Machine Learning
Artur d'Avila Garcez
6th Tutorial – Logistic Regression

Background

Logistic regression seeks to model the relationship between a dependent variable and one or more independent variables. As in the case of *linear regression*, logistic regression allows us to look at the fit of the model as well as the significance of the relationships between dependent and independent variables. However, while linear regression uses *least squares* to find a best fitting curve and come up with coefficients that predict the change in the dependent variable given changes in the independent variables, logistic regression estimates the probability of an event occurring (e.g. the probability of a person staying in education post 16 years of age). What one wants to predict from knowledge of the relevant independent variables is not a precise numerical value of a dependent variable, but rather the probability (p) that an event will occur. This means that, in logistic regression, the following function is used:

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

where *P* is the probability of the event, *e* is the base of the natural logarithm (approximately 2.718) and α and β are the parameters of the model (as in linear regression). The value of α yields *P* when *x* is zero, and β tells us how *P* should change with *x*.

An important concept in logistic regression is that of *odds ratios*. Logistic regression, being based on the probability of an event occurring, allows us to calculate these, which are defined as the ratio between the probabilities of an event occurring to it not occurring. For example, suppose that the probability p of staying in education post 16 is 0.8. The probability of not staying in education will be q = 1 - p = 0.2. The odds of staying on are then defined as:

odds(staying on) = p/q = 4 (or 4 to 1)

The odds of not staying on would be:

odds(not staying on) = q/p = 0.25 (1 to 4, or 4 to 1 against[1])

[1] The use of the term in statistics and gambling is not consistent (with the exception of horse racing). Gambling odds are expressed in the form "X to Y" and it is implied that the odds are odds against, i.e. with X possible outcomes in which the event will not take place.

In logistic regression, the dependent variable is a *logit*, which is a log of odds:

$$logit(P) = \ln(\frac{P}{1-P})$$

Since:

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

we can get back to our original logistic regression function:

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

showing how we can get from odds ratios to probabilities and vice versa.

Setup – Matlab

We will run a simple example of logistic regression over a binomial data distribution.

1. Unzip “logisticRegression.zip” into any folder of your choice;
2. Open Matlab and inside it, locate the folder where the contents of “logisticRegression.zip” have been unzipped;
3. Double click on “logisticRegression.m” inside Matlab;
4. Left-click the editor and press F5 to run the code.

You should see a logistic curve and the fitted points as circles, around the edges of the curve.

In the above example, the distribution is “binomial” and the link function is “logit”; the link function f does the following to the linear models seen in class:

$$y = f(w,\phi(x))$$

Using link function “identity” allows you to simulate the linear models seen in class.

Exercises

1. Read the description of the Matlab functions:

   glmfit (http://uk.mathworks.com/help/stats/glmfit.html)
   glmval (http://uk.mathworks.com/help/stats/glmval.html),

   used on lines 13 and 17 of “logisticRegression.m”.

2. Change the distribution/link functions used on the .m code (it was “logit” for logistic regression; you can change the parameters of functions glmfit and glmval accordingly to each of a, b and c below):

   a. “normal/identity”;
   b. “Poisson/log”;
   c. “gamma/reciprocal”.

   Notice that the binomial data used originally has an additional parameter, n. Therefore, changing the distribution used (which was ‘binomial’) requires several steps:

   - Line 9 needs to be commented out;
   - On line 13, [y n] becomes y;
   - On line 17, ***‘size’, n*** needs to be deleted from *glmval*;
   - On line 18, ***./n*** needs to be deleted from *plot*;

3. Change the data points used in lines 7 to 10 and redo exercise 2, using e.g.:
   a. Linear points (write down similarly spaced x and y values)
   b. Polynomial points (write down values for x and use any polynomial function e.g. y = x^2 + e, where e is a small integer of your choice)
   c. Logarithmic points (write down values for x and use any logarithmic function e.g. y = ln(x) + e, where e is a small integer of your choice)

4. Open and inspect file “logisticRegressionNEW.m” before running it. Would you expect the choice of distribution/link functions to work well? Now run it and check the result.