Bayesian Ingredients

Hedibert Freitas Lopes

The University of Chicago Booth School of Business
5807 South Woodlawn Avenue, Chicago, IL 60637
http://faculty.chicagobooth.edu/hedibert.lopes

hlopes@ChicagoBooth.edu

Outline

Example i. Sequential learning

- John claims some discomfort and goes to the doctor.
- The doctor believes John may have the disease A.
- $\theta = 1$: John has disease A; $\theta = 0$: he does not.
- The doctor claims, based on his expertise (H), that

$$P(\theta = 1|H) = 0.70$$

- Examination X is related to θ as follows

$$\begin{cases} P(X = 1|\theta = 0) = 0.40, & \text{positive test given no disease} \\ P(X = 1|\theta = 1) = 0.95, & \text{positive test given disease} \end{cases}$$

Observe $X = 1$

Exam's result: $X = 1$

$$
\begin{aligned}
P(\theta = 1|X = 1) &\propto P(X = 1|\theta = 1)P(\theta = 1) \\
&\propto (0.95)(0.70) = 0.665 \\
P(\theta = 0|X = 1) &\propto P(X = 1|\theta = 0)P(\theta = 0) \\
&\propto (0.40)(0.30) = 0.120
\end{aligned}
$$

Consequently

$$
\begin{aligned}
P(\theta = 0|X = 1) &= 0.120/0.785 = 0.1528662 \text{ and} \\
P(\theta = 1|X = 1) &= 0.665/0.785 = 0.8471338
\end{aligned}
$$

The information $X = 1$ increases, for the doctor, the probability that John has the disease A from 70% to 84.71%.

Posterior predictive

John undertakes the test Y, which relates to θ as follows[1]

$$P(Y=1|\theta=1)=0.99 \quad \text{and} \quad P(Y=1|\theta=0)=0.04$$

Then, the predictive of $Y=0$ given $X=1$ is given by

$$\begin{aligned}
P(Y=0|X=1) &= P(Y=0|X=1,\theta=0)P(\theta=0|X=1) \\
&+ P(Y=0|X=1,\theta=1)P(\theta=1|X=1) \\
&= P(Y=0|\theta=0)P(\theta=0|X=1) \\
&+ P(Y=0|\theta=1)P(\theta=1|X=1) \\
&= (0.96)(0.1528662)+(0.01)(0.8471338) \\
&= 15.52\%
\end{aligned}$$

Key condition: X and Y are conditionally independent given θ.

[1] Recall that $P(X=1|\theta=1)=0.95$ and $P(X=1|\theta=0)=0.40$.

Model criticism

Suppose the observed result was $Y = 0$. This is a reasonably unexpected result as the doctor only gave it roughly 15% chance.

He should at least consider rethinking the model based on this result. In particular, he might want to ask himself

1. Did 0.7 adequately reflect his $P(\theta = 1|H)$?
2. Is test X really so unreliable?
3. Is the sample distribution of X correct?
4. Is the test Y so powerful?
5. Have the tests been carried out properly?

Observe $Y = 0$

Let $H_2 = \{X = 1, Y = 0\}$. Then, Bayes theorem leads to

$$\begin{aligned}
P(\theta = 1|H_2) &\propto P(Y = 0|\theta = 1)P(\theta = 1|X = 1) \\
&\propto (0.01)(0.8471338) = 0.008471338 \\
P(\theta = 0|H_2) &\propto P(Y = 0|\theta = 0)P(\theta = 0|X = 1) \\
&\propto (0.96)(0.1528662) = 0.1467516
\end{aligned}$$

Therefore,

$$P(\theta = 1|X = 1, Y = 0) = \frac{P(Y = 0, \theta = 1|X = 1)}{P(Y = 0|X = 1)} = 0.0545753$$

$$P(\theta = 1|H_i) = \begin{cases} 0.7000 & , H_0\text{: before X and Y} \\ 0.8446 & , H_1\text{: after X=1 and before Y} \\ 0.0546 & , H_2\text{: after X=1 and Y=0} \end{cases}$$

Example *ii*. Normal model and normal prior

Let us now consider a simple measurement error model with normal prior for the unobserved measurement.

$$\begin{aligned} X|\theta &\sim N(\theta, \sigma^2) \\ \theta &\sim N(\theta_0, \tau_0^2) \end{aligned}$$

with σ^2, θ_0 and τ_0^2 known for now. It is easy to show that the posterior of θ given $X = x$ is also normal.

More precisely, $(\theta|X = x) \sim N(\theta_1, \tau_1^2)$ where

$$\begin{aligned} \theta_1 &= w\theta_0 + (1 - w)x \\ \tau_1^{-2} &= \tau_0^{-2} + \sigma^{-2} \\ w &= \tau_0^{-2}/(\tau_0^{-2} + \sigma^{-2}) \end{aligned}$$

w measures the relative information contained in the prior distribution with respect to the total information (prior plus likelihood).

