Serik Sagitov, Chalmers Tekniska Högskola, October 20, 2004

Ch 15. Decision theory and Bayesian inference

1. Minimax

Decision rule d

choose an action $a = d(x)$

using uncertain measurement X with pmf/pdf $f(x|\theta)$

of unknown state of nature θ

Loss function $l(\theta, a)$ determines

risk function $R(\theta, d) = \mathrm{E}(l(\theta, d(X))|\theta)$

Minimax decision rule

for each decision rule d compute $\max_\theta R(\theta, d)$

choose d minimizing $\max_\theta R(\theta, d)$

Ex 1: steel section length

Two possible actions

steel section length $a = 40$ or $a = 50$ ft

Two possible states of nature

depth of a firm stratum $\theta = 40$ or $\theta = 50$ ft

Data $x = 45$ ft, uncertain measurement of θ

conditional pmf $f(x|\theta) = \mathrm{P}(X = x|\theta)$

	$x = 40$	$x = 45$	$x = 50$	total
θ=40	0.6	0.3	0.1	1
θ=50	0.1	0.2	0.7	1

$f(x|\theta)$ is the likelihood for fixed x and variable θ

Four possible decision rules

	$x=40$	$x=45$	$x=50$
$d_1(x)$	40	40	40
$d_2(x)$	40	40	50
$d_3(x)$	40	50	50
$d_4(x)$	50	50	50

Conditional distributions of $d_2(X)$ and $d_3(X)$

	$\mathrm{P}_\theta(d_2=40)$	$\mathrm{P}_\theta(d_2=50)$	$\mathrm{P}_\theta(d_3=40)$	$\mathrm{P}_\theta(d_3=50)$
θ=40	0.9	0.1	0.6	0.4
θ=50	0.3	0.7	0.1	0.9

Loss function $l(\theta, a)$

$l(40,40) = l(50,50) = 0$

$l(40,50) = \$100$, $l(50,40) = \$400$

Risk function $R(\theta, d) = \mathrm{E}(l(\theta, d(X))|\theta)$

	$d=d_1$	d_2	d_3	d_4	minimax
$\theta=40$	0	10	40	100	
$\theta=50$	400	120	40	0	
max risk	400	120	40	100	d_3

Minimax action: $d_3(45) = 50$ ft

2. Bayesian approach

Population parameter θ is a random variable Θ

prior distribution - knowledge before new measurement

Posterior distribution

reflects knowledge updated after measurement using Bayes formula $\mathrm{P}(\Theta=\theta|X=x) = \frac{\mathrm{P}(X=x|\Theta=\theta)\mathrm{P}(\Theta=\theta)}{\mathrm{P}(X=x)}$

Prior distribution of Θ $\quad g(\theta) = \mathrm{P}(\Theta = \theta)$

Joint distribution of (X, Θ)

$f(x, \theta) = \mathrm{P}(X = x, \Theta = \theta) = f(x|\theta)g(\theta)$

Marginal distribution of X

$\phi(x) = \mathrm{P}(X = x) = \sum_\theta f(x, \theta)$

Posterior distribution of Θ

$h(\theta|x) = \frac{1}{\phi(x)} f(x|\theta)g(\theta)$

$$\boxed{\text{Posterior} \propto \text{likelihood} \times \text{prior}}$$

Bayes action minimizes posterior risk

$$PR(a|x) = \mathrm{E}(l(\Theta, a)|X = x) = \sum_\theta l(\theta, a)h(\theta|x)$$

Ex 1: steel section length

Given the prior probabilities $g(40) = 0.8$ and $g(50) = 0.2$

joint distribution / posterior distribution

$f(x,\theta)$	$x = 40$	45	50	$h(\theta\|x)$	$x = 40$	45	50
$\theta = 40$	0.48	0.24	0.08	$\theta = 40$	0.96	0.86	0.36
$\theta = 50$	0.02	0.04	0.14	$\theta = 50$	0.04	0.14	0.64
$\phi(x)$	0.50	0.28	0.22	Total	1.00	1.00	1.00

Posterior risk

$PR(a|x) = l(40, a) \cdot h(40|x) + l(50, a) \cdot h(50|x)$

	$a = 40$	$a = 50$	min PR	Bayes action
$x = 40$	16	96	16	$a = 40$
$x = 45$	56	86	56	$a = 40$
$x = 50$	256	36	36	$a = 50$

