Bayesian Statistics for Computer intensive methods sf2955

Timo Koski

April 29, 2009

KTH Matematik

Contents

(a) Statistical Inference
(b) Parametric statistical model
(c) Bayes rule, Posterior Distributions
(d) Bayes' Billiard Balls
(e) Predictive Distributions
(f) Asymptotic form of the posterior density

KTH Matematik

Learning/inference from data

By learning/inference from data one often means the process of inferring a general law or principle from the observations of particular instances. The general law is a piece of knowledge about the mechanism of nature that generates the data.

KTH Matematik

Learning from data

The intended learning is done by use of 'MODELS', which serve as the language in which the constraints predicated on the data can be described. We deal here with parametric statistical models.

Parametric statistical model

x is an observation of a random variable (X).

$$x \in f\left(x|\theta\right)$$

$f\left(x|\theta\right)$ is a probability density on R^p. $f\left(x|\theta\right)$ is a known function of x and θ.

θ is an unknown parameter.

KTH Matematik

Parametric statistical model

x can be of the form $x^{(n)} = (x_1, \ldots, x_n)$. x can be continuous or discrete variate or a mix thereof.
θ is an unknown parameter $\in \Theta =$ a vector space of finite dimension.
Hence we exclude, e.g., non-parametric statistics.

KTH VETENSKAP OCH KONST
KTH Matematik

Parametric statistical model

$$x \in f\left(x|\theta\right)$$

- x is distributed according to f,
- x is an observation from the distribution f.

An outcome x of a random variable (r.v.) X.

Parametric statistical model: Examples; Normal distribution

$$\theta = \left(\mu, \sigma^2\right) \in \Theta = R \times (0, \infty) \ .$$

$$f\left(x|\theta\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{\sigma^2}(x-\mu)^2}, -\infty < x < \infty.$$

We say that x is an observation from the normal distribution $N\left(\mu, \sigma^2\right)$.

KTH Matematik

Parametric statistical model: Examples; Bernoulli Distribution

Consider r.v. X with values $0, 1$, $0 < \theta < 1$ and

	$x = 1$	$x = 0$
$f(x \mid \theta)$	θ	$1 - \theta$

then we say x is distributed according to the Bernoulli distribution with the parameter θ.

$$X = x \in Be(\theta),$$

KTH Matematik

Parametric statistical model

$f(x|\theta)$ is a probabilistic mechanism of generating data, characterizes the behaviour of future observations conditional on θ.

KTH Matematik

Learning/inference from data: Inversion

Retrieve the parameters of the probabilistic generating mechanism using x. $f(x|\theta)$ is a probabilistic generating mechanism of data, characterizes the behaviour of future observations conditional on θ, but in inference the roles of x and θ are inverted.

KTH Matematik

Inversion and Bayes' Rule

Since

$$p(A \mid E) \cdot p(E) = p(E \mid A) \cdot p(A)$$

we have in a formal way Bayes' Rule of inversion

$$p(A \mid E) = \frac{p(E \mid A) \cdot p(A)}{p(E)}.$$

$$p(E) = p(E \mid A)p(A) + p\left(E \mid A^c\right) p\left(A^c\right).$$

A^c is the complement set of A.

KTH VETENSKAP OCH KONST
KTH Matematik

Bayes' rule

Bayes' rule extended to continuous random variables:

$$g\left(y|x\right) = \frac{f\left(x|y\right) \cdot g\left(y\right)}{\int f\left(x|y\right) \cdot g\left(y\right) dy},$$

Due to the standardization $g\left(y|x\right)$ is a probability density; $g\left(y|x\right) \geq 0$, $\int g\left(y|x\right) dy = 1$.

Bayes' rule: parametric model

$$\pi(\theta|x) = \frac{f(x \mid \theta) \cdot \pi(\theta)}{\int_\Theta f(x \mid \theta) \cdot \pi(\theta)\, d\theta}$$

Terminology for Bayes' Rule:

- $\pi(\theta)$: **prior distribution** on Θ.
- $\pi(\theta|x)$: **posterior distribution** on Θ.
- $m(x) = \int_\Theta f(x \mid \theta) \cdot \pi(\theta)\, d\theta$: **marginal distribution** of x.

