Serik Sagitov, Chalmers and GU, March 14, 2016

Introduction to Bayesian inference

1 Bayesian approach

The main idea of the Baysian approach is to treat the population parameter θ as a random variable, where the source of randomness is the luck of knowledge. Two distributions of θ

prior distribution density $g(\theta)$ brings into the model the knowledge on θ before data is collected,
posterior distribution $h(\theta|x)$ updates the knowledge on θ using the collected data x.

$\boxed{\text{Bayes formula } h(\theta|x) = \frac{f(x|\theta)g(\theta)}{\phi(x)}}$ $\boxed{\text{Posterior} \propto \text{likelihood} \times \text{prior}, \propto \text{means proportional.}}$

Marginal distribution of the data X has density $\phi(x) = \int f(x|\theta)g(\theta)d\theta$. For a given x, the constant $\phi(x)$ is the likelihood $f(x|\theta)$ of the data value x averaged over different values of θ using the prior distribution.

Uninformative prior: when we have no prior knowledge of θ, the prior distribution is often modelled by the uniform distribution. In the uniform case, since $g(\theta) \propto$ constant, we have $h(\theta|x) \propto f(x|\theta)$ so that all the posterior knowledge comes from the likelihood function.

Example (IQ measurement)
A randomly chosen individual has an unknown true intelligence quotient value θ. Its prior distribution is $\theta \sim \mathrm{N}(100, 225)$. This normal distribution describes the whole population with mean IQ of $m = 100$ and standard deviation $v = 15$.
Given a true personal value θ, the result of an IQ measurement has distribution $X \sim \mathrm{N}(\theta, 100)$, with no systematic error and a random error $\sigma = 10$. Since

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

and the posterior is proportional to $g(\theta)f(x|\theta)$, we find that $h(\theta|x)$ is proportional to

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

where $\gamma = \frac{\sigma^2}{\sigma^2+v^2}$ is the so-called shrinkage factor. We conclude that the posterior distribution is normal $h(\theta|x) = \frac{1}{\sqrt{2\pi\gamma}v}e^{-\frac{(\theta-\gamma m-(1-\gamma)x)^2}{2\gamma v^2}}$ with mean $\gamma m + (1-\gamma)x$ and variance γv^2.

Suppose that the observed IQ result is $x = 130$, then the posterior distribution becomes $\mathrm{N}(120.7, 69.2)$. We see that the prior expectation $m = 100$ has corrected the observed result $x = 130$ down to 120.7. The posterior variance 69.2 is smaller than that of the prior distribution 225 by the shrinkage factor $\gamma = 0.308$: the updated knowledge is less uncertain than the prior knowledge.

2 Conjugate priors

Suppose we have two parametric families of probability distributions $\mathcal{G}$ and $\mathcal{H}$.

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

Beta distribution $\mathrm{Beta}(a, b)$
has density, mean, and variance

$$f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}, \quad 0 < p < 1, \qquad \mu = \frac{a}{a+b}, \quad \sigma^2 = \frac{\mu(1-\mu)}{a+b+1}.$$

Parameters $a > 0$, $b > 0$ determining the shape of the distribution are called pseudo-counts. Uniform distribution is obtained with $a = b = 1$.
Exercise: verify that for given $a > 1$ and $b > 1$, the maximum of density function $f(p)$ is attained at

$$\hat{p} = \frac{a-1}{a+b-2}.$$

Dirichlet distribution $\mathrm{Dir}(\alpha_1, \ldots, \alpha_r)$
has 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 pseudo-counts $\alpha_1, \ldots, \alpha_r$, $\alpha_0 = \alpha_1 + \ldots + \alpha_r$.
Dirichlet distribution is a multivariate extension of the beta distribution
marginal distributions $p_j \sim \mathrm{Beta}(\alpha_j, \alpha_0 - \alpha_j)$, $j = 1, \ldots, r$,
negative covariances $\mathrm{Cov}(p_1, p_2) = -\frac{\alpha_1\alpha_2}{\alpha_0^2(\alpha_0+1)}$.

