Structure Learning: Parameter Estimation II

Bayesian Inference

- The MLE is a frequentist inference method. There is another approach to inference called Bayesian inference.
- The key differences between frequentist and Bayesian approaches are shown in the next slides
- See "A primer on Bayesian statistics in Health Economics and Outcomes research" by Anthony O'Hagan and Bryan R. Luce

Bayesian Inference

The Nature of Probability

Frequentist	**Bayesian**
Probability is a limiting, long-run frequency	Probability measures a personal degree of belief
It only applies to events that are (at least in principle) repeatable	It applies to any event or proposition about which we are uncertain

Bayesian Inference

The Nature of Parameters

Frequentist	**Bayesian**
Parameters are not repeatable or random	Parameters are unknown
They are therefore not random variables, but fixed (unknown) quantities	They are therefore random variables

Bayesian Inference

The Nature of Inference

Frequentist	Bayesian
Does not (although it appears to) make statements about parameters	Makes direct probability statements about parameters
Interpreted in terms of long-run repetition	Interpreted in terms of evidence from the observed data

Bayesian inference

Bayesian inference:

1. Choose probability density $f(\theta)$ – called the prior distribution that expresses our beliefs about a parameter θ before we see any data.
2. We choose a statistical model $f(x|\theta)$
3. After observing data $X_1, \ldots, X_n$, we update our beliefs and calculate the posterior distribution $f(\theta|X_1, \ldots, X_n)$

Bayesian Inference

Suppose we have n independent, identically distributed observations $X_1, \ldots, X_n$. The joint density of the data is:

$$f(x_1,\ldots,x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta) = L_n(\theta)$$

$$f(\theta \mid x_1,\ldots,x_n) = \frac{f(x_1,\ldots,x_n \mid \theta) f(\theta)}{f(x_1,\ldots,x_n)} = \frac{f(x_1,\ldots,x_n \mid \theta) f(\theta)}{\int f(x_1,\ldots,x_n \mid \theta) f(\theta) d\theta}$$

$$= \frac{L_n(\theta) f(\theta)}{\int L_n(\theta) f(\theta) d\theta} = \alpha L_n(\theta) f(\theta)$$

$$\therefore\ f(\theta \mid x_1,\ldots,x_n) \propto L_n(\theta) f(\theta)$$

Posterior Distribution

Likelihood

Prior (Note: We are not committing to a particular θ but using the entire distribution of θ)

Bayesian Inference

What do you do with the posterior distribution?

- Use the entire distribution (can be clumsy sometimes)
- Get a point estimate by summarizing the center of the posterior – use the mean or mode
- The posterior mean is:

$$\bar{\theta}_n = E[\theta] = \int \theta f(\theta \mid x_1,\ldots,x_n) d\theta = \frac{\int \theta L_n(\theta) f(\theta)}{\int L_n(\theta) f(\theta) d\theta}$$

Conjugate Priors

Let's redo the first candy example except this time, we will put a $Beta(\alpha, \beta)$ prior on θ. Recall that θ is the probability a candy will be cherry flavored. The posterior has the form:

$$f(\theta|x_1, \dots, x_n) = \frac{f(\theta)L_n(\theta)}{\int f(\theta)L_n(\theta)d\theta}$$

$$f(\theta) = Beta(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

where $\Gamma(z) = (z-1)!$

9

Conjugate Priors

$$f(\theta \mid x_1, \dots, x_n) = \frac{f(\theta)L_n(\theta)}{\int f(\theta)L_n(\theta)d\theta}$$

$$= \frac{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{c}(1-\theta)^{l}}{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int \theta^{c}(1-\theta)^{l}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta}$$

$$= \frac{\theta^{c}(1-\theta)^{l}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int \theta^{c}(1-\theta)^{l}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta} = \frac{\theta^{c+\alpha-1}(1-\theta)^{l+\beta-1}}{\int \theta^{c+\alpha-1}(1-\theta)^{l+\beta-1}d\theta}$$

Let's take a look at this term in the denominator

10

Conjugate Priors

Below is the Beta distribution with alpha parameter = c + α and beta parameter = l + β. Since it is a known pdf, it will integrate to 1.

$$\int Beta(c+\alpha, l+\beta)\, d\theta = \int \frac{\Gamma(c+\alpha+l+\beta)}{\Gamma(c+\alpha)\Gamma(l+\beta)}\theta^{c+\alpha-1}(1-\theta)^{l+\beta-1}d\theta = 1$$

This is the term in the denominator from the previous page. It is almost a Beta distribution except it is missing the normalization constant in front.

$$\int \theta^{c+\alpha-1}(1-\theta)^{l+\beta-1}d\theta$$

Let's call the normalization constant (the expression with the Gammas) c. The expression above becomes:

$$\int \theta^{c+\alpha-1}(1-\theta)^{l+\beta-1}d\theta = \frac{1}{c}\int c\theta^{c+\alpha-1}(1-\theta)^{l+\beta-1}d\theta = \frac{1}{c}$$