KTH VETENSKAP OCH KONST
KTH Matematik

Uncertainty

Uncertainty about the unknown θ is modeled by a probability distribution $\pi(\theta)$, and $\pi(\theta|x)$ expresses the uncertainty about the unknown θ after the observation of x.
We use probability as tool for all parts of our analysis. This is called coherence.
Mathematically: the unknown θ becomes a random variable. (x, θ) will have a joint distribution.

KTH Matematik

Distributions

$$\pi(\theta|x) = \frac{f(x \mid \theta) \cdot \pi(\theta)}{m(x)},$$

Terminology:

- $m(x) = \int_\Theta f(x \mid \theta) \cdot \pi(\theta)\, d\theta$
- $\phi(x, \theta)$: **joint distribution** of (x, θ).
-

$$\pi(\theta|x)\, m(x) = \phi(x, \theta) = f(x|\theta)\, \pi(\theta)$$

KTH VETENSKAP OCH KONST
KTH Matematik

Notation

The notation

$$\int_{\Theta} f\left(x \mid \theta\right) \cdot \pi\left(\theta\right) d\theta$$

is imprecise by intent, as it can mean both a single integral and a multiple integral.

KTH Matematik

Bayesian Parametric Statistical Model

A Bayesian parametric statistical model consists of

- a parametric model

$$x \in f\left(x|\theta\right)$$

- a prior distribution

$$\theta \in \pi(\theta)$$

The quantity of interest

$$\theta|x \in \pi\left(\theta|x\right)$$

KTH Matematik

Prior Density

Any function $\pi(\cdot)$ such that

$$\pi(\theta) \geq 0,$$

and

$$\int_{\Theta} \pi(\theta)\, d\theta = 1,$$

can serve as a prior distribution.

KTH VETENSKAP OCH KONST
KTH Matematik

Improper Prior Densities

But even functions with the properties

$$\pi(\theta) \geq 0,$$

and

$$\int_{\Theta} \pi(\theta)\, d\theta = \infty,$$

are also invoked as priors, and are called improper priors.

KTH VETENSKAP OCH KONST
KTH Matematik

An Example

$X_i \mid M = m \in \mathrm{N}\left(m, \sigma_0^2\right)$, $M \in \mathrm{N}\left(\mu, s^2\right)$. $x^{(n)} = (x_1, \ldots, x_n)$ a sample of I.I.D. X_i, $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

$$M \mid (X_1, \ldots, X_n) \in \mathrm{N}\left(\frac{n\overline{x}/\sigma_0^2 + \mu/s^2}{n/\sigma_0^2 + 1/s^2}, \frac{1}{n/\sigma_0^2 + 1/s^2}\right)$$

i.e., $\pi\left(m|x^{(n)}\right)$ is the density of this normal distribution. Here μ and s^2 are *hyperparameters*.

KTH Matematik

Confidence interval

- $P\left(a(x) \leq \theta \leq b(x)\right) = \int_{a(x)}^{b(x)} \pi\left(\theta|x\right) d\theta$
- $\underbrace{P\left(a(x) \leq \theta \leq b(x)\right)}$

  This is a probability, not a degree of confidence

Confidence interval: Example

Take the example above

$$\mathrm{N}\left(\frac{n\overline{x}/\sigma_0^2+\mu/s^2}{n/\sigma_0^2+1/s^2},\frac{1}{n/\sigma_0^2+1/s^2}\right)$$

Let $s\to\infty$ (the prior becomes improper). Then

$$\mathrm{N}\left(\frac{n\overline{x}/\sigma_0^2+\mu/s^2}{n/\sigma_0^2+1/s^2},\frac{1}{n/\sigma_0^2+1/s^2}\right)\to\mathrm{N}\left(\overline{x},\frac{\sigma_0^2}{n}\right)$$

But then the Bayesian confidence interval $P\left(a(x)\leq\theta\leq b(x)\right)=0.95$ becomes the familiar $\overline{x}\pm\lambda_{0.025}\frac{\sigma_0}{\sqrt{n}}$, and the common mistaken but natural interpretation of the confidence interval is correct !

