APPM2720 Lecture 11.1

Prediction for [0,1] responses.

This lecture is about handling a prediction problem where one wants to use a set of variables to predict a binary outcome. For example, given different features in an email message predict if it is spam or not spam. This kind of method would be the basis the basis of building a spam filter for an email system. The statistical model that can be used in this case is termed logistic regression and is a flexible way to relate a binary response to continuous variables.

The **spam.RData** from the **Week11** folder is a data frame **dat** along with a logical varible **train** that divides the data frame into two parts (training an testing). **dat** is data on 4601 emails and **spamMetaData.txt** has information about the variables.

Binary variables, probabilities and the logistic tranformation.

Since binary variables only take on two aspects we can just code them as being **0** and **1**. In particular `dat$spam` is a [0,1] variable where 0 is not spam and 1 is spam. The variable **capLong** appears to be useful in predicting spam and the goal is to find the *probability* that a email is spam given that the value for capLong. Probabilities are always between [0,1] and the logistic is a useful way to convert arbitrary values to this range.

Logistic function: $\phi(u) = \frac{e^u}{1+e^u}$

Converting a linear function to a probability:

probability = $p(X) = \phi(\alpha + \beta X) = \frac{e^{(\alpha+\beta X)}}{1+e^{(\alpha+\beta X)}}$

See Chapter 4 of ISLR for examples why this is better than just

probability $= \alpha + \beta X$

In the logistic regression the interpretation is the probability that one observes a "1" for a given value of X is $P(X) = \phi(\alpha + \beta X)$. The probability of being a "0" is just $1 - P(X)$.

fitting the model

The **glm** function in R can be used to find α and β. It uses the formula syntax also found in **lm** but uses maximum likelihood (ML) to fit these parameters. ML is a more general method of estimating parameters than least squares and can be interpreted as finding parameters that maximize the probability of having observed the data.

To illustrate this concept here is an example where the probability is estimated without any covariates.

Finding the the probability for binomial sample.

The setup is we have data $Y_1, Y_2, Y_3, \ldots, Y_n$ where the probability/likelihood function is

$L(Y, \theta) = \theta$ if $Y = 1$

and

$= 1 - \theta$ if $Y = 0$.

For a given value of θ this is the probability observing the data value, Y. For a given value of the data this is how likely it is given a choice of θ.

Assuming the sample values are independent (unrelated to each other) the likelihood for the sample is the product of the individuals.

$L(Y_1, \theta) \times L(Y_2, \theta) \ldots \times L(Y_n, \theta)$

the data is fixed and θ is found by maximizing this expression. It is easier to maximize the log likelihood and one will get the same answer:

$log(L(Y_1, \theta) + log(L(Y_2, \theta) + \ldots + log(L(Y_n, \theta)$

This is analagous to the sum of squares in least squares fitting and how one figures out what the equivalent idea is for a residual.

The maximum likelihood estimate

The last sum can be simplified as $Mlog(\theta) + (n - M)log(1 - theta)$
where M is the number of data values that are 1. (n-M are equal to zero). Taking derivatives and setting equal to zero we have $M/\theta - (n - M)/(1 - \theta) = 0$

and so $\hat{\theta} = M/n$.

The probability that Y=1 is estimated by the fraction of the sample that is equal to 1.

Logistic regression

Note that in the simple example above it is equivalent to $\beta = 0$ and $\theta = \phi(\alpha)$. With an additional variable one has more parameters one has to use the logistic transform to keep the probabilities between 0 and 1. One still maximizes a sum like the log likelihood above but each term can be different because observation can depend on a different value for the variable.

$$log(L(Y_1, \phi(\alpha + X_1\beta)) \quad + \quad log(L(Y_2, \phi(\alpha + X_2\beta)) \quad + \ldots + \quad log(L(Y_n, \phi(\alpha + X_n\beta))$$