Stat 535 C - Statistical Computing & Monte Carlo Methods

Arnaud Doucet

Email: arnaud@cs.ubc.ca

25th January 2007

- Suggested Projects:

`www.cs.ubc.ca/~arnaud/projects.html`

- Assignement on the web: please give it back before the break.

2.1– Outline

- Bayesian Statistics.
- Testing Hypotheses: The Bayesian way.
- Bayesian Model Selection.

2.2– Simple Binomial example

• (Bayes, 1764): A billiard ball W is rolled on a line of length one, with a uniform probability of stopping anywhere. It stops at $\theta = p$. A second ball O is then rolled n times under the same assumptions and X denotes the number of times the ball O stopped on the left of W. Given X, what inference can we make on p?

• We select $\theta \sim \mathcal{U}[0,1]$ and $X|\theta \sim \mathcal{B}(n,\theta)$ binomial distribution

$$\Pr(X = x|\theta) = \begin{pmatrix} n \\ x \end{pmatrix} \theta^x (1-\theta)^{n-x} \Rightarrow \pi(\theta|x) = \frac{\begin{pmatrix} n \\ x \end{pmatrix} \theta^x (1-\theta)^{n-x} 1_{[0,1]}(\theta)}{\int_0^1 \begin{pmatrix} n \\ x \end{pmatrix} \theta^x (1-\theta)^{n-x} d\theta}$$

- We have

$$\pi\left(x\right)=\int_{0}^{1}\Pr\left(\left.X=x\right|\theta\right)\pi\left(\theta\right)d\theta=\frac{1}{n+1}\text{ for }x=0,...,n$$

- It follows that $\pi\left(\left.\theta\right|x\right)=\mathcal{B}e\left(x+1,n+1-x\right)$ where for $\alpha,\beta>0$

$$\mathcal{B}e\left(\theta;\alpha,\beta\right)=\frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}\theta^{\alpha-1}\left(1-\theta\right)^{\beta-1}1_{[0,1]}\left(\theta\right)$$

- *Prediction.* Given $X = x$, the distribution of the outcome Y is given by

$$
\begin{aligned}
\Pr\left(Y = left \middle| x\right) &= \int \Pr\left(Y = left \middle| \theta, x\right) \pi\left(\theta \middle| x\right) d\theta \\
&= \int \Pr\left(Y = left \middle| \theta\right) \pi\left(\theta \middle| x\right) d\theta \\
&= \int \theta \pi\left(\theta \middle| x\right) d\theta = \mathbb{E}\left[\theta \middle| x\right] = \frac{x+1}{n+2}.
\end{aligned}
$$

2.2– Simple Binomial example

• *Application.* Laplace developed independently such a model. From 1745 to 1770, x =241,945 girls and $n-x$ =251,527 boys were born in Paris. Let θ be the probability that any birth is female, then $n = 251,527 + 241,945$

$$\Pr\left(\theta \leq 0.5 \middle| x = 241,945\right) = 1 - 1.15 \times 10^{-42} \approx 1.$$

• *Remark*: This is completely different from a p-value. We do not integrate over observations we have never seen.

2.3– A Simple Poisson Model

- Assume you have some couting observations $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{P}(\theta)$; i.e.

$$f(x_i|\theta) = e^{-\theta}\frac{\theta^{x_i}}{x_i!}$$

- Assume we adopt a Gamma prior for θ; i.e. $\theta \sim \mathcal{G}a(\alpha, \beta)$

$$\pi(\theta) = \mathcal{G}a(\theta; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}.$$

- We have

$$\pi(\theta| x_1, ..., x_n) = \mathcal{G}a\left(\theta; \alpha + \sum_{i=1}^{n} x_i, \beta + n\right).$$

2.4– Simple Gaussian example

- Consider $X_1|\,\theta \sim \mathcal{N}\left(\theta, \sigma^2\right)$ and $\theta \sim \mathcal{N}\left(m_0, \sigma_0^2\right)$

$$\begin{aligned}
\pi\left(\theta|\, x_1\right) &\propto f\left(x_1|\,\theta\right)\pi\left(\theta\right) \propto \exp\left(-\frac{\left(x_1-\theta\right)^2}{2\sigma^2} - \frac{\left(\theta - m_0\right)^2}{2\sigma_0^2}\right) \\
&\propto \exp\left(-\frac{\theta^2}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}\right) + \theta\left(\frac{x_1}{\sigma^2}+\frac{m}{\sigma^2}\right)\right) \\
&\propto \exp\left(-\frac{1}{2\sigma_1^2}\left(\theta - m_1\right)^2\right)
\end{aligned}$$

