Week 7 Topics

1. Chapter 10: Introduction Logistic Regression

Logistic regression extends the ideas of linear regression to the situation where the outcome variable, Y, is categorical. Logistic regression can be used for classifying a new record, where the class is unknown.

Logistic regression can be seen as a special case of generalized linear regression model and thus analogous to linear regression. The model of logistic regression, however, is based on quite different assumptions (about the relationship between dependent and independent variables) from those of linear regression. In particular the key differences of these two models can be seen in the following two features of logistic regression

* First, the conditional distribution  is a Bernoulli distribution rather than a Gaussian distribution, because the dependent variable is binary.
* Second, the predicted values are probabilities and are therefore restricted to (0,1) through the logistic distribution function because logistic regression predicts the probability of particular outcomes.

In this section, we focus on the use of logistic regression for classification. We deal only with binary outcome variable having two possible classes. Popular example of binary outcomes are success/failure, yes/no, and similar.

In some cases we may choose to convert a continuous outcome variable or an outcome variable with multiple classes into a binary outcome variable for the purpose of simplification. As with the multiple linear regression, the predictor variables X1, X2, …, Xk may be categorical variables, continuous variables, or a mixture of these two types. While in multiple linear regression, the aim is to predict the value of the continuous Y for a new record, in logistic regression the goal is to predict which class a new record will belong to, or simply to classify the new record into one of the classes.

In logistic regression, we take two steps:

* The first step yields estimates of the *propensities* or *probabilities* of the belonging to each class. In the binary case, we get an estimate of p = P(Y=1). The probability of belonging to class 1 (which automatically tells us the probability of belonging to class 0, q = 1-p).
* In the second step, we use a cutoff value on these probabilities in order to classify each case into one of the classes. For example, a cutoff of 0.5 means that cases with an estimated probability of P(Y=1) >= 0.5 are classified as belonging to class 1, whereas cases with P(Y=1)< 0.5 are classified as belonging to class 0.

1. The Logistic Regression Model

The idea behind the logistic regression is straightforward. Instead of using *Y* directly as the outcome variable, we use a function of it, which is called ***logit.*** The logit, it turns out, can be modeled as a linear function of predictors. Once the logit has been predicted, it can be mapped back to a probability.

To understand the logit, we take several intermediate steps. These steps are:

* First, we look at the *p* = P(Y=1), the probability of belonging to class 1 (as opposed to class 0). In contrast to the binary variable Y, which only takes the value0 and 1, *p* can take any value in the interval of [0,1]. However, if we express p as a linear function of k predictors in the form

It is not guaranteed that the right-hand side will lead to values within the interval [0, 1].

The solution is to use a nonlinear function of the predictors in the form

This is called the logistic response function, for any values of X1, X2, …, Xk , the right-hand side will be always lead to values in the [0, 1] interval.

* Next, we look at a difference measures of belonging a certain class, knowns as **Odds**. In probability of a particular event, odds means the ratio between the number of favorable outcomes to the number of unfavorable outcomes. Odds in favor and odds in against.

**Odds in favor**: Odds in favor of a particular event are given by Number of favorable outcomes to Number of unfavorable outcomes.

**Odds against**: Odds against is given by Number of unfavorable outcomes to number of favorable outcomes.

In probability, the odds (favorable odds) of belonging to class 1 are defined as the ratio of the probability of belonging to class 1 to the *probability of belonging to class 0*.

And the odds of unfavorable outcome is

To explain better the meaning of Odds, let’s consider this scenario: Instead of talking about the *probability* of contracting a disease, people talking about the odds of contracting the disease. So, if the probability of contracting a disease is 35% the odds of contracting the disease is which is about 54%.

We can use the Odds to calculate the probability as

Let’s use the above to calculate the odds based on the *p* function

Let’s assume Q = then we can write the Odds function as:

Which is:

Substituting Q we will have:

And the log in base *e* of Odds is the regression function!

We can see also, the log(Odds) is **logit**

1. Classification with Logistic Regression

As we learned in the previous section the logistic regression function is written as follow:

Where P is a probability distribution and we know the logarithm of Odds is

Which is called logit. So we can rewrite the probability outcome as:

This way we can calculate the probability outcome and with using a cutoff decide what is the class of the outcome .

1. Example

Probability of passing an exam versus hours of study

A group of 20 students spend between 0 and 6 hours studying for an exam. How does, the number of hours spent studying, affect the probability that the student will pass the exam?

The table below (Figure 5.1) shows the number of hours each student spent studying, and whether they passed (1) or failed (0).

Remembering Bernoulli probability distribution from stat course which states:  is the probability distribution of a random variable which takes the value 1 with success probability of ***P*** and the value 0 with failure probability of ***q = 1 - P***. The Bernoulli distribution is a special case of the “two point distribution”, for which the two possible outcomes need not be 0 and 1. Bernoulli distribution is also a special case of the binomial distribution; the Bernoulli distribution is a binomial distribution where n=1 (number of try is 1).

Therefore, when we are using the R package for logistic regression we choose family = “binomial”

1. Logistic Regression with R

We use a small dataset with one predictor. This dataset, contains information about a group of 20 students spend between 0 and 6 hours studying for an exam and their success or failure in that exam. We want to know how does, the number of hours spent studying, affect the probability that the student will pass the exam? and use the model to predict new student success or failure in exam.