List of conjugate prior models

Data distribution	Prior	Posterior distribution	Comments
$(X_1, \ldots, X_n), X_i \sim \mathrm{N}(\theta, \sigma^2)$	$\mu \sim \mathrm{N}(m, v^2)$	$\mathrm{N}(\gamma_n m + (1-\gamma_n)\bar{x}; \gamma_n v^2)$	(1), (3), (4)
$X \sim \mathrm{Bin}(n, p)$	$p \sim \mathrm{Beta}(a, b)$	$\mathrm{Beta}(a + x, b + n - x)$	(2), (3), (4)
$(X_1, \ldots, X_r) \sim \mathrm{Mn}(n; p_1, \ldots, p_r)$	$\mathrm{Dir}(\alpha_1, \ldots, \alpha_r)$	$\mathrm{D}(\alpha_1 + x_1, \ldots, \alpha_r + x_r)$	(2), (3), (4)
$X \sim \mathrm{Pois}(\mu)$	$\mu \sim \Gamma(\alpha, \lambda)$	$\Gamma(\alpha + x, \lambda + 1)$	(3), (4)
$X \sim \mathrm{Exp}(\rho)$	$\rho \sim \Gamma(\alpha, \lambda)$	$\Gamma(\alpha + 1, \lambda + x)$	(3), (4)

(1) the shrinkage factor for n measurements is $\gamma_n = \frac{\sigma^2}{\sigma^2 + nv^2}$
(2) the update rule: posterior pseudo-counts = prior pseudo-counts plus sample counts
(3) posterior variance is always smaller than the prior variance
(4) the contribution of the prior distribution becomes smaller for larger samples

Example (beta-binomial model)
Consider the probability p of a thumbtack landing on its base. Uninformative prior for p: the uniform over [0,1] distribution. Data: the number of base landings $X \sim \mathrm{Bin}(n, p)$ for n tossings of the thumbtack.
Experiment 1: $n_1 = 10$ tosses, counts $x_1 = 2$, $n_1 - x_1 = 8$, prior distribution $\mathrm{Beta}(1, 1)$ with mean $\mu_0 = 0,5$ and standard deviation $\sigma_0 = 0.29$, posterior distribution Beta(3, 9) with mean $\hat{p} = \frac{3}{12} = 0.25$ and standard deviation $\sigma_1 = 0.12$.
Experiment 2: $n_2 = 40$ tosses, counts $x_2 = 9$, $n_2 - x_2 = 31$, prior distribution Beta(3, 9), posterior distribution Beta(12, 40) with mean $\hat{p} = \frac{12}{52} = 0.23$ and standard deviation $\sigma_2 = 0.06$.

3 Bayesian estimation

In the language of decision theory we are searching for an optimal action
{assign value a to unknown parameter θ}.
The optimal a depends on the choice of the loss function $l(\theta, a)$. Bayes action minimises posterior risk

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

We consider two loss functions leading to two Bayesian estimators.

$\boxed{\text{Zero-one loss function: } l(\theta, a) = 1_{\{\theta \neq a\}}}$ $\boxed{\text{Squared error loss: } l(\theta, a) = (\theta - a)^2}$

MAP (maximum a posteriori probability)
Using the zero-one loss function we find that the posterior risk is the probability of misclassification
$R(a|x) = \sum_{\theta \neq a} h(\theta|x) = 1 - h(a|x)$.
To minimise the risk we have to maximise the posterior probability: define $\hat{\theta}_{\text{map}}$ as the value of θ that maximises $h(\theta|x)$. With the uninformative prior, $\hat{\theta}_{\text{map}} = \hat{\theta}_{\text{mle}}$.

PME (posterior mean estimate)
Using the squared error loss function we find that the posterior risk is a sum of two components
$R(a|x) = \mathrm{E}((\theta - a)^2|x) = \mathrm{Var}(\theta|x) + [\mathrm{E}(\theta|x) - a]^2$.
We minimise the posterior risk by putting $\hat{\theta}_{\text{pme}} = \mathrm{E}(\theta|x)$.