$$\Rightarrow \theta|\, x_1 \sim \mathcal{N}\left(m_1, \sigma_1^2\right)$$

$$\begin{aligned}
\text{with } \frac{1}{\sigma_1^2} &= \frac{1}{\sigma_0^2}+\frac{1}{\sigma^2} \Rightarrow \sigma_1^2 = \frac{\sigma_0^2\sigma^2}{\sigma_0^2+\sigma^2}, \\
m_1 &= \sigma_1^2\left(\frac{x_1}{\sigma^2}+\frac{m}{\sigma_0^2}\right).
\end{aligned}$$

2.4– Simple Gaussian example

- To predict the distribution of a new observation $X|\,\theta \sim \mathcal{N}\left(\theta, \sigma^2\right)$ in light of x_1 we use the predictive distribution

$$f\left(x|\, x_1\right) = \int f\left(x|\, \theta\right) \pi\left(\theta|\, x_1\right) d\theta$$

We can do direct calculations or alternatively use the fact that $f\left(x|\, x_1\right)$ is Gaussian so characterized by its mean and variance

$$\begin{aligned} \mathbb{E}\left[X|\, x_1\right] &= \mathbb{E}\left[\theta + V|\, x_1\right] = \mathbb{E}\left[\theta|\, x_1\right] = m_1, \\ var\left[X|\, x_1\right] &= var\left[\theta + V|\, x_1\right] = var\left[\theta|\, x_1\right] + var\left[V\right] = \sigma_1^2 + \sigma^2. \end{aligned}$$

2.4– Simple Gaussian example

• Now assume that you observe a realization x_2 of $X_2|\,\theta \sim \mathcal{N}\left(\theta, \sigma^2\right)$. Then you are interested now in

$$\begin{aligned}\pi\left(\theta|\,x_1, x_2\right) &\propto f\left(x_2|\,\theta\right) f\left(x_1|\,\theta\right) \pi\left(\theta\right) \\ &\propto f\left(x_2|\,\theta\right) \pi\left(\theta|\,x_1\right) \\ &\propto f\left(x_1|\,\theta\right) \pi\left(\theta|\,x_2\right).\end{aligned}$$

• Updating the prior one observation at a time, or all observations together, does not matter.

• The sequential approach can be useful for massive dataset. In this case at time n

$$\pi\left(\theta|\,x_1, ..., x_n\right) \propto f\left(x_n|\,\theta\right) \pi\left(\theta|\,x_1, ..., x_{n-1}\right);$$

i.e. ‘the prior at time n is the posterior at time $n-1$’.

2.5– Simple Gaussian example: Bayes vs ML

- ML estimate of θ at time n is simply

$$\theta_{ML} = \arg\sup_{\theta} \prod_{i=1}^{n} f\left(x_i \middle| \theta \right) = \frac{1}{n} \sum_{i=1}^{n} x_i .$$

- Posterior of θ at time n is

$$\theta | \, x_1, ..., x_n \sim \mathcal{N}\left(m_n, \sigma_n^2 \right)$$

where

$$\begin{aligned} \frac{1}{\sigma_n^2} &= \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \Rightarrow \sigma_n^2 = \frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2} \underset{n\to\infty}{\sim} \frac{\sigma^2}{n}, \\ m_n &= \sigma_n^2 \left(\frac{\sum_{i=1}^{n} x_i}{\sigma^2} + \frac{m}{\sigma_0^2} \right) \underset{n\to\infty}{\sim} \frac{\sum_{i=1}^{n} x_i}{n}. \end{aligned}$$

- Asymptotically in n the prior is washed out by the data and $E\left[\theta | \, x_1, ..., x_n \right] = m_n \approx \theta_{ML}$.

2.6– Bayes vs ML

• However, keep in mind that information provided by a Bayesian approach is much richer.

• You can compute for example posterior probabilities

$$\Pr\left(\theta \in A \right| x_1, ..., x_n) \text{ or } var\left(\theta \right| x_1, ..., x_n)$$

or compute the distributions of future observations

$$f\left(x \right| x_1, ..., x_n).$$

• ML can be reassuring because of consistency and efficiency.
For finite sample sizes, do you really care?
For time series models for example, there is no such thing.