3. 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-data give a G-posterior

Data distribution	Prior	Posterior distribution
$X \sim \mathrm{N}(\mu, \sigma^2)$	$\mu \sim \mathrm{N}(\mu_0, \sigma_0^2)$	$\mathrm{N}(c_1\mu_0 + (1-c_1)x; c_1\sigma_0^2)$
$X \sim \mathrm{Bin}(n, p)$	$p \sim \mathrm{B}(a, b)$	$\mathrm{B}(a+x, b+n-x)$
$\mathrm{Mn}(n; p_1, \ldots, p_r)$	$\mathrm{D}(\alpha_1, \ldots, \alpha_r)$	$\mathrm{D}(\alpha_1 + x_1, \ldots, \alpha_r + x_r)$
$X \sim \mathrm{Pois}(\mu)$	$\mu \sim \Gamma(\alpha, \lambda)$	$\Gamma(\alpha + x, \lambda + 1)$
$X \sim \mathrm{Exp}(\rho)$	$\rho \sim \Gamma(\alpha, \lambda)$	$\Gamma(\alpha + 1, \lambda + x)$

Normal/normal model with $n = 1$ observation

shrinkage factor $c_1 = \frac{\sigma^2}{\sigma_0^2 + \sigma^2}$

Beta distribution $\mathrm{B}(a, b)$

pdf $f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}$, $0 < p < 1$

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

Dirichlet distribution $\mathrm{D}(\alpha_1, \ldots, \alpha_r)$

pdf $f(p_1, \ldots, p_r) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_r)} p_1^{\alpha_1 - 1} \ldots p_r^{\alpha_r - 1}$

$p_1 + \ldots + p_r = 1$

positive pseudocounts $\alpha_1, \ldots, \alpha_r$, $\alpha_0 = \alpha_1 + \ldots + \alpha_r$

Marginal distr. $p_j \sim \mathrm{Beta}(\alpha_j, \alpha_0 - \alpha_j)$, $j = 1, \ldots, r$

$\mathrm{Cov}(p_1, p_2) = -\frac{\alpha_1\alpha_2}{\alpha_0^2(\alpha_0 + 1)}$

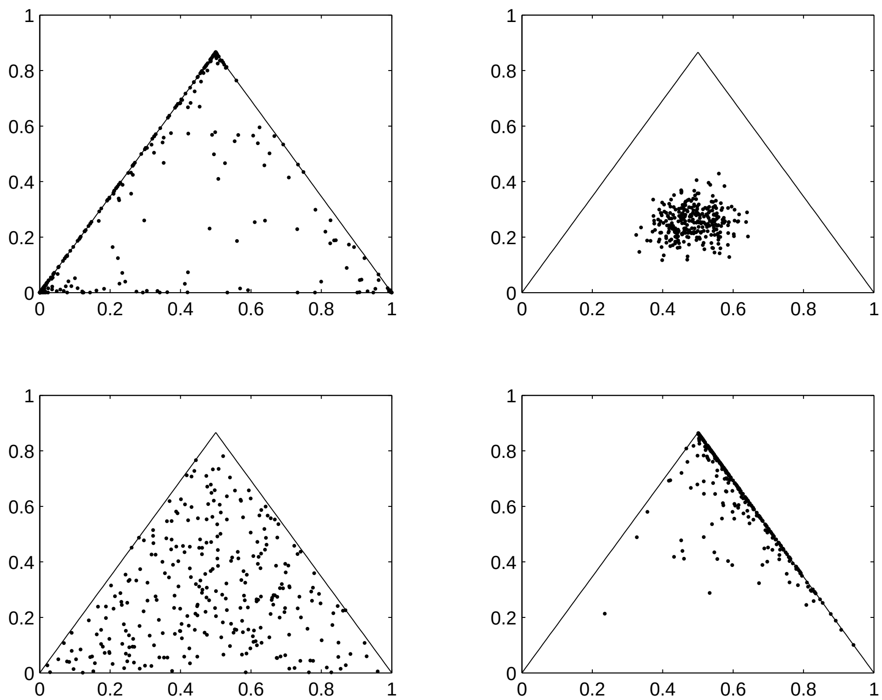

Figure 1: *Dirichlet distributiton profiles. Four scatter plots each depicting* $n = 300$ *points generated using different sets of parameters: upper left* $(0.3, 0.3, 0.1)$, *upper right (13,16,15), lower left (1,1,1), lower right (3,0.1,1). Here we use a triangle picture of three-dimentional Dirichlet distributiton with the first parameter controlling the distance to the bottom edge of the triangle, the second parameter controlling the distance to the right edge of the triangle, and the third parameter controlling the distance to the left edge of the triangle.*

