STA613/CBB540: Statistical methods in computational biology

Bayesian Regression (1/31/13)

Lecturer: Barbara Engelhardt *Scribe: Amanda Lea*

1 Bayesian Paradigm

Bayesian methods ask: given that I have observed data, what is the probability of a latent parameter. Here, the data are fixed and the parameter value we are trying to estimate is unknown. This uncertainty is captured in a probability distribution, known as the posterior probability distribution, describing the distribution over possible values of the estimated parameter given the observed data. In addition, 'beliefs' about the parameter's likely value before observing any data can be incorporated into the model through a distribution known as a *prior*. It is possible to have no prior information about the parameter, and this situation can also be dealt with (for example, by using an uninformative prior).

Formally, relationship is represented as the following:

$$P(\theta|\mathbf{D}) \propto P(\theta)P(\mathbf{D}|\theta)$$

where θ is the model parameters, $\mathbf{D}$ is the observed data, $P(\theta)$ is the prior distribution on the model parameters, $P(\mathbf{D}|\theta)$ is the likelihood of the data, and $P(\theta|\mathbf{D})$ is the posterior probability.

In the Bayesian setting, we are trying to maximize the posterior probability (i.e., maximize the probability of our estimated parameter), instead of the likelihood.

$$\mathbf{argmax}_\theta[P(\theta|\mathbf{D})] \propto \mathbf{argmax}_\theta[P(\theta)P(\mathbf{D}|\theta)]$$

Bayesian methods have a number of advantages with respect to traditional frequentist methods:

- the Bayesian paradigm handles small amounts of data well
- outliers that would otherwise skew results do not contribute as much to parameter estimates (if they are not likely under the prior distribution)
- beliefs about the parameters being estimated can be incorporated into the model explicitly
- the data, rather than the parameters, are thought of as fixed. Parameters are unknown random variables, and described probabilistically.

2 Bayesian regression

Let's work toward a description of regression in the Bayesian paradigm.

$$Y|X,\beta \sim N(X^T\beta, \sigma^2\mathbf{I})$$

$$Y|X,\beta \propto (\sigma^2)^{-n/2}\exp\left\{\frac{-1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)\right\}$$

For the Bayesian approach, we need to put a prior probability on our parameters β and σ^2.

$$P(\beta,\sigma^2) = p(\beta|\sigma^2)p(\sigma^2)$$

We can choose to set any prior distribution on our parameters, but certain choices will make the posterior probability distribution easier to work with (and will give a closed form solution for the *maximum a posteriori* (MAP) estimate of the parameters). We can return to the exponential family form equations to help us think about appropriate priors.

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

We can write our prior probability $(\eta|\lambda)$ in terms of the exponential family. Here, $(\eta|\lambda)$ is the prior distribution of the (natural) parameters for our data, conditioned on prior parameter λ.

$$\begin{aligned}
x|\eta &\sim P(x|\eta)\\
\eta|\lambda &\sim F(\eta|\lambda)\\
F(\eta|\lambda) &= h_c(\eta)\exp\{\lambda_1^T\eta - \lambda_2^T(A(\eta)-A_c(\lambda))\}\\
P(\eta|x_1...x_n,\lambda) &\propto F(\eta|\lambda)\Pi_{i=1}^n P(x_i|\eta)\\
P(\eta|x_1...x_n,\lambda) &\propto h_c(\eta)\exp\left\{\left[\lambda_1 + \sum_{i=1}^n T(x_i)\right]^T\eta + (\lambda_2+n)(-A(\eta))\right\}
\end{aligned}$$

In the posterior, the sufficient statistics are $T_p(x) = \begin{bmatrix} \eta \\ -A(\eta) \end{bmatrix}$

When the posterior distribution and the prior probability distribution are in the same family, then the prior distribution is called the conjugate prior for the likelihood distribution. Every member of the exponential family has a conjugate prior. For example, the Beta distribution is the conjugate prior of the Bernoulli distribution and the Gaussian distribution is conjugate to itself in the mean parameter. Choosing the appropriate conjugate prior distribution ensures that your posterior has a predictable distribution (in terms of the exponential family) and the MAP parameter estimates will have a closed form solution.

