intercepts in the logistic regression equation. Outcome

one is called the reference value. The regression coefficients 11 12 1p β ,β ,Λ,β for the reference value are set to zero. The choice of the reference value is arbitrary. Usually, it is the most frequent value or a control outcome to which the other outcomes are to be compared. This leaves G-1 logistic regression equations in the logistic model. The β 's are population regression coefficients that are to be estimated from the data. Their estimates are represented by b's. The β 's represents unknown parameters to be estimated, while the b's are their estimates. These equations are linear in the logits of p. However, in terms of the probabilities, they are nonlinear. The corresponding nonlinear equations are

$$\mathrm{p_g} = \mathrm{Prob}(Y = g \mid \mathrm{X}) = \frac{e^{\mathrm{XB}_g}}{1 + e^{\mathrm{XB}_2} + e^{\mathrm{XB}_3} + \cdots + e^{\mathrm{XB}_G}}$$

since 1 X 1 = B e because all of its regression coefficients are zero. A note on the names of the models. Often, all of these models are referred to as logistic regression models. However, when the independent variables are coded as ANOVA type models, they are sometimes called logit models.

A note about the interpretation of e XB may be useful. Using the fact that ( )( ) a b a b e = e e + , XB e may be reexpressed as follows

$$e^{\mathrm{XB}} = e^{\beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \cdots + \beta_p \mathrm{X}_p}$$
$$= e^{\beta_1 \mathrm{X}_1} e^{\beta_2 \mathrm{X}_2} \cdots e^{\beta_p \mathrm{X}_p}$$

Working on a Movies Dataset:

Our project works on a dataset that contains more than 650 movies which contains the name, directors and a lot of infos about the achievement, our aim is to predict if a given movies won an Oscar depending in this dataset.

Step 1:

We first removes all infos that are not related on our goal (like the date of publish) to reduce the amount of extra info that can affect the performance of the code and the null values that is not important for us.

Step 2:

Then we convert all the categorical features to a numerical one (for example if a attribute ‘nominated for an award’ occurs, we assign to is 1 for true and 0 for false).

Step 3:

After we have a numerical dataset, the aim is to reduce the processing time by scaling our numbers (means normalize the huge numbers to have smallest one),

This could help us to process our dataset with the lowest cost.

Step 4:

Our dataset is now ready, we need now to start learning on it, so we split it into 20% for testing and 80% for training, the model is now learned the dataset and ready to predict if a given movie won a Oscar or not.

Step 5:

We now have to start to test the model and check its performance and feature just to make sure that everything is going well, if not, we try to improve the dataset by removing annoying data or inserting important one.