11

Conjugate Priors

Continuing from where we left off…

$$f(\theta \mid x_1, \dots, x_n) = \frac{\theta^{c+\alpha-1}(1-\theta)^{l+\beta-1}}{\int \theta^{c+\alpha-1}(1-\theta)^{l+\beta-1}d\theta}$$

$$= \frac{\theta^{c+\alpha-1}(1-\theta)^{l+\beta-1}}{\frac{\Gamma(c+\alpha)\Gamma(l+\beta)}{\Gamma(c+\alpha+l+\beta)}} = \frac{\Gamma(c+\alpha+l+\beta)}{\Gamma(c+\alpha)\Gamma(l+\beta)}\theta^{c+\alpha-1}(1-\theta)^{l+\beta-1}$$

$$= Beta(c+\alpha, l+\beta)$$

Conjugate Priors

- A conjugate prior is a family of prior probability distributions with the property that the posterior also belongs to that family.
- eg. the conjugate prior for a Bernoulli is a Beta distribution
- Other useful conjugate priors:

Likelihood	Conjugate Prior	Posterior
Normal	Normal	Normal
Binomial	Beta	Beta
Poisson	Gamma	Gamma
Multinomial	Dirichlet	Dirichlet

13

Conjugate Priors

Why are they useful?

- Since we know the form of the posterior, we can easily calculate statistics such as the mean.
- For example, we know:

$$E[Beta(\alpha,\beta)] = \frac{\alpha}{\alpha+\beta}$$

- Thus, the mean for the candy example above is:

$$E[Beta(c+\alpha, l+\beta)] = \frac{\alpha+c}{\alpha+\beta+l+c}$$

14

Conjugate Priors

- You can think of α and β in the posterior distribution as "virtual counts"
- eg. Using a uniform prior Beta(1,1), the mean of the posterior becomes:

$$E[Beta(c+1, l+1)] = \frac{\alpha+c}{\alpha+\beta+l+c} = \frac{1+c}{2+l+c}$$

15

Conjugate Priors

$$E[Beta(c+1, l+1)] = \frac{\alpha+c}{\alpha+\beta+l+c} = \frac{1+c}{2+l+c}$$

- If we observe no data, ie. *c=0*, *l=0*, the posterior mean is ½, which is what we would expect since we have to pick between the two flavors of lime and cherry
- If we observe lots of data, then the *c* term in the numerator and the *l+c* term in the denominator dominate the prior

16

Conjugate Priors

- The conjugate prior that is of most relevance to parameter estimation is the Multinomial-Dirichlet
- Recall that a Dirichlet distribution is a generalization of a Beta distribution
- And a Multinomial distribution is a generalization of a Binomial distribution
- If a node in a Bayesian network can take 2 values, the analysis is just like the Beta-Binomial example in previous slides
- If it takes more than 2 values, then you have to use a Multinomial-Dirichlet

Conjugate Priors

Multinomial

$$f(x_1,\ldots,x_k \mid n, p_1,\ldots,p_k) = \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

for $\Sigma x_i = n$, $p_i \,\varepsilon\, [0,1]$, $\Sigma p_i = 1$

Note: The parameters $p_1, \ldots, p_k$ from the multinomial are now the random variables in the Dirichlet prior

Dirichlet

$$f(p_1,\ldots,p_k \mid \alpha_1,\ldots,\alpha_k) = \frac{\Gamma(\alpha_1+\ldots+\alpha_k)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)} p_1^{\alpha_1-1} \cdots p_k^{\alpha_k-1}$$

for $p_i \geq 0$, $\Sigma\, p_i = 1$

Conjugate Priors

Likelihood	Conjugate Prior	Posterior
Binomial(x | n, p)	Beta(α, β)	Beta(x+α, n-x+β)
Multinomial(x_1,...,x_k | n, p_1, ..., p_k)	Dirichlet(p_1, ..., p_k | α_1, ..., α_k)	Dirichlet(x_1+ α_1,...,x_k + α_k)

For Beta-Binomial posterior: $E[p] = \frac{x+\alpha}{n+\alpha+\beta}$

For Dirichlet-Multinomial posterior: $E[p_i] = \frac{x_i+\alpha_i}{n+\sum_j \alpha_j}$

Conjugate Priors

Suppose you were asked to estimate *P(Price =Low | Type = Sedan, Color = Silver)*.

Notice that this distribution is a multinomial distribution with n = 2 (because there are 2 records with Color=Silver, Type=Sedan) and p_{low}, p_{medium}, p_{high} corresponding to when *Price* is low, medium, and high.

Now suppose I tell you to use a Dirichlet prior where all the α_i are 1.

Color	Type	Price
Silver	Sedan	Low
Black	Sedan	Medium
Silver	Pickup	High
Silver	Sedan	Low
Red	SUV	High

Estimate P(Price = Low | Color = Silver, Type = Sedan)

$$= \frac{\#(Color = Silver \text{ AND } Type = Sedan \text{ AND } Price = Low)+1}{\#(Color = Silver \text{ AND } Type = Sedan)+3}$$

$$= \frac{2+1}{2+3} = \frac{3}{5}$$