Serik Sagitov, Chalmers Tekniska Högskola, February 13, 2013

Introduction to Bayesian inference

1 Bayesian approach

Main idea of the Baysian approach: treat the population parameter θ is a random variable. Two distributions of θ

prior distribution density $g(\theta)$ = knowledge on θ before data is collected,
posterior distribution $h(\theta|x)$ = knowledge on θ updated after the data x is collected.

$$\boxed{\text{Bayes formula } h(\theta|x) = \tfrac{f(x|\theta)g(\theta)}{\phi(x)}} \qquad \boxed{\text{Posterior} \propto \text{likelihood} \times \text{prior}}$$

Marginal distribution of X has density $\phi(x) = \int f(x|\theta)g(\theta)d\theta$. This is the likelihood $f(x|\theta)$ of the data weighed over different values of θ using the prior distribution.

Example. IQ measurement.
A randomly chosen individual has IQ θ. Its prior distribution is $\theta \sim \mathrm{N}(100{,}225)$ describing population as a whole: average IQ is $m = 100$ and standard deviation $v = 15$. The result of an IQ measurement has distribution $X \sim \mathrm{N}(\theta,\, 100)$: no systematic error and random error $\sigma = 10$. We have

$$g(\theta) = \frac{1}{\sqrt{2\pi}v} e^{-\frac{(\theta-m)^2}{2v^2}}, \quad f(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\theta)^2}{2\sigma^2}},$$

and $h(\theta|x)$ is proportional to $g(\theta)f(x|\theta)$. Put $\gamma = \frac{\sigma^2}{\sigma^2+v^2}$, shrinkage factor. Since

$$e^{-\frac{(\theta-m)^2}{2v^2}} e^{-\frac{(x-\theta)^2}{2\sigma^2}} = \exp\left\{-\frac{(\theta-m)^2}{2v^2} - \frac{(x-\theta)^2}{2\sigma^2}\right\} = \exp\left\{-\frac{(\theta-\gamma m-(1-\gamma)x)^2}{2\gamma v^2}\right\},$$

we conclude that the posterior distribution is normal

$$h(\theta|x) = \frac{1}{\sqrt{2\pi\gamma}v} e^{-\frac{(\theta-\gamma m-(1-\gamma)x)^2}{2\gamma v^2}}.$$

If observed IQ is $x = 130$, then the posterior distribution is $\theta \sim \mathrm{N}(120.7,\, 69.2)$.

2 Conjugate priors

Two families of probability distributions G and H

$$\boxed{G \text{ is a family of conjugate priors to } H\text{, if a } G\text{-prior and a } H\text{-likelihood give a } G\text{-posterior}}$$

Examples of conjugate priors

Data distribution	Prior	Posterior distribution	Comments
$(X_1,\ldots,X_n), X_i \sim \mathrm{N}(\theta,\sigma^2)$	$\mu \sim \mathrm{N}(m,v^2)$	$\mathrm{N}(\gamma_n m + (1-\gamma_n)\bar{x}; \gamma_n v^2)$	$\gamma_n = \frac{\sigma^2}{\sigma^2+nv^2}$
$X \sim \mathrm{Bin}(n,p)$	$p \sim \mathrm{Beta}(a,b)$	$\mathrm{Beta}(a+x, b+n-x)$	counts plus ...
$(X_1,\ldots,X_r) \sim \mathrm{Mn}(n;p_1,\ldots,p_r)$	$\mathrm{D}(\alpha_1,\ldots,\alpha_r)$	$\mathrm{D}(\alpha_1+x_1,\ldots,\alpha_r+x_r)$	... pseudocounts
$X \sim \mathrm{Pois}(\mu)$	$\mu \sim \Gamma(\alpha,\lambda)$	$\Gamma(\alpha+x, \lambda+1)$	posterior variance ...
$X \sim \mathrm{Exp}(\rho)$	$\rho \sim \Gamma(\alpha,\lambda)$	$\Gamma(\alpha+1, \lambda+x)$	... is always smaller

Beta distribution $\mathrm{Beta}(a, b)$ density $f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}$, $0 < p < 1$.

Mean and variance $\mu = \frac{a}{a+b}$, $\sigma^2 = \frac{\mu(1-\mu)}{a+b+1}$, pseudocounts $a > 0$, $b > 0$.