Example from Box & Tiao (1973)

Prior A: Physicist A (large experience): $\theta \sim N(900, (20)^2)$

Prior B: Physicist B (not so experienced): $\theta \sim N(800, (80)^2)$.

Model: $(X|\theta) \sim N(\theta, (40)^2)$.

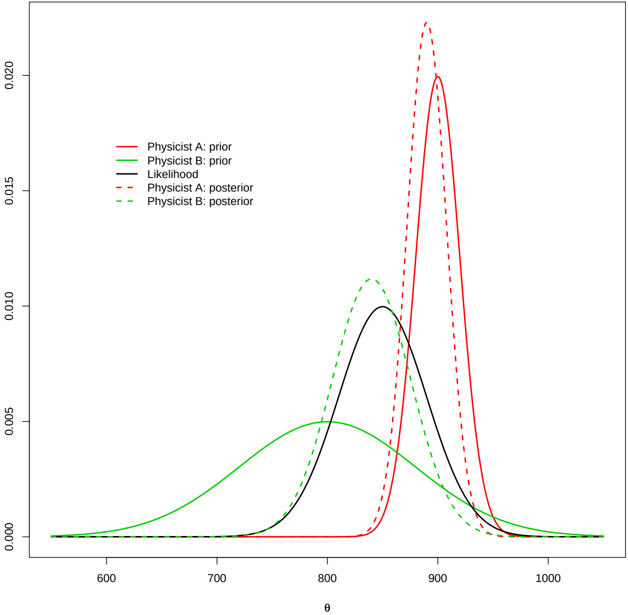
Observation: $X = 850$

$$\begin{aligned}(\theta|X = 850, H_A) &\sim N(890, (17.9)^2)\\(\theta|X = 850, H_B) &\sim N(840, (35.7)^2)\end{aligned}$$

Information (precision)

Physicist A: from 0.002500 to 0.003120 (an increase of 25%)

Physicist B: from 0.000156 to 0.000781 (an increase of 400%)


Physicist A: prior
Physicist B: prior
Likelihood
Physicist A: posterior
Physicist B: posterior
0.000
0.005
0.010
0.015
0.020
600
700
800
900
1000
θ

Two additional examples

Observations $\mathbf{x}$, parameters $\boldsymbol{\theta}$ and history H.

Likelihood functions/models - $p(\mathbf{x}|\boldsymbol{\theta}, H)$

$$\begin{array}{lcll} \text{Example iii} & : & x_i|\theta, H \sim N(\theta z_i; \sigma^2) & i = 1, \ldots, n \\ \text{Example iv} & : & x_t|\theta, H \sim N(0; e^{\theta_t}) & t = 1, \ldots, T \end{array}$$

Prior distributions - $p(\boldsymbol{\theta}|H)$

$$\begin{array}{lcll} \text{Example iii} & : & \theta|H \sim N(\theta_0, \tau_0^2) & \\ \text{Example iv} & : & \theta_t|H \sim N(\alpha + \beta\theta_{t-1}, \sigma^2) & t = 1, \ldots, T \end{array}$$

Assume, for now, that $(z_1, \ldots, z_n, \sigma^2, \nu_0, \theta_0, \tau_0^2)$ and $(\alpha, \beta, \sigma^2, \theta_0)$ are known and belong to H.

Turning the Bayesian crank

Posterior (Bayes' Theorem):

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{x}, H) &= \frac{p(\boldsymbol{\theta}, \mathbf{x}|H)}{p(\mathbf{x}|H)} \\
&= \frac{p(\mathbf{x}|\boldsymbol{\theta}, H)p(\boldsymbol{\theta}|H)}{p(\mathbf{x}|H)} \\
&\propto p(\mathbf{x}|\boldsymbol{\theta}, H)p(\boldsymbol{\theta}|H)
\end{aligned}
$$

Prior predictive distribution:

$$
p(\mathbf{x}|H) = \int_{\Theta} p(\mathbf{x}|\boldsymbol{\theta}, H)p(\boldsymbol{\theta}|H)\, d\boldsymbol{\theta} = E_{\boldsymbol{\theta}}[p(\mathbf{x}|\boldsymbol{\theta}, H)]
$$

Posterior predictive distribution

Let $\mathbf{y}$ be a new set of observations conditionally independent of $\mathbf{x}$ given $\boldsymbol{\theta}$. Then,

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}, H) &= \int_{\Theta} p(\mathbf{y}, \boldsymbol{\theta}|\mathbf{x}, H) d\boldsymbol{\theta} = \int_{\Theta} p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{x}, H) p(\boldsymbol{\theta}|\mathbf{x}, H) d\boldsymbol{\theta} \\ &= \int_{\Theta} p(\mathbf{y}|\boldsymbol{\theta}, H) p(\boldsymbol{\theta}|\mathbf{x}, H) d\boldsymbol{\theta} = E_{\boldsymbol{\theta}|\mathbf{x}} [p(\mathbf{y}|\boldsymbol{\theta}, H)] \end{aligned}$$

Note 1: In general, but not always (time series, for example) $\mathbf{x}$ and $\mathbf{y}$ are independent given $\boldsymbol{\theta}$.