3.1– Testing hypotheses in a Bayesian framework

• Consider the problem where we have $\theta \sim \mathcal{U}[0,1]$ and

$$\Pr(X = x|\theta) = \begin{pmatrix} n \\ \\ x \end{pmatrix} \theta^x (1-\theta)^{n-x} \text{ then } \pi(\theta|x) = \mathcal{B}e(x+1, n+1-x).$$

• If we want to test $H_0 : \theta \geq \frac{1}{2}$ vs $H_1 : \theta < \frac{1}{2}$ then, in a Bayesian approach, you can simply compute

$$\pi(H_0|x) = 1 - \pi(H_1|x) = \int_{1/2}^{1} \pi(\theta|x)\, d\theta.$$

• Golden rule of Bayesians: **Thou shalt not integrate with respect to observations** (except for design...)
$\Rightarrow$ Contrary to frequentists, your test is never based on observations you do not observe.

• More generally, ones wants to compare two hypothesis: $H_0 : \theta \sim \pi_0$ versus $H_1 : \theta \sim \pi_1$. You can think of it as putting a prior given by

$$\begin{aligned} \pi(\theta) &= \pi(H_0)\,\pi(\theta|\,H_0) + \pi(H_1)\,\pi(\theta|\,H_1) \\ &= \pi(H_0)\,\pi_0(\theta) + \pi(H_1)\,\pi_1(\theta) \end{aligned}$$

where $\pi(H_0) + \pi(H_1) = 1$.

• In the previous example, $\pi_0(\theta) = \mathcal{U}\left(\theta; \left[\frac{1}{2}, 1\right]\right)$ and $\pi_1(\theta) = \mathcal{U}\left(\theta; \left[0, \frac{1}{2}\right)\right)$ and $\pi(H_0) = \pi(H_1) = \frac{1}{2}$.

• To compare H_0 versus H_1, we typically compute the *Bayes factor* which partially eliminated the influence of the prior modelling (i.e. $\pi(H_i)$!)

$$
\begin{aligned}
B_{10}^{\pi} &= \frac{\pi(x|H_1)}{\pi(x|H_0)} \\
&= \frac{\int f(x|\theta)\,\pi(\theta|H_1)\,d\theta}{\int f(x|\theta)\,\pi(\theta|H_0)\,d\theta} \\
&= \frac{\int f(x|\theta)\,\pi_1(\theta)\,d\theta}{\int f(x|\theta)\,\pi_0(\theta)\,d\theta} \\
&= \frac{\pi(H_1|x)}{\pi(H_0|x)}\frac{\pi(H_0)}{\pi(H_1)}
\end{aligned}
$$

- Bayes procedures can be directly used to test point null hypothesis; i.e. $H_0 : \theta = \theta_0$ (that is $\pi_0(\theta) = \delta_{\theta_0}(\theta)$) versus $H_1 : \theta \sim \pi_1$ where the prior is then defined as

$$\pi(\theta) = \pi(H_0)\, \delta_{\theta_0}(\theta) + \pi(H_1)\, \pi_1(\theta)$$

- The associated Bayes factor is simply

$$B_{10}^{\pi} = \frac{\pi(x|\, H_1)}{\pi(x|\, H_0)} = \frac{\int \pi(x|\, \theta)\, \pi_1(\theta)\, d\theta}{\pi(x|\, \theta_0)}.$$

3.4– Example

- Assume you have an coin, you toss it 10 times and gets x heads. Is it biased?

- Let θ be the proba of having a head then we can test $H_0 : \theta = \frac{1}{2}$ versus $H_1 : \theta \sim \mathcal{U}\left(\frac{1}{2}, 1\right]$ using

$$B_{10}^{\pi} = \frac{\frac{1}{2}\int_{\frac{1}{2}}^{1} \begin{pmatrix} 10 \\ \\ x \end{pmatrix} \theta^{x}\left(1-\theta\right)^{10-x} d\theta}{\begin{pmatrix} 10 \\ \\ x \end{pmatrix} \left(\frac{1}{2}\right)^{x}\left(1-\frac{1}{2}\right)^{10-x}}.$$

or $H_0 : \theta = \frac{1}{2}$ versus $H_1 : \theta \sim \mathcal{U}\left[0,1\right]$ using

$$B_{10}^{\pi} = \frac{\int_0^1 \begin{pmatrix} 10 \\ \\ x \end{pmatrix} \theta^x \left(1-\theta\right)^{10-x} d\theta}{\begin{pmatrix} 10 \\ \\ x \end{pmatrix} \left(\frac{1}{2}\right)^x \left(1-\frac{1}{2}\right)^{10-x}}.$$

3.5– Towards Bayes Model Selection

- Bayes factors are not limited to the comparison of models with the same parameter space.

