STA613/CBB540: Statistical methods in computational biology

Logistic Regression (1/24/13)

Lecturer: Barbara Engelhardt *Scribe: Dinesh Manandhar*

1 Introduction

Logistic regression is model for regression used in categorical prediction of a dependent variable based on its association with one or more independent (continuous or discrete) predictor variables. The probability of how well the independent predictor variable(s) explain the dependent response variable is calculated using the logistic function, a general sigmoid function whose range is between 0 and 1.

2 Exponential Family

Exponential family represents a general class of distributions on finite dimensional Euclidean spaces parameterized by a finite dimensional parameter vector. The exponential family formulation unifies a number of discrete and continuous distributions used for practical modelling, such as Normal, Poisson, Beta, Binomial, Exponential, Dirichlet, and Gamma distributions. (Some of the distributions that are not in the exponential family include mixture model densities, F-distributions, Cauchy distribution, finite or infinite mixtures of other distributions such as beta-binomial, etc.)

2.1 Density of an exponential family distribution

$$P(x|\eta) = h(x)\exp\{\eta(\theta)^T T(x) - A(\eta)\}. \tag{1}$$

Here, $\eta(\theta)$ represents the *natural parameter* (for most of this discussion we will refer to this parameter simply as η), $T(x)$ is the *sufficient statistic*, $h(x)$ is a *normalizing constant* (which can be thought of as a regularizer), and $A(\eta)$ is the *log partition function*.

2.2 Representing the Bernoulli distribution in the exponential family form

For a Bernoulli distribution, with $x \in \{0, 1\}$ representing either success (1) or failure (0) of a trial and μ representing the probability of a success, $0 \leq \mu \leq 1$, we have,

$$\begin{aligned} P(x|\mu) &= \mu^x(1-\mu)^{(1-x)} \\ &= \exp\{\log(\mu^x(1-\mu)^{(1-x)})\} \\ &= \exp\{x\log\mu) + (1-x)\log(1-\mu)\} \\ &= \exp\left\{(\log\frac{\mu}{1-\mu})x + \log(1-\mu)\right\}. \end{aligned}$$

Comparing the final expression with equation 1, we have,

$$\begin{aligned} \eta &= \log \frac{\mu}{1-\mu} \\ T(x) &= x \\ h(x) &= 1 \\ A(\eta) &= -\log(1-\mu) = \log(1+e^{\eta}), \end{aligned}$$

where the last expression for $A(\eta)$ can be obtained by using the expression for μ (the logistic function) derived below.

3 The logistic function

In the Bernoulli distribution, in the exponential family, note that the logit function (i.e., log odds function) maps the *mean parameter* vector, μ, to the natural parameter, η. The function that maps η to μ is the logistic function, which is the inverse of the logit function as shown below:

$$\begin{aligned} \eta &= \log \frac{\mu}{1-\mu} \\ \Rightarrow \mu &= \frac{1}{1+\exp\{-\eta\}}, \text{ the logistic function.} \end{aligned}$$

4 Logistic regression model

As in linear regression, we have pairs of observed variables $D = \{(x_1, y_1), ..., (x_n, y_n)\}$. Now, our variables $y_i \in \{0, 1\}$, are modeled by a conditional Bernoulli distribution. For a generalized linear model (GLM), we have the following model assumptions:

- observed input x assumed to enter the model via a linear combination: θx,
- the conditional mean μ is represented as a function of θx,
- the response y is characterized by an exponential family distribution with conditional mean μ.

For a GLM, we have two choices:

- choose an exponential family distribution (this is often constrained by the form of the response variable y)
- choose a response function $f : \eta \to \mu$, which maps the natural parameter η to the conditional mean μ. If we set $\eta = \theta^T x$, then this function, for a particular choice of exponential family distribution, is given, and called the *canonical response function.*