Ex 2: IQ measurement

IQ distribution of a person $X \sim \mathrm{N}(\theta,\, 100)$

prior distr $\theta \sim \mathrm{N}(100{,}225)$, population as a whole

If observed IQ is $x = 130$, then

posterior distribution $\theta \sim \mathrm{N}(120.7,\, 69.2)$

4. Bayesian updating

Normal/normal model n observations

$$\begin{aligned}(\mu_0, \sigma_0^2) &\overset{x_1}{\to} (c_1\mu_0 + (1-c_1)x_1; c_1\sigma_0^2) \overset{x_2}{\to} \ldots \\ &\overset{x_n}{\to} (c_n\mu_0 + (1-c_n)\bar{x}; c_n\sigma_0^2)\end{aligned}$$

shrinkage factor $c_n = \frac{\sigma^2}{\sigma^2 + n\sigma_0^2} \to 0$ as $n \to \infty$

Ex 3: thumbtack experiment

Beta/binomial model

number of base landings $X \sim \mathrm{Bin}(n, p)$

n tossings, $p = \mathrm{P}$(landing on base)

My personal Beta prior $p \sim \mathrm{B}(a_0, b_0)$

$\mu_0 \approx 0.4$, $\sigma_0 \approx 0.1 \Rightarrow$ pseudocounts $a_0 = 10$, $b_0 = 15$

Experiment 1: $n_1 = 10$ tosses

counts $x_1 = 2$, $n_1 - x_1 = 8$, posterior distr $\mathrm{B}(12,\, 23)$

PME $\hat{p} = \frac{12}{35} = 0.34$, $\sigma_1 = 0.08$

Experiment 2: $n_2 = 40$ tosses

counts $x_2 = 9$, $n_2 - x_2 = 31$, posterior distr $\mathrm{B}(21,\, 54)$

PME $\hat{p} = \frac{21}{75} = 0.28$, $\sigma_2 = 0.05$

5. Bayesian estimation

Action $a = \{\text{assign value } a \text{ to unknown parameter } \theta\}$

optimal action depends on the choice of loss function

MAP = max a posteriori probability

$$\boxed{\text{0-1 loss function: } l(\theta, a) = 1_{\{\theta \neq a\}}}$$

Posterior risk = probability of misclassification

$PR(a|x) = \sum_{\theta \neq a} h(\theta|x) = 1 - h(a|x)$

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

If noninformative prior $g(\theta) = \text{const}$, then

$h(\theta|x) \propto f(x|\theta)$ and $\hat{\theta}_{\text{map}} = \hat{\theta}_{\text{mle}}$

PME = posterior mean estimate

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

$PR(a|x) = \mathrm{E}((\Theta - a)^2|x) = \mathrm{Var}(\Theta|x) + [\mathrm{E}(\Theta|x) - a]^2$

$\hat{\theta}_{\text{pme}} = \mathrm{E}(\Theta|X = x)$

Ex 4: loaded die experiment

a possibly loaded die

is rolled 18 times: 211 453 324 142 343 515

If the prior distribution is D(1,1,1,1,1,1), then

MAP = MLE = sample proportions $(\frac{4}{18}, \frac{3}{18}, \frac{4}{18}, \frac{4}{18}, \frac{3}{18}, 0)$

PME

$\hat{p}_1 = \frac{5}{24} = 0.208$, $\hat{p}_2 = \frac{4}{24} = 0.167$, $\hat{p}_3 = \frac{5}{24} = 0.208$

$\hat{p}_4 = \frac{5}{24} = 0.208$, $\hat{p}_5 = \frac{4}{24} = 0.167$, $\hat{p}_6 = \frac{1}{24} = 0.042$

6. Interval estimation

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

$\mathrm{P}(\theta_0(x) < \Theta < \theta_1(x) | X = x) = 1 - \alpha$

Ex 2: IQ measurement

$n = 1$, $\bar{X} \sim \mathrm{N}(\mu; 100)$

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 $120.7 \pm 1.96 \cdot \sqrt{69.2} = 120.7 \pm 16.3$

7. 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

Rejection region minimizing the average cost

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

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 $\frac{h(\theta_0|x)}{h(\theta_1|x)} < \frac{l_{\mathrm{II}}}{l_{\mathrm{I}}}$

Ex 5: a rape case study

http://www.law.umich.edu/thayer/redmay.htm

The defendant A, age 37, local, is charged with rape

H_0: A is innocent, H_1: A is guilty

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

prior probability $\mathrm{P}(H_1) = \frac{1}{200{,}000}$

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

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

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 $\frac{l_{\mathrm{II}}}{l_{\mathrm{I}}} > 0.018$

error I: a nonguilty is convicted

error II: a guilty is unpunished

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