- Assume you have some data and two statistical models.
Under H_0, $\theta_0 \in \Theta_0$, the prior is $\pi_0(\theta_0)$ and the likelihood is $f_0(x|\theta_0)$.
Under H_1, $\theta_1 \in \Theta_1$, the prior is $\pi_1(\theta_1)$ and the likelihood is $f_1(x|\theta_1)$
then

$$B_{10}^{\pi} = \frac{\pi(x|H_1)}{\pi(x|H_0)} = \frac{\int f_1(x|\theta_1)\,\pi_1(\theta_1)\,d\theta_1}{\int f_0(x|\theta_0)\,\pi_0(\theta_0)\,d\theta_0}$$

- One can have $\Theta_0 = \mathbb{R}$ and $\Theta_1 = \mathbb{R}^{1000}$.

- In this case you have a parameter space $\Theta = \Theta_0 \cup \Theta_1$ which is the union two subspaces of different dimensions.

3.5– Towards Bayes Model Selection

- Jeffreys' scale of evidence says that

  - if $\log_{10}(B_{10}^{\pi})$ varies between 0 and 0.5, the evidence against H_0 is poor,
  - if it is between 0.5 and 1, it is substantial,
  - if it is between 1 and 2, it is strong, and
  - if it is above 2, it is decisive.

- Bayes factor tell you where one should prefer H_0 to H_1: it does NOT tell you whether any of these models are sensible!

3.5– Towards Bayes Model Selection

- You can also compute the posterior probabilities of H_0 and H_1

$$\pi(H_0|x) = \frac{\pi(x|H_0)\,\pi(H_0)}{\pi(x)}$$

$$= \frac{\pi(x|H_0)\,\pi(H_0)}{\pi(x|H_0)\,\pi(H_0) + \pi(x|H_1)\,\pi(H_1)}.$$

- The posterior probabilities satisfy

$$\frac{\pi(H_1|x)}{\pi(H_0|x)} = \frac{\pi(x|H_1)}{\pi(x|H_0)}\frac{\pi(H_1)}{\pi(H_0)} = B_{10}^{\pi}\frac{\pi(H_1)}{\pi(H_0)}.$$

- Testing hypothesis in a Bayesian way is attractive.... but be careful to vague priors!!!

- Assume you have $X|\left(\mu,\sigma^2\right)\sim\mathcal{N}\left(\mu,\sigma^2\right)$ where σ^2 is assumed known but μ (the parameter θ) is unknown. We want to test $H_0:\mu=0$
vs $H_1:\mu\sim\mathcal{N}\left(\xi,\tau^2\right)$ then

$$\begin{aligned} B_{10}^{\pi}(x) &= \frac{\pi\left(x|\,H_1\right)}{\pi\left(x|\,H_0\right)}=\frac{\int\mathcal{N}\left(x;\mu,\sigma^2\right)\mathcal{N}\left(\mu;\xi,\tau^2\right)d\mu}{f\left(x|\,0\right)} \\ &= \frac{\sigma}{\sqrt{\sigma^2+\tau^2}}\exp\left(\frac{\tau^2x^2}{2\sigma^2\left(\sigma^2+\tau^2\right)}\right)\underset{\tau^2\to\infty}{\longrightarrow}0 \end{aligned}$$

3.5– Towards Bayes Model Selection

• The Bayes factor depends heavily on τ^2. As $\tau^2 \to \infty$, the prior becomes uniformative but then $B_{10}^{\pi}(x) \to 0$ whatever being x and $\pi(H_0|x) \to 1$.

• We will see that next week but using vague priors for model selection is a very very bad idea... (Lindley's paradox).

3.6– Bayesian Polynomial Regression Example

• In practice, you might have more than 2 potential models/hypothesis for your data.

• Consider the following polynomial regression problem where $D = \{x_i, y_i\}_{i=1}^{n}$ where $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$.

$$\begin{aligned} Y &= \sum_{i=0}^{M} \beta_i X^i + \sigma V \text{ where } V \sim \mathcal{N}(0,1) \\ &= \beta_{0:M}^{\mathrm{T}} f_M(X) + \sigma V \end{aligned}$$

• Here the problem is that if M is too large, then the MLE estimate will overfit.

As M increases, the model overfits.