Dirichlet distribution $\mathrm{D}(\alpha_1, \ldots, \alpha_r)$ density $f(p_1, \ldots, p_r) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_r)} p_1^{\alpha_1 - 1} \ldots p_r^{\alpha_r - 1}$

with non-negative $p_1 + \ldots + p_r = 1$. Positive pseudocounts $\alpha_1, \ldots, \alpha_r$, $\alpha_0 = \alpha_1 + \ldots + \alpha_r$.

Marginal distributions

$p_j \sim \mathrm{Beta}(\alpha_j, \alpha_0 - \alpha_j)$, $j = 1, \ldots, r$, and negative covariances $\mathrm{Cov}(p_1, p_2) = -\frac{\alpha_1\alpha_2}{\alpha_0^2(\alpha_0+1)}$.

Example. Thumbtack experiment. Beta-binomial model: number of base landings $X \sim \mathrm{Bin}(n, p)$ for n tossings of the thumbtack with $p = \mathrm{P}(\text{landing on base})$.

My personal Beta prior $p \sim \mathrm{B}(a_0, b_0)$ with $\mu_0 \approx 0.4$, $\sigma_0 \approx 0.1 \Rightarrow$ pseudocounts $a_0 = 10$, $b_0 = 15$.

Experiment 1: $n_1 = 10$ tosses, counts $x_1 = 2$, $n_1 - x_1 = 8$, posterior distribution Beta(12, 23) with mean $\hat{p} = \frac{12}{35} = 0.34$ and standard deviation $\sigma_1 = 0.08$.

Experiment 2: $n_2 = 40$ tosses, counts $x_2 = 9$, $n_2 - x_2 = 31$, posterior distribution Beta(21, 54) with mean $\hat{p} = \frac{21}{75} = 0.28$ and standard deviation $\sigma_2 = 0.05$.

3 Bayesian estimation

Action $a = \{\text{assign value } a \text{ to unknown parameter } \theta\}$. Optimal action depends on the choice of the loss function $l(\theta, a)$. Bayes action minimizes posterior risk

$$R(a|x) = \int l(\theta, a)h(\theta|x)d\theta \quad \text{or} \quad R(a|x) = \sum_\theta l(\theta, a)h(\theta|x).$$

MAP = maximum a posteriori probability estimate is based on

$$\boxed{\text{Zero-one loss function: } l(\theta, a) = 1_{\{\theta \neq a\}}}$$

Posterior risk = probability of misclassification $R(a|x) = \sum_{\theta \neq a} h(\theta|x) = 1 - h(a|x)$

$\hat{\theta}_{\text{map}} = \theta$ that maximizes $h(\theta|x)$.

For the non-informative prior $g(\theta) = \text{const}$, we get $h(\theta|x) \propto f(x|\theta)$ and $\hat{\theta}_{\text{map}} = \hat{\theta}_{\text{mle}}$.

PME = posterior mean estimate $\hat{\theta}_{\text{pme}} = \mathrm{E}(\theta|x)$ is based on

$$\boxed{\text{Squared error loss: } l(\theta, a) = (\theta - a)^2}$$

$R(a|x) = \mathrm{E}((\theta - a)^2|x) = \mathrm{Var}(\theta|x) + [\mathrm{E}(\theta|x) - a]^2$.

Example. Loaded die experiment. A possibly loaded die is rolled 18 times:

211 453 324 142 343 515.

If the prior distribution is non-informative D(1,1,1,1,1,1), then MAP = MLE are given by the sample proportions $(\frac{4}{18}, \frac{3}{18}, \frac{4}{18}, \frac{4}{18}, \frac{3}{18}, 0)$. Not good: it excludes sixes in the future.

With the same prior D(1,1,1,1,1,1) the PME are

$\hat{p}_1 = \frac{5}{24} = 0.21$, $\hat{p}_2 = \frac{4}{24} = 0.17$, $\hat{p}_3 = \frac{5}{24} = 0.21$, $\hat{p}_4 = \frac{5}{24} = 0.21$, $\hat{p}_5 = \frac{4}{24} = 0.17$, $\hat{p}_6 = \frac{1}{24} = 0.04$.

4 Credibility interval

Confidence interval : θ is an unknown constant and a CI is random
$\mathrm{P}(\theta_0(X) < \theta < \theta_1(X)) = 1 - \alpha$.
Credibility interval: θ is random and a CrI is nonrandom. It is computed from the posterior distribution $\mathrm{P}(\theta_0(x) < \theta < \theta_1(x)) = 1 - \alpha$.