KTH VETENSKAP OCH KONST
KTH Matematik

General Aspects of Bayesian Inference

- inference is based on the observed x, not on an unobserved sample space.
- $\pi(\theta|x)$ is the only quantity evaluated for inference about θ.

On the other hand , evaluation of $\pi(\theta|x)$ is not in general possible by explicit means of integral calculus. This is where statistical inference needs Markov chain Monte Carlo (McMC).

KTH VETENSKAP OCH KONST
KTH Matematik

Likelihood

The distribution $f(x \mid \theta)$ regarded as a function of θ is known as the *likelihood function*

$$l(x;\theta) = f(x \mid \theta).$$

The likelihood function $l(\theta|x)$ thus compares the plausibilities of different parameter values for given x.

KTH Matematik

Log likelihood

$$ll(x;\theta) = -\log l(x;\theta).$$

is called the log likelihood function.

Likelihood Principle

The information brought by x about θ is entirely contained in the likelihood function $l(\theta \mid x)$.
If x_1 and x_2 are two observations depending on the same parameter θ such that there exists a constant c such that

$$l_1(x_1;\theta) = c \cdot l_2(x_2;\theta)$$

for every θ, then they bring the same information about θ and must lead to identical inferences.

KTH VETENSKAP OCH KONST
KTH Matematik

Likelihood Principle

Conditions for the likelihood principle:

- inference should be about the same parameter set
- θ should include every unknown factor in the model.

KTH Matematik

Bayes' rule & likelihood

$$\pi(\theta|x) = \frac{f(x \mid \theta) \cdot \pi(\theta)}{\int_\Theta f(x \mid \theta) \cdot \pi(\theta)\, d\theta'}$$

$$= \frac{l(x;\theta) \cdot \pi(\theta)}{\int_\Theta l(x \mid \theta) \cdot \pi(\theta)\, d\theta}$$

Hence Likelihood Principle is satisfied by Bayesian inference. There are ways of implementing the likelihood principle: MLE and MAP $\Rightarrow$

KTH VETENSKAP OCH KONST
KTH Matematik

The Maximum Likelihood Estimate (MLE)

The **maximum likelihood estimate** MLE, $\widehat{\theta}_{\mathrm{ML}}$ of θ, is defined by

$$\widehat{\theta}_{\mathrm{ML}} = \mathrm{argmax}_{\theta \in \Theta} f\left(x \mid \theta\right)$$

$$= \mathrm{argmin}_{\theta \in \Theta} ll\left(x; \theta\right)$$

MLE *is a parameter value that gives the observed* x *the highest possible probability*.

KTH Matematik

The Maximum A Posterior Estimate (MAP)

The **maximum a posterior estimate** MAP $\widehat{\theta}_{\mathrm{MAP}}$ of θ is defined by

$$\widehat{\theta}_{\mathrm{MAP}} = \mathrm{argmax}_{\theta \in \Theta} \pi\left(\theta \mid x\right)$$

KTH Matematik

Definition: Conjugate Family of Priors

A family $\mathcal{F}$ of probability distributions on Θ is said to be **conjugate** or **closed under sampling** for a likelihood function

$$l(x;\theta) = f(x \mid \theta).$$

if for every $\pi \in \mathcal{F}$, the posterior distribution $\pi(\theta|x)$ also belongs to $\pi \in \mathcal{F}$. □

Above we have seen an example with normal prior density on the mean.

