Bayesian Graphical Models

Aarhus University, Fall 2003,
Lectures 9 and 10

Steffen L. Lauritzen, Aalborg University

Bayesian inference

Parameter θ, data $X = x$, likelihood

$$L(\theta \mid x) \propto p(x \mid \theta) = \frac{dP_\theta(x)}{d\mu(x)}.$$

Express knowledge about θ through *prior distribution* π on θ. Use also π to denote density of prior w.r.t. some measure ν.

Inference about θ from x is then represented through *posterior distribution* $\pi^*(\theta) = p(\theta \mid x)$. Then, from Bayes' formula

$$\pi^*(\theta) = p(x \mid \theta)\pi(\theta)/p(x) \propto L(\theta \mid x)\pi(\theta)$$

so the *likelihood function is equal to the density of the posterior w.r.t. the prior* modulo a constant.

Bayesian graphical models

Represent statistical models as *Bayesian networks with parameters included as nodes,* i.e. for expressions as

$$p(x_v \,|\, x_{\mathrm{pa}(v)}, \theta_v)$$

include θ_v as additional parent of v.

Then Bayesian inference about θ can in principle be calculated by probability propagation as in general Bayesian networks.

This is true for θ_v discrete.

For θ continuous, we must develop other computational techniques.

Bernoulli experiments

Data $X_1 = x_1, \ldots, X_n = x_n$ independent and Bernoulli distributed with parameter θ, i.e.

$$P(X_i = 1 \,|\, \theta) = 1 - P(X_i = 0) = \theta.$$

Represent as a Bayesian network with θ as only parent to all nodes $x_i, i = 1, \ldots, n$. Use a beta prior:

$$\pi(\theta \,|\, a, b) \propto \theta^{a-1}(1-\theta)^{b-1}.$$

If we let $x = \sum x_i$, we get the posterior:

$$\begin{aligned} \pi^*(\theta) &\propto \theta^x(1-\theta)^{n-x}\theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{x+a-1}(1-\theta)^{n-x+b-1} \end{aligned}$$

So the posterior is also beta with parameters $(a + x, b + n - x)$.

Conjugate families

A family $\mathcal{P}$ of distributions on Θ is said to be *conjugate* under sampling from x if

$$\pi \in \mathcal{P} \implies \pi^* \in \mathcal{P}.$$

The family of beta distributions is conjugate under Bernoulli sampling.

If the family of priors is parametrised:

$$\mathcal{P} = \{P_\alpha, \alpha \in \mathcal{A}\}$$

we sometimes say that α is a *hyperparameter.* Then, Bayesian inference can be made by just updating hyperparameters. Terminology of hyperparameter breaks down in complex models.

Conjugate exponential families

For a k-dimensional exponential family

$$p(x \mid \theta) = b(x) e^{\theta^\top t(x) - \psi(\theta)}$$

the *standard conjugate family* is given as

$$\pi(\theta \mid a, \kappa) \propto e^{\theta^\top a - \kappa \psi(\theta)}$$

for $(a, \kappa) \in \mathcal{A} \subseteq \mathcal{R}^k \times \mathcal{R}_+$, where $\mathcal{A}$ is determined so that the normalisation constant is finite.

Posterior updating from $(x_1, \ldots, x_n)$ with $t = \sum_i t(x_i)$ is then made as $(a^*, \kappa^*) = (a + t, \kappa + n)$.

The family of Beta distributions is a standard conjugate family.

Markov chain Monte Carlo

When exact computation is infeasible, Markov chain Monte Carlo (MCMC) methods are used.

An MCMC method for the *target distribution* π^* on $\mathcal{X} = \mathcal{X}_V$ constructs a Markov chain $X^0, X^1, \ldots, X^k, \ldots$ with π^* as *equilibrium distribution.*

For the method to be useful, π^* must be the *unique* equilibrium, and the Markov chain must be *ergodic* so that for all relevant A

$$\pi^*(A) = \lim_{n\to\infty} \pi^*_n(A) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=m+1}^{m+n} \chi_A(X^i)$$

where χ_A is the indicator function of the set A.