Example. IQ measurement.
Given $n = 1$, $\bar{X} \sim \mathrm{N}(\mu; 100)$ a 95% CI for μ is $130 \pm 1.96 \cdot 10 = 130 \pm 19.6$.
Posterior distribution of μ is $\mathrm{N}(120.7; 69.2)$
95% CrI for μ is $120.7 \pm 1.96 \cdot \sqrt{69.2} = 120.7 \pm 16.3$.

5 Hypotheses testing

Choose between H_0: $\theta = \theta_0$ and H_1: $\theta = \theta_1$
given prior probabilities $\mathrm{P}(H_0) = \pi_0$, $\mathrm{P}(H_1) = \pi_1$ and the likelihoods $f(x|\theta_0)$, $f(x|\theta_1)$.
Cost function: l_{I} = error type I cost, l_{II} = error type II cost.
Average cost for given a RR = rejection region

$$l_{\mathrm{I}}\pi_0 \mathrm{P}(X \in \mathrm{RR}|\theta_0) + l_{\mathrm{II}}\pi_1 \mathrm{P}(X \notin \mathrm{RR}|\theta_1) = l_{\mathrm{II}}\pi_1 + \int_{x \in \mathrm{RR}} \Big(l_{\mathrm{I}}\pi_0 f(x|\theta_0) - l_{\mathrm{II}}\pi_1 f(x|\theta_1) \Big) dx,$$

where the integral is taken over the RR. The rejection region minimizing the average cost is

$$\boxed{\mathrm{RR} = \{x\colon l_{\mathrm{I}}\pi_0 f(x|\theta_0) < l_{\mathrm{II}}\pi_1 f(x|\theta_1)\}}$$

Reject H_0 if small likelihood ratio $\frac{f(x|\theta_0)}{f(x|\theta_1)} < \frac{l_{\mathrm{II}}\pi_1}{l_{\mathrm{I}}\pi_0}$ or small posterior odds $\frac{h(\theta_0|x)}{h(\theta_1|x)} < \frac{l_{\mathrm{II}}}{l_{\mathrm{I}}}$.

Example. Rape case study.
The defendant A, age 37, local, is charged with rape, H_0: A is innocent, H_1: A is guilty,
error I: a non-guilty is convicted, against error II: a guilty is unpunished.
Prior probability $\mathrm{P}(H_1) = \frac{1}{200{,}000}$.
Evidence:
E_1: DNA match, $\mathrm{P}(E_1|H_0) = \frac{1}{200{,}000{,}000}$, $\mathrm{P}(E_1|H_1)$=1,
E_2: A is not recognized by the victim,
E_3: alibi supported by the girlfriend.
Assumptions
$\mathrm{P}(E_2|H_1) = 0.1$, $\mathrm{P}(E_2|H_0) = 0.9$
$\mathrm{P}(E_3|H_1) = 0.25$, $\mathrm{P}(E_3|H_0) = 0.5$
Posterior probabilities
$\mathrm{P}(H_1|E_1) = \frac{\mathrm{P}(E_1|H_1)\mathrm{P}(H_1)}{\mathrm{P}(E_1|H_1)P(H_1)+\mathrm{P}(E_1|H_0)P(H_0)} = \frac{1000}{1001}$
$\mathrm{P}(H_1|E_1, E_2) = \frac{\mathrm{P}(E_2|H_1)\mathrm{P}(H_1|E_1)}{\mathrm{P}(E_2|H_1)\mathrm{P}(H_1|E_1)+\mathrm{P}(E_2|H_0)\mathrm{P}(H_0|E_1)} = \frac{1000}{1009}$
$\mathrm{P}(H_1|E_1, E_2, E_3) = \frac{\mathrm{P}(E_3|H_1)\mathrm{P}(H_1|E_1,E_2)}{\mathrm{P}(E_3|H_1)\mathrm{P}(H_1|E_1,E_2)+\mathrm{P}(E_3|H_0)\mathrm{P}(H_0|E_1,E_2)} = \frac{1000}{1018}$.

Posterior odds $\frac{\mathrm{P}(H_0|E_1,E_2,E_3)}{\mathrm{P}(H_1|E_1,E_2,E_3)} = \frac{18}{1000} = 0.018$, reject H_0 if $\frac{l_{\mathrm{II}}}{l_{\mathrm{I}}} > 0.018$.

Is it better for fifty guilty people to go unpunished than for one innocent man to be convicted?