Note 2: It might be more useful to concentrate on prediction rather than on estimation since the former is *verifiable*. In other words, $\mathbf{x}$ and $\mathbf{y}$ can be observed; not $\boldsymbol{\theta}$.

Sequential Bayes theorem: A rule for updating probabilities

Experimental result: $\mathbf{x}_1 \sim p_1(\mathbf{x}_1 \mid \boldsymbol{\theta})$

$$p(\boldsymbol{\theta} \mid \mathbf{x}_1) \propto l_1(\boldsymbol{\theta}; x_1) p(\boldsymbol{\theta})$$

Experimental result: $\mathbf{x}_2 \sim p_2(\mathbf{x}_2 \mid \boldsymbol{\theta})$

$$\begin{aligned} p(\boldsymbol{\theta} \mid \mathbf{x}_2, \mathbf{x}_1) &\propto l_2(\boldsymbol{\theta}; \mathbf{x}_2) p(\boldsymbol{\theta} \mid \mathbf{x}_1) \\ &\propto l_2(\boldsymbol{\theta}; \mathbf{x}_2) l_1(\boldsymbol{\theta}; \mathbf{x}_1) p(\boldsymbol{\theta}) \end{aligned}$$

Experimental results: $\mathbf{x}_i \sim p_i(\mathbf{x}_i \mid \boldsymbol{\theta})$, for $i = 3, \cdots, n$

$$\begin{aligned} p(\boldsymbol{\theta} \mid \mathbf{x}_n, \cdots, \mathbf{x}_1) &\propto l_n(\boldsymbol{\theta}; \mathbf{x}_n) p(\boldsymbol{\theta} \mid \mathbf{x}_{n-1}, \cdots, \mathbf{x}_1) \\ &\propto \left[\prod_{i=1}^{n} l_i(\boldsymbol{\theta}; \mathbf{x}_i) \right] p(\boldsymbol{\theta}) \end{aligned}$$

Example iii. Simple linear regression

Model, prior and posterior:

$$\begin{aligned} x_i|\theta, H &\sim N(\theta z_i; \sigma^2) \qquad i = 1, \ldots, n \\ \theta|H &\sim N(\theta_0, \tau_0^2) \\ \theta|\mathbf{x}, H &\sim N(\theta_1, \tau_1^2) \end{aligned}$$

where

$$\tau_1^{-2} = \tau_0^{-2} + \mathbf{z}'\mathbf{z}/\sigma^2 \quad \text{and} \quad \theta_1 = \tau_1^2\left(\theta_0\tau_0^{-2} + \mathbf{z}'\mathbf{x}/\sigma^2\right)$$

Note 1: As n increases, $\tau_1 \to \sigma^2(\mathbf{z}'\mathbf{z})^{-1}$ and $\theta_1 \to (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{x}$.

Note 2: The same applies when $\tau_0^{-2} \to 0$, i.e. with 'little' prior knowledge about θ.

Example iv. SV model

Model, prior and posterior:

$$
\begin{aligned}
x_t|\theta_t, H &\sim N(0; e^{\theta_t}) \qquad t = 1, \ldots, T \\
\theta_t|H &\sim N(\alpha + \beta\theta_{t-1}, \sigma^2) \qquad t = 1, \ldots, T \\
p(\boldsymbol{\theta}|\mathbf{x}, H) &\propto \prod_{t=1}^{T} e^{-\theta_t/2} \exp\left\{-\frac{1}{2} x_t e^{-\theta_t}\right\} \\
&\times \prod_{t=1}^{T} \exp\left\{-\frac{1}{2\sigma^2}(\theta_t - \alpha - \beta\theta_{t-1})^2\right\}
\end{aligned}
$$

Unfortunately, closed form solutions are rare.

- How to compute $E(\theta_{43}|\mathbf{x}, H)$ or $V(\theta_{11}|\mathbf{x}, H)$?
- How to obtain a 95% credible region for $(\theta_{35}, \theta_{36}|\mathbf{x}, H)$?
- How to sample from $p(\theta|\mathbf{x}, H)$?
- How to compute $p(\mathbf{x}|H)$ or $p(x_{T+1}, \ldots, x_{T+k}|\mathbf{x}, H)$?