For logistic regression, we set our natural parameter $\eta = \theta^T x$. Therefore, for our regression model where the conditional probability is modeled as a Bernoulli distribution, the parameter $\mu = E[Y|X, \theta]$ can be obtained from the logistic function,

$$\mu = \frac{1}{1+\exp\{-\eta\}} = \frac{1}{1+\exp\{-\theta^T x\}}.$$

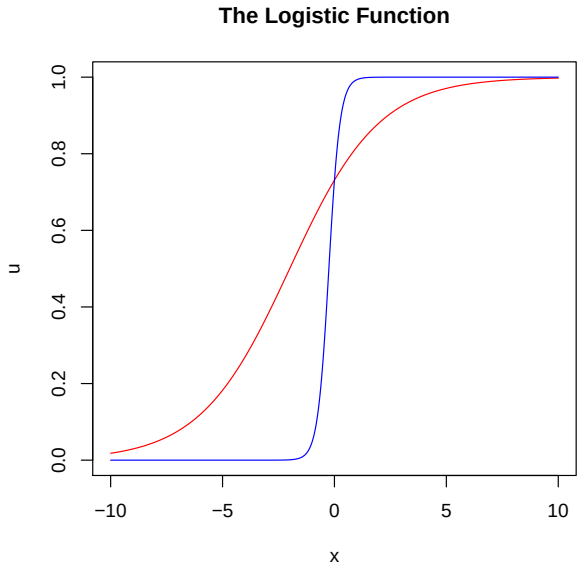

Figure 1: The logistic function $\mu = \frac{1}{1+exp(-\theta^T x)}$, plotted for a range of x values, given two different $\theta = (\theta_\mathbf{0}, \theta_\mathbf{1})^\mathbf{T}$ vectors: $\theta = (1, 0.5)^T$ outputs the slowly increasing red sigmoid curve while $\theta = (1, 4)^T$ outputs the steeper blue curve. For either of the curves, the x-coordinate corresponding to $\mu = 0.5$ is where $\theta_0 = -\theta_1 x$.

The logistic function is thus our canonical response function for logistic regression. Note that the range of a logistic function is (0, 1), i.e. $0 < \mu < 1$, which is what we want in this case.

Figure 1 plots two different logistic functions for two different $\theta = (\theta_0, \theta_1)^T$ values. When $\theta = (1, 0.5)^T$, we get the red plot looking like a sigmoid function, and when $\theta = (1, 4)^T$, we get the steeper (blue) curve. A larger coefficient value for a covariate means that that covariate plays larger role in shaping the regression.

5 Estimating the coefficients of regression (θ)

We have decided that $\eta = \theta^T x$, and thus $\mu = \dfrac{1}{1 + \exp\{-\theta^T x\}}$. Say we have a set of data points, $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. Then we have the likelihood function as

$$P(y_1, y_2, \ldots, y_n | x_1, x_2, \ldots, \theta) = \prod_{i=1}^{n} \mu^{y_i} (1 - \mu)^{1 - y_i}$$

The log-likelihood function is

$$l(D|\theta) = \sum_{i=1}^{n} [y_i \log \mu_i + (1 - y_i) \log(1 - \mu_i)],$$

which means

$$\begin{aligned}\frac{dl}{d\theta} &= \sum_{i=1}^{n}\left(\frac{y_i}{\mu_i} - \frac{1-y_i}{1-\mu_i}\right)\frac{d\mu_i}{d\eta_i}x_i \\ &= \sum_{i=1}^{n}(y_i-\mu_i)x_i, \qquad \text{using the fact that } \frac{d\mu_i}{d\eta_i} = \mu_i(1-\mu_i) \\ &= \mathbf{x^T}(\mathbf{y}-\mu).\end{aligned} \tag{2}$$

We could try to estimate the maximum likelihood estimate (MLE) of θ by setting the derivative of likelihood with respect to θ equal to 0 and solving for θ. However, it turns out that there is no analytic solution for θ. Instead, we can use one of the following gradient-descent type methods to estimate θ:

5.1 Online method

This is a stochastic gradient ascent algorithm, where the θ is estimated using each data point (x_i, y_i) one at a time until it converges. At each iteration of the gradient ascent, the θ is updated as follows:

$$\theta^{(t+1)} = \theta^{(t)} + \rho(y_i - \mu_i^{(t)})x_i,$$

where ρ is a predefined step size. Some drawbacks of this model is that the choice of the step size ρ is arbitrary, and for given a fairly small step size, the time for convergence may be long.

5.2 Iteratively reweighted least squares (IRLS)

Another method for estimating θ uses the Newton-Raphson formula. First, let's calculate the second derivative (also called the Hessian, H) of the log likelihood. We know,

$$\begin{aligned}H &= -\sum_{i=1}^{n}\frac{d\mu_i}{d\eta_i}x_i(x_i)^T, \\ &= -\sum_{i=1}^{n}\mu_i(1-\mu_i)x_i(x_i)^T, \\ &= -\mathbf{x^T}\mathrm{W}\mathbf{x},\end{aligned}$$

where W is an $n \times n$ diag$(\mu_i(1-\mu_i))$ matrix. Since $Y \sim Bern(\mu) \implies Var(Y) = \mu(1-\mu)$, the Hessian weighs each data point by a function of its variance.

