Lecture 101

Date: 20 March 2020
Scribes: Akash Gupta
Topics: Machine Learning : Classification

0.1 Overview

We start out with an overview of classification techniques which are essentially techniques developed where our response variable is a categorical or qualitative variable. This overview contains the bullet points and definitions for condensed clarity.

- Classication is process by which we predict qualitative variables. In essence predicting a qualitative response for an observation means classifying that observation - grouped into a category or class.
- The basis for applying classification to an observation involves predicting the probability of this observation belonging to the possible classes.
- As an example we would like to know things like :- If there is a banking service and it has information of customer data like account information, personal details, etc. That whether some person would be grouped into a 'high risk of default' or 'low risk of default' catefory.
- As demonstrations, we will use the default data set wherein we will attempt to predict whether a given individual will default on his credit card payment based on annual income and monthly credit card balance.
- At the outset it appears that those who defaulted on credit card payments also had a high monthly credit card balance.
- Y : default, X_1 : balance, X_2 : income
- It is generally not advisable to use linear regression techniques to predict qualitative responses since if we code our response variable outcomes differently, we will get different results in the linear regression model. This will lead to wrong predictions.

0.2 Logistic Regrssion

Default takes on either 'Yes' or 'No' and instead of modelling the response directly, logistic regression models that probability that the response variable belongs to either category :- models the probability of default. Probability of default given a certain value of balance can be denoted as :

$$Pr[default = yes|balance] = p(balance) \tag{1}$$

$$Pr[Y = 1|X_i] = p(X_i) \tag{2}$$

These values will typically lie between 0 and 1 and we will in most cases assign 'Yes' to default if $p(balance) > 0.5$. We use here a general 0/1 coding.

0.2.1 Logistic Model

In a typical linear model like : $p(Y) = \beta_0 + \beta_1 X$ we would obtain negative or greater than one values of probability which is basically nonsense values. So we model the probability using logistic function that always gives us values between 0 and 1 and is an 'S' shaped curve. It would give the average fitted probability as 0.033 which is approximately equal to the default rate in our dataset. It is given by :

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} \tag{3}$$

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X} \tag{4}$$

Here equation 4 LHS represents 'odds'. An example of how to read odds : on an average 1 in 5 people will default means an odds of 1/4 of default. This is given by :

$$\frac{0.2}{1 - 0.2} \rightarrow p = 0.2 \tag{5}$$

In equation 4 when we take log on both sides we obtain :

$$\log(\frac{p(X)}{1 - p(X)}) = \beta_0 + \beta_1 X \tag{6}$$

Where the LHS is called **Logit** or 'log-odds'. Logistic regression model has a logit that is linear in X. Our interpretation of β_1 here is : a one unit change in X causes β_1 unit change in log-odds, or a one unit change in X causes a : $odds \times e^{\beta_1}$ change in odds. Note that a change in $p(X)$ due to X depends on the value of X.

0.2.2 Estimating regression coefficients

Here we will be using Maximum likelihood method to estimate our regression coefficients. While finding coefficient estimates, we are essentially trying to find $\hat{p}(X)$ such that these predicted probability measures are as close as possible to the actual observed default status values of different observations. We are basically selecting β_0 and β_1 such that when we put these values into our model, we get $\hat{p}(X)$ as close as possible to 1 for defaulters and close to 0 for non defaulters. The likelihood function is given by :

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(X_i) \prod_{i':y'_i=0} (1 - p(X'_i)) \tag{7}$$

The estimates $\hat{\beta_0}$ and $\hat{\beta_1}$ are chosen such that likelihood function is maximised. After fitting the data we get the coefficient values shown in the table below. Note that a $\hat{\beta_1}$ value of 0.0055 is interpretted as :- a unit increase in X leads to a 0.0055 increase in log-odds of default.

Attributes	Coefficient	Std. Error	Z-stat	p-value
Intercept	-10.6513	0.3612	-29.5	< 0.0001
Balance(X)	0.0055	0.0002	24.9	< 0.0001

We test significance of our coefficients through a z test statistic given by :

$$z = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \tag{8}$$

Our sample size generally is quite large so we can apply the t formula to a z stat and compute the z value for testing the hypothesis that : $H_0 : \beta_1 = 0$ which if true, would imply that balance does not affect default probabilities at all. Here our p value is extremely small so we conclude that there is a significant relation between balance and default probability.

0.2.3 Making predictions

We can make predictions of probability of default by simply plugging in the values of estimated coefficients into the equation as given below. We try to estimate the probability of default for someone with a balance of \$1000 :

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}} \tag{9}$$

$$\hat{p}(X) = \frac{e^{-10.6513 + 0.0055 * 1000}}{1 + e^{-10.6513 + 0.0055 * 1000}} = 0.00576 \tag{10}$$

In a similar fashion, we can even make predictions for probability of default using a qualitative predictor like - Student (1 or 0). As before, this will give the average probability of default if someone is a student or non-student. Generally it is seen in our model that students have a higher probability of default. Note for future reference that coefficient for Student[yes] is 0.4049 and the average levels of default probability for students and non students respectively is : 0.0431 and 0.0292 respectively.

0.2.4 Multiple logistic regression

The multiple logistic regression equations for logit and probability are given as follows respectively :

$$\log(\frac{p(X)}{1 - p(X)}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{11}$$

$$p(X) = \frac{e^{\beta_0 + \beta_1 X + \beta_2 X_2 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X + \beta_2 X_2 + \cdots + \beta_p X_p}} \tag{12}$$

We now present our coefficient estimates from the multiple logistic regression with three predictor variables (balance (X_1), income (X_2), student (X_3)) here :

Attributes	Coefficient	Std. Error	Z-stat	p-value
Intercept	-10.8690	0.4923	-22.08	< 0.0001
Balance(X_1)	0.0057	0.0002	24.74	< 0.0001
Income(X_2)	0.0030	0.0082	0.37	0.7115
Student[yes](X_3)	-0.6468	0.2362	-2.74	< 0.0062

Note a surprising result here that the coefficient value for students in multi regression is negative while in single regression is positive. This would suggest that students are less likely to default than non-students. The negative coefficient means that keeping balance and income constant/fixed, a student is less likely to default than a non-student. But in the single regression case, average overall default rates of students (0.04)are higher than that of non-students (0.02) - hence the positive single coefficient. This contradiction is primarily because balance is correlated with student. This means that since students tend to hold higher levels of debt - tend to have a higher credit card balance and since we know that balance has a pretty strong association with default probability, overall student default rate tends to be higher. **Note that even though overall default rate for students is higher than that of non-students, default rate for a given value of balance, the probability of default among students is seen to be lesser than that of non-students**. When such correlations among predictors exist, it is called **confounding**. Finally we can find predicted probabilities of default by simply putting the values for the estimated coefficients and observations in the model equations. We will now look at how to classify among more than 2 categories with **Discriminant analysis** later on.