Serik Sagitov, Chalmers Tekniska Högskola, November 7, 2014

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.

Bayes formula $h(\theta|x) = \frac{f(x|\theta)g(\theta)}{\phi(x)}$ Posterior $\propto$ likelihood $\times$ 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 $\theta \sim$ 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 $X \sim$ N(θ, 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 $\theta \sim$ N(120.7, 69.2).

2 Conjugate priors

Two families of probability distributions G and H

G is a family of conjugate priors to H, if a G-prior and a H-likelihood give a G-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 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 $p=\mathrm{P}$(landing on base).
My personal Beta prior $p\sim\mathrm{B}(a_0,b_0)$ with $\mu_0\approx0.4$, $\sigma_0\approx0.1\Rightarrow$ 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 $a=\{$assign value a 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:

Measurement outcome	Decision	H_0 true	H_1 true
$X \in \mathrm{RR}$	Accept H_0	0	c_1
$X \notin \mathrm{RR}$	Accept H_1	c_0	0

Average cost for a given rejection region RR

$$c_0\pi_0\mathrm{P}(X \in \mathrm{RR}|\theta_0) + c_1\pi_1\mathrm{P}(X \notin \mathrm{RR}|\theta_1) = c_1\pi_1 + \int_{x\in\mathrm{RR}} \Big(c_0\pi_0 f(x|\theta_0) - c_1\pi_1 f(x|\theta_1)\Big)dx,$$

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

$$\boxed{\mathrm{RR} = \{x\colon c_0\pi_0 f(x|\theta_0) < c_1\pi_1 f(x|\theta_1)\}}$$

Reject H_0 if small likelihood ratio $\frac{f(x|\theta_0)}{f(x|\theta_1)} < \frac{c_1\pi_1}{c_0\pi_0}$ or small posterior odds $\frac{h(\theta_0|x)}{h(\theta_1|x)} < \frac{c_1}{c_0}$.

Example. Rape case study.
The defendant A, age 37, local, is charged with rape, H_0: A is innocent, H_1: A is guilty.
Prior probability $\pi_1 = \frac{1}{200{,}000}$.
Evidence E with conditionally independent components

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

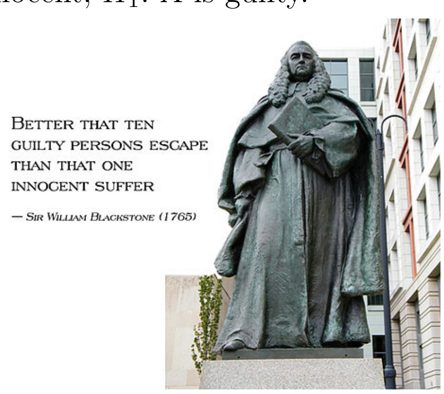
Assumptions

$\mathrm{P}(E_2|H_1) = 0.1$, $\mathrm{P}(E_2|H_0) = 0.9$

$\mathrm{P}(E_3|H_1) = 0.25$, $\mathrm{P}(E_3|H_0) = 0.5$

Posterior odds ratio

$\frac{\mathrm{P}(H_0|E)}{\mathrm{P}(H_1|E)} = \frac{\pi_0\mathrm{P}(E|H_0)}{\pi_1\mathrm{P}(E|H_1)} = \frac{\pi_0\mathrm{P}(E_1|H_0)\mathrm{P}(E_2|H_0)\mathrm{P}(E_3|H_0)}{\pi_1\mathrm{P}(E_1|H_1)\mathrm{P}(E_2|H_1)\mathrm{P}(E_3|H_1)} = 0.018$.



Reject H_0 if

$\frac{c_1}{c_0} = \frac{\text{cost for unpunished crime}}{\text{cost for punishing an innocent}} > 0.018$.

Prosecutor's fallacy: $\mathrm{P}(H_0|E) = \mathrm{P}(E|H_0)$, which is only true if $\mathrm{P}(E) = \pi_0$.
Example: $\pi_0 = \pi_1 = 1/2$, $\mathrm{P}(E|H_0) \approx 0$, $\mathrm{P}(E|H_1) \approx 1$.