Geometric ergodicity

Using simulations from Markov chain constructed, we estimate expectations as averages:

$$\bar{g} = \int_{\mathcal{X}} g(x)\, d\pi^*(x) \approx \bar{g}_n = \frac{1}{n} \sum_{i=m+1}^{m+n} g(x^i).$$

The values $x^0, \dots, x^m$ are discarded and m is referred to as length of the *burn-in period*.

If the Markov chain is *geometrically ergodic*, i.e.

$$||\pi^* - \mathcal{L}(X^n \,|\, x^0)||_{\mathsf{totvar}} \leq c(x^0)\psi^n \text{ for some } \psi < 1$$

and $\int g^2 \, d\pi^* < \infty$, there is also a central limit theorem so

$$\bar{g}_n \overset{\mathrm{a}}{\sim} \mathcal{N}(\bar{g}, \sigma_g^2/n).$$

The standard Gibbs sampler

A simple MCMC method is made as follows.

1. Enumerate $V = \{1, 2, \ldots, |V|\}$
2. choose starting value $x^0 = x_1^0, \ldots, x_{|V|}^0$.
3. Update now x^0 to x^1 by replacing x_i^0 with x_i^1 for $i = 1, \ldots, |V|$, where x_i^1 is chosen from 'the full conditionals'

$$\pi^*(X_i \,|\, x_1^1, \ldots, x_{i-1}^1, x_{i+1}^0, \ldots x_{|V|}^0).$$

4. Continue similarly to update x^k to x^{k+1} and so on.

Properties of Gibbs sampler

With positive joint target density $\pi^*(x) > 0$, the Gibbs sampler Markov chain is ergodic with π^* as the unique equilibrium distribution.

In this case the distribution of $X(n)$ converges to π^* for n tending to infinity.

Geometric ergodicity is not generally satisfied and a generally applicable condition for this to hold is not known (to me at least).

Full conditional distributions

For a directed graphical model, the density of full conditional distributions are:

$$\begin{aligned} f(x_i \mid x_{V\setminus i}) &\propto \prod_{v\in V} f(x_v \mid x_{\mathrm{pa}(v)}) \\ &\propto f(x_i \mid x_{\mathrm{pa}(i)}) \prod_{v\in \mathrm{ch}(i)} f(x_v \mid x_{\mathrm{pa}(v)}) \\ &= f(x_i \mid x_{\mathrm{bl}(i)}), \end{aligned}$$

x where $\mathrm{bl}(i)$ is the *Markov blanket* of node i:

$$\mathrm{bl}(i) = \mathrm{pa}(i) \cup \left\{\cup_{v\in \mathrm{ch}(i)} \mathrm{pa}(v) \setminus \{i\}\right\} = \mathrm{ne}^m(i)$$

where $\mathrm{ne}^m(i)$ are the neighbours of i in the moral graph.

Envelope sampling

In many cases, the conditional distributions further simplify (by local conjugacy). If not, there are many ways of sampling from a general density $f(x)$ which is known up to a proportionality factor, i.e. $f(x) \propto h(x)$.

One is using an *envelope* $g(x) \geq h(x)$, where g(x) is a known density and then performing rejection sampling as follows:

1. Choose $X = x$ from distribution with density g
2. Choose $U = u$ uniform on the unit interval.
3. If $u > g(x)/h(x)$, then reject x and repeat step 1, else return x.

Metropolis within Gibbs

If no envelope is known, an alternative is to use one step of a Metropolis–Hastings sampler.

Here g is known density, $f \propto h$ and x is a current value (of x_i during the Gibbs updating).x

1. Choose $Y = y$ from distribution with density g
2. Choose $U = u$ uniform on the unit interval.
3. If $u > \min\{1, \frac{g(x)h(y)}{g(y)h(x)}\}$, then keep x, else replace x with y.

Note that here g only needs to be known up to a constant factor.