Now we can use the Newton-Raphson formula:

$$\theta^{t+1} = \theta^t - H^{-1}\frac{dl}{d\theta},$$

to obtain the value of θ that maximizes the likelihood. Essentially, the Newton-Raphson formula, through an implementation of an iterative-update-process approximates the zero of the function $\frac{dl}{d\theta}$.

$$\begin{aligned}\theta^{(t+1)} &= \theta^t + (\mathbf{x^T}\mathrm{W}\mathbf{x})^{-1}\mathbf{x^T}(\mathbf{y}-\mu^{(\mathbf{t})}) \\ &= (\mathbf{x^T}W^{(t)}\mathbf{x})^{-1}(\mathbf{x^T}W^{(t)}\mathbf{x})\theta^{(t)} + (\mathbf{x^T}W^{(t)}\mathbf{x})^{-1}\mathbf{x^T}(\mathbf{y}-\mu^{(\mathbf{t})}) \\ &= (\mathbf{x^T}W^{(t)}\mathbf{x})^{-1}\mathbf{x^T}W^{(t)}[\mathbf{x}\theta^{(\mathbf{t})} + (W^{(t)})^{-1}(y-\mu^{(t)})]\end{aligned}$$

As can be observed from the first line of the update above, at each iteration the θ is scaled according to the variance of the gradient. Therefore, if the variance (contained in the W matrix) is big, the step size is smaller, and if the variance is small, θ is larger.

6 Type II Diabetes Paper

A genome-wide association study identifies novel risk loci for type 2 diabetes -Sladek et. al., 2007

The paper studies the effects of almost 400 thousand single nucleotide polymorphisms (SNPs) in quest of explaining the heretability of Type 2 diabetes mellitus (T2DM) through a case-control setting experiment. Of primary interest in the study is the discovery of SNPs with statistically significant high odds ratios, which is a measure of the likelihood of association of the particular SNP of interest to T2DM. (Using the idea of odds ratio, the authors did end up finding seven novel SNPs (with odds ratio calculated as high as 1.7) that have high degree of association with T2DM.)

Heritability of a trait is defined as the proportion of observable differences in that trait between individuals within a population that is due to genetic differences. A phenotypic trait is a function of both genetic and environmental factors, and heritability of the trait quantifies the fraction of phenotypic variation explained by the genetic variation. So, if a trait is 100% heritable, then we would not see any difference in the trait between twins raised in two different environments, whereas if the trait was just say, 50% heritable, the genotypic contribution only explains a fraction of the total phenotypic variation between the individuals. Since heritability is a measure of phenotypic variation due to genetic differences, it has to be defined with respect to a population, and phenotypic variation of the trait within that population. As an aside, the heritability of a phenotype, even for traits like height, which are highly heritable, cannot yet be fully explained by a set of SNPs; this is known as "the missing heritability problem".

In the paper, in order to enrich for genetic causes to T2DM, a number of constraints (like the requirement of at least one affected first degree relative, age onset under 45 years, and BMI $< 30kgm^{-2}$) were used to select the participants. The patients used in the study (1363 in total) were in either case group (meaning that they have T2DM) or in the control group (meaning that they have never been diagnosed with T2DM, not that they might not have it any time in the future); so the control group represents the background population distribution of the disease.

Since the authors are interested in the odds ratio for a SNP, they used logistic regression. In this setting, let $Y = \{1, 0\}$ be a binary indicator of a subject having T2DM (represented by 1) or not. Let X represent the number of minor alleles, i.e.$X \in \{0, 1, 2\}$. Then, the $E[Y|X, \theta] = \mu$ which tells us the likelihood of getting T2DM, and the odd ratio is $\dfrac{\mu}{1-\mu}$.

The p-value in this categorical setting of the variables can be obtained by using the *Armitage trend test,* which is explained here in brief.

MAF	0	1	2
Cases			
Control			

Once we have the counts for each phenotype-genotype combination in the table above, a summary statistic and the variance in the statistic is calculated. With a large sample approximation, the variance is assumed to be under normal distribution, and the p-value can be obtained as the probability of seeing the observed variance (and the ones more extreme) under normal distribution.

(Just a note: If we wanted to create and proceed with the model for recessive trait, we can just bin the columns with 0 and 1 minor allele frequencies and proceed as above.)