Example (loaded die experiment)
A possibly loaded die is rolled 18 times, 211 453 324 142 343 515. Parameter of interest $\theta = (p_1, \ldots, p_6)$.
Take the uninformative prior distribution Dir(1,1,1,1,1,1) and compare two Bayesian estimates
$\hat{\theta}_{\text{map}} = \hat{\theta}_{\text{mle}} = (\frac{4}{18}, \frac{3}{18}, \frac{4}{18}, \frac{4}{18}, \frac{3}{18}, 0)$ is based only on the sample counts,
$\hat{\theta}_{\text{pme}} = (\frac{5}{24}, \frac{4}{24}, \frac{5}{24}, \frac{5}{24}, \frac{4}{24}, \frac{1}{24})$ uses pseudo-counts.
Observe that the maximum likelihood estimate assigns value zero to p_6, thereby excluding sixes in future observations.

4 Credibility interval

Confidence interval formulas: θ is an unknown constant and a the confidence interval is random
$\mathrm{P}(\theta_0(X) < \theta < \theta_1(X)) = 1 - \alpha$.
A credibility interval (CrI) is treated as a nonrandom interval while θ is a random variable. A CrI 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 Bayesian hypotheses testing

We consider the case of two simple hypotheses. Choose between H_0: $\theta = \theta_0$ and H_1: $\theta = \theta_1$ using not only the likelihoods of the data $f(x|\theta_0)$, $f(x|\theta_1)$ but also prior probabilities $\mathrm{P}(H_0) = \pi_0$, $\mathrm{P}(H_1) = \pi_1$. The rejection region $\mathcal{R}$ for the data X is found in terms of a cost function:

		Decision	H_0 true	H_1 true
Cost values	$X \notin \mathcal{R}$	Accept H_0	0	c_1
	$X \in \mathcal{R}$	Accept H_1	c_0	0

For a given set $\mathcal{R}$, the average cost is the weighted mean of two values c_0 and c_1

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

It follows that the rejection region minimising the average cost is $\mathcal{R} = \{x : c_0\pi_0 f(x|\theta_0) < c_1\pi_1 f(x|\theta_1)\}$. The optimal decision rule:

reject H_0 for small values of the likelihood ratio $\frac{f(x|\theta_0)}{f(x|\theta_1)} < \frac{c_1\pi_1}{c_0\pi_0}$,

or in other terms, for small posterior odds $\frac{h(\theta_0|x)}{h(\theta_1|x)} < \frac{c_1}{c_0}$.

Example (rape - a case study)
The defendant A, age 37, local, is charged with rape.
The jury have to choose between two alternative hypotheses H_0: A is innocent, H_1: A is guilty.

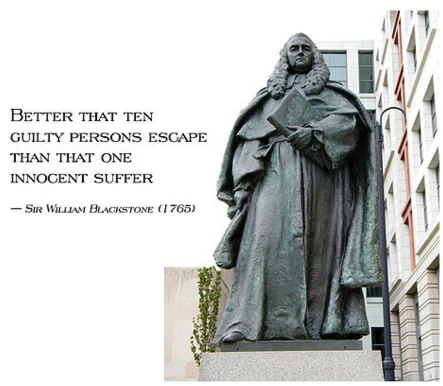
Uninformative prior probability $\pi_1 = \frac{1}{200{,}000}$. Prior to the evidence is taken into account any of 200 000 males in the appropriate group could be guilty.

Three pieces of evidence which are conditionally independent

E_1: strong DNA match, $\mathrm{P}(E_1|H_0) = \frac{1}{200{,}000{,}000}$, $\mathrm{P}(E_1|H_1)$=1,

E_2: defendant A is not recognised by the victim,

E_3: an 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 $\quad \frac{c_1}{c_0} = \frac{\text{cost for unpunished crime}}{\text{cost for punishing an innocent}} > 0.018.$

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