Here, we will choose a Gaussian distribution for the prior on β and an inverse gamma distribution for the prior on σ^2.

$$\begin{aligned} P(\beta|\sigma^2) &\propto (\sigma^2)^{-k/2}\exp\{\frac{-1}{2\sigma^2}(\beta-\mu_0)^T\Lambda_0(\beta-\mu_0)\} \\ &\sim N(\mu_0,\sigma^2\Lambda_0) \\ P(\sigma^2|a,b) &\sim (\sigma^2)^{-a-1}\exp\left\{\frac{-b}{\sigma^2}\right\} \end{aligned}$$

Where μ_0 and Λ_0 are parameters for the Gaussian distribution on β, and a and b are the shape and scale parameters of the inverse gamma distribution. We can write the posterior distribution together as:

$$P(\beta,\sigma^2|Y,X) \propto P(Y|X,\beta,\sigma^2)P(\beta|\sigma^2)P(\sigma^2)$$

Where $P(Y|X,\beta,\sigma^2)$ is the likelihood and $P(\beta|\sigma^2)P(\sigma^2)$ is the prior. In this equation, X is a $k \times n$ matrix and y is an $n \times k$ matrix. Here, we are only working with one dimension of X. Now we will substitute the above probability distribution equations into the full equation:

$$P(\beta,\sigma^2|Y,X) \propto$$

$$(\sigma^2)^{-n/2}\exp\left\{\frac{-1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)\right\} * (\sigma^2)^{-k/2}\exp\left\{\frac{-1}{2\sigma^2}(\beta-\mu_0)^T\Lambda_0(\beta-\mu_0)\right\} * (\sigma^2)^{-a-1}\exp\left\{\frac{-b}{\sigma^2}\right\}$$

We can further simply to get the posterior (written below as an expectation). This form of Bayesian regression is known as ridge regression.

$$E[\beta|\sigma^2,Y,X] = (X^TX+\Lambda_0)^{-1}(X^TY+\Lambda_0\mu_0)$$

This is very similar to the normal equations we encountered in linear regression.

$$\beta = (X^TX)^{-1}(X^TY)$$

However, in this form of Bayesian regression, also known as *ridge regression*, we add a covariance matrix, Λ_0, that helps make X^TX invertible if it is singular. Usually, $\mu_0 = 0$ and $\Lambda_0 = \mathbf{cI}$, where $\mathbf{c}$ is a constant and $\mathbf{I}$ is the identity matrix. Note that in the second term in this equation, the impact of the prior parameters when $n \to \infty$ (or the size of the data grows to infinity) is reduced to zero.

3 Stephens and Balding 2009 - Bayesian statistical methods for genetic association studies

In this paper, the authors compare Bayesian regression and testing paradigms in the context of association studies and identifying quantitative trait loci. Bayes factors, similar to likelihood ratios, describe the tradeoff between the alternative and the null model.

What are the features of Bayes factors as compared to p-values?

- P values are affected by power and are therefore not comparable across studies with different sample sizes

- there are many implicit assumptions when using P-values; BFs makes many of those assumptions explicit
- p-values cannot be combined (in an inutitive, easily worked out way)
- multiple hypothesis testing is a problem when using P-values, as compared to Bayes factors.

Let's look at an example of the steps for computing Bayes factors and testing hypotheses in a Bayesian framework. For correlating genotype with phenotype, we normally use the following null hypothesis (that a SNP genotype is uncorrelated with phenotype) and alternative (that a SNP genotype is not uncorrelated with phenotype) hypotheses:

$$H_0 : \beta = 0$$
$$H_a : \beta \neq 0$$

Using a Bayesian approach, we would set the posterior probability of association according to the null or alternate hypothesis:

$$P(\beta \neq 0|X, Y)$$

Then, we would chose our prior odds. For this example (and many genetic association examples), the prior odds indicate how likely it is that a given SNP will be associated with the phenotype of interest. This value may vary across SNPs, perhaps depending on the minor allele frequency. Here, we will choose 10^{-6}.

$$\Pi : \frac{P(H_a)}{P(H_0)} \approx 10^{-6}$$

Now, we compute the Bayes factor. This allows us to quantify the ratio of the observed data likelihood under the alternative model H_a versus the null model H_0. For a Bayes factor less than or equal to 1, we cannot reject the null hypothesis. For a Bayes factor greater than 1, we may reject the null hypothesis (depending on our threshold for significance/rejection of the null). Larger Bayes factors indicate stronger support that the data fit H_a rather than H_0. The form of a Bayes factor looks similar to a likelihood ratio test, but in this case, the likelihoods need not come from the same model (with different numbers of parameters).

$$BF = \frac{P(D|H_a)}{P(D|H_0)}$$

Next, we compute the posterior odds.

$$PO = BF * \Pi$$
$$PO = \frac{P(D|H_a)}{P(D|H_0)} * \frac{P(H_a)}{P(H_0)}$$

And from there, we can compute the posterior probability of association (PPA), where values closer to one indicate a higher chance of association:

$$PPA = \frac{PO}{1 + PO}$$

The posterior probability of association can be interpreted directly as a probability, irrespective of power or sample size. PPAs for a list of SNPs can be ranked and used in an FDR-type framework to chose a threshold for significance.