The table below (Figure 1) shows the number of hours each student spent studying, and whether they passed (1) or failed (0).

|  |  |
| --- | --- |
| Hours | Pass Exam(No=0, Yes=1) |
| 0.50 | 0 |
| 0.75 | 0 |
| 1.00 | 0 |
| 1.25 | 0 |
| 1.50 | 0 |
| 1.75 | 1 |
| 2.00 | 0 |
| 2.25 | 1 |
| 2.50 | 0 |
| 2.75 | 1 |
| 3.00 | 0 |
| 3.25 | 1 |
| 3.50 | 0 |
| 3.75 | 1 |
| 4.00 | 1 |
| 4.25 | 1 |
| 4.50 | 1 |
| 4.75 | 1 |
| 5.00 | 1 |
| 5.50 | 1 |

Figure 1: Hours of Study and Success in Exam

The following steps and R code will help us to build a model based on Logistic Regression, evaluate it, and use it with a level of confidence for predicting new cases

1. Loading data into R

ex<-read.csv("Logistics Regression Example.csv")

View(ex)

1. The target attribute (PassExam) is numeric with change it into categorical

ex$PassExam<-factor(ex$PassExam)

If the data was large we should partition it into training and validation datasets. But we go directly to model building. Make sure the algorithm uses the binomial family approach!

install.packages("glm")

logit.reg<-glm(ex$PassExam ~ ., data = ex, family = "binomial")

options(scipen = 999) # to avoid scientific notation numbers

1. View the outcome summary

summary(logit.reg)

Call:

glm(formula = ex$PassExam ~ ., family = "binomial", data = ex)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.7704 -0.5147 0.1969 0.5071 1.7562

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.9359 1.7804 -2.211 0.0271 \*

Hours 1.5055 0.6271 2.401 0.0164 \*

Logit = -3.9359 +1.5055\*Hours

---

Signif. codes: 0 ë\*\*\*í 0.001 ë\*\*í 0.01 ë\*í 0.05 ë.í 0.1 ë í

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 27.526 on 19 degrees of freedom

Residual deviance: 15.964 on 18 degrees of freedom

AIC: 19.964

Number of Fisher Scoring iterations: 5

As you see the effect of the hours of study on success is low (only one star)

1. Using the model to predict the same data (as you know this is not a good method but here we just limited to small dataset. We do it to just learn the method. In the real case we validate it using validation dataset ☺)

ex.predicted<-predict(logit.reg, ex)

T<-data.frame(ex, ex.predicted)

T

Hours PassExam ex.predicted

1 0.50 0 -3.1831967

2 0.75 0 -2.8068259

3 1.00 0 -2.4304552

4 1.25 0 -2.0540844

5 1.50 0 -1.6777136

6 1.75 1 -1.3013429

7 2.00 0 -0.9249721

8 2.25 1 -0.5486014

9 2.50 0 -0.1722306

10 2.75 1 0.2041401

11 3.00 0 0.5805109

12 3.25 1 0.9568816

13 3.50 0 1.3332524

14 3.75 1 1.7096231

15 4.00 1 2.0859939

16 4.25 1 2.4623646

17 4.50 1 2.8387354

18 4.75 1 3.2151061

19 5.00 1 3.5914769

20 5.50 1 4.3442184

1. Create a plot of the results. First we get the range of the predictor of interest (here is only one, Hours). Then we create a list of this range and use the list to predict and plot the prediction values.

range(ex$Hours)

[1] 0.5 5.5

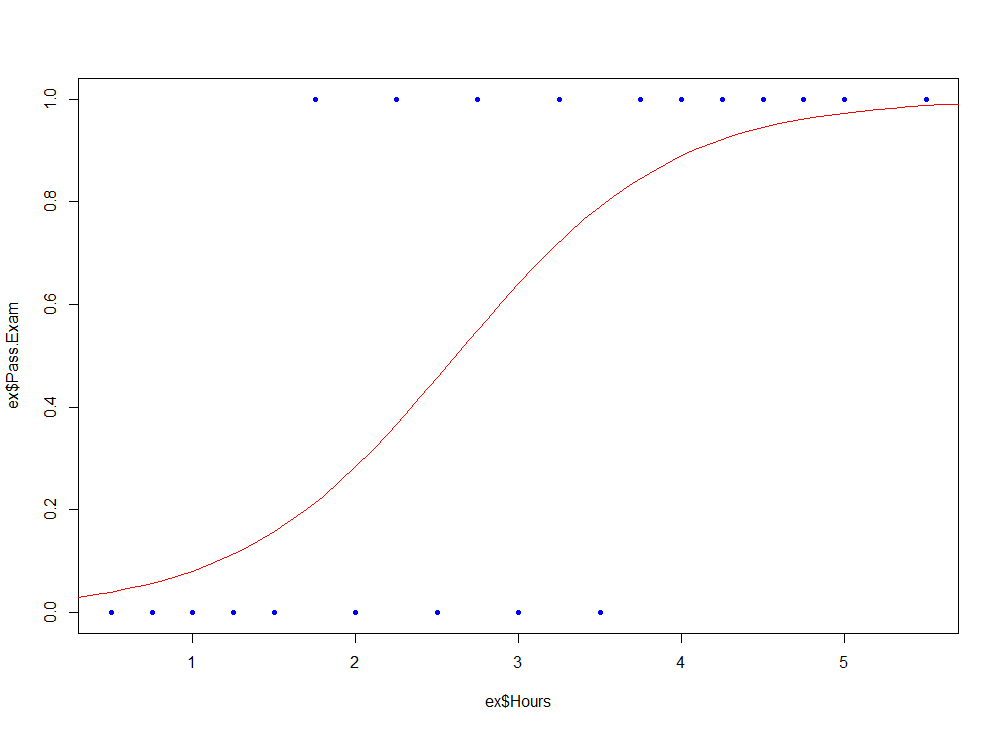
xex<-seq(0, 6, 0.1)

yex<-predict(logit.reg, list(Hours = xex), type = "response")

plot(ex$Hours, ex$PassExam)

lines(xex, yex)

As you see the list made by the predictor is used as the X axis values and predicted values for Y Axis.



It seems students studying more than 2.5H has more chance to success.