Example v. Multiple linear regression

The standard Bayesian approach to multiple linear regression is

$$(y|X, \beta, \sigma^2) \sim N(X\beta, \sigma^2 I_n)$$

where $y = (y_1, \ldots, y_n)$, $X = (x_1, \ldots, x_n)'$ is the $(n \times q)$, design matrix and $q = p + 1$.

The prior distribution of (β, σ^2) is $NIG(b_0, B_0, n_0, S_0)$, i.e.

$$\begin{aligned} \beta|\sigma^2 &\sim N(b_0, \sigma^2 B_0) \\ \sigma^2 &\sim IG(n_0/2, n_0 S_0/2) \end{aligned}$$

for known hyperparameters b_0, B_0, n_0 and S_0.

Example v. Conditionals and marginals

It is easy to show that (β, σ^2) is $NIG(b_1, B_1, n_1, S_1)$, i.e.

$$
\begin{aligned}
(\beta|\sigma^2, y, X) &\sim N(b_1, \sigma^2 B_1) \\
(\sigma^2|y, X) &\sim IG(n_1/2, n_1 S_1/2)
\end{aligned}
$$

where

$$
\begin{aligned}
B_1^{-1} &= B_0^{-1} + X'X \\
B_1^{-1} b_1 &= B_0^{-1} b_0 + X'y \\
n_1 &= n_0 + n \\
n_1 S_1 &= n_0 S_0 + (y - Xb_1)'y + (b_0 - b_1)' B_0^{-1} b_0.
\end{aligned}
$$

It is also easy to derive the full conditional distributions, i.e.

$$
\begin{aligned}
(\beta|y, X) &\sim t_{n_1}(b_1, S_1 B_1) \\
(\sigma^2|\beta, y, X) &\sim IG(n_1/2, n_1 S_{11}/2)
\end{aligned}
$$

where

$$n_1 S_{11} = n_0 S_0 + (y - X\beta)'(y - X\beta).$$

Example v. Ordinary least squares

It is well known that

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'y \\ \hat{\sigma}^2 &= \frac{S_e}{n-q} = \frac{(y-X\hat{\beta})'(y-X\hat{\beta})}{n-q}\end{aligned}$$

are the OLS estimates of β and σ^2, respectively.

The conditional and unconditional sampling distributions of $\hat{\beta}$ are

$$\begin{aligned}(\hat{\beta}|\sigma^2, y, X) &\sim N(\beta, \sigma^2(X'X)^{-1}) \\ (\hat{\beta}|y, X) &\sim t_{n-q}(\beta, S_e)\end{aligned}$$

respectively, with

$$(\hat{\sigma}^2|\sigma^2) \sim IG\left((n-q)/2, ((n-q)\sigma^2/2\right).$$

Example v. Sufficient statistics

Recall $(y_t|x_t, \beta, \sigma^2) \sim N(x_t'\beta, \sigma^2)$ for $t = 1, \ldots, n$, with prior $\beta|\sigma^2 \sim N(b_0, \sigma^2 B_0)$ and $\sigma^2 \sim IG(n_0/2, n_0 S_0/2)$.

Then, for $y^t = (y_1, \ldots, y_t)$ and $X^t = (x_1, \ldots, x_t)'$, it follows:

$$
\begin{aligned}
(\beta|\sigma^2, y^t, X^t) &\sim N(b_t, \sigma^2 B_t) \\
(\sigma^2|y^t, X^t) &\sim IG(n_t/2, n_t S_t/2)
\end{aligned}
$$

where $n_t = n_{t-1} + 1$, $B_t^{-1} = B_{t-1}^{-1} + x_t x_t'$, $B_t^{-1} b_t = B_{t-1}^{-1} b_{t-1} + y_t x_t$ and $n_t S_t = n_{t-1} S_{t-1} + (y_t - b_t' x_t) y_t + (b_{t-1} - b_t)' B_{t-1}^{-1} b_{t-1}$.

The only ingredients needed are: $x_t x_t'$, $y_t x_t$ and y_t^2.

These recursions will play an important role later on when deriving **sequential Monte Carlo** methods for conditionally Gaussian dynamic linear models, like many stochastic volatility models.

Example v. Predictive

The predictive density can be seen as the *marginal likelihood*, i.e.

$$p(y|X) = \int p(y|X,\beta,\sigma^2)p(\beta|\sigma^2)p(\sigma^2)d\beta d\sigma^2$$

or, by Bayes' theorem, as the *normalizing constant*, i.e.

$$p(y|X) = \frac{p(y|X,\beta,\sigma^2)p(\beta|\sigma^2)p(\sigma^2)}{p(\beta|\sigma^2,y,X)p(\sigma^2|y,X)}$$

which is valid for all (β,σ^2).

Closed form solution is available for the multiple normal linear regression:

$$(y|X) \sim t_{n_0}(Xb_0, S_0(I_n + XB_0X')).$$