KTH VETENSKAP OCH KONST
KTH Matematik

Definition: Conjugate Family of Priors

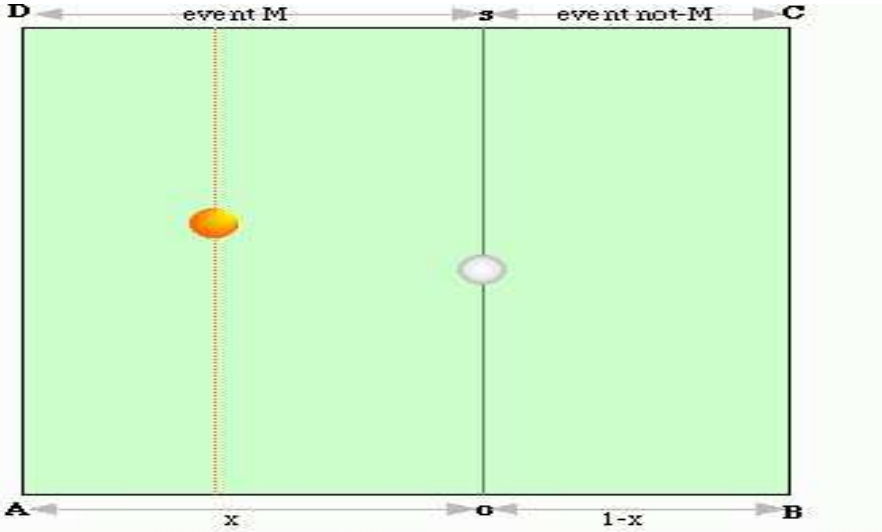
An intuitive way of understanding conjugate priors is that with conjugate priors the prior knowledge can be translated into equivalent sample information. See, e.g.,

$$\mathrm{N}\left(\frac{n\overline{x}/\sigma_0^2+\mu/s^2}{n/\sigma_0^2+1/s^2}, \frac{1}{n/\sigma_0^2+1/s^2}\right).$$

Next we reconsider another problem with a conjugate family of priors.

KTH VETENSKAP OCH KONST
KTH Matematik

Bayes' Billiard Ball



The square table is made in such a way that, if the White or Orange ball be thrown on it, the probability that it rests on any part of the plane is the same. First, the White ball is thrown, and suppose it rests on line **os**; then the Orange ball is thrown n times. We define event M as any event the Orange ball rests between **A** and **o**, and not-M as its resting between **o** and **B**. What this means is this:

The first throw of the White ball determines the value of probability x (i.e. the probability of an unknown event) from a uniform distribution between 0 and 1; and then a series of trials, with probability x of success (i.e., M) is generated (this provides the data on which to infer the correct value of x).

Then, thanks to the geometrical representation of the problem in the Figure, we can obtain the solution to the initial problem, by calculating integration. Although we have omitted mathematical formulas, the preceding is the central idea.

KTH VETENSKAP OCH KONST
KTH Matematik

Bayes' Billiard Ball

A billiard ball W is rolled on a line of length one, with a uniform probability of stopping anywhere. It stops at p, not disclosed to us. A second ball O is rolled n times under the same assumptions and X denotes the number of times O stops to the left of W. Given $X = x$, what inference can we make on p ?
(In the figure above $x \leftrightarrow p$.)

KTH Matematik

Modeling and Learning for Bayes' Billiard Ball

We let $\mathbf{P}$ be a random variable, whose values are denoted by p, $0 \leq p \leq 1$.
Parametric statistical model for rolls of Bayes' Billiard Ball O:
Conditional on $\mathbf{P} = p$, the rolls are outcomes of I.I.D $\mathrm{Be}(p)$ R.V's.

KTH VETENSKAP OCH KONST
KTH Matematik

Modeling and Learning for Bayes' Billiard Ball

Hence for $x = 0, 1, 2, \ldots, n$,

$$f(x|p) = P(X = x \mid \mathbf{P} = p)$$

$$= \left(\begin{array}{c} n \\ x \end{array} \right) p^x \cdot (1-p)^{n-x},$$

(the Binomial distribution)

KTH VETENSKAP OCH KONST
KTH Matematik

The Posterior Density

Bayes' rule

$$\pi\left(p \mid x\right) = \frac{f\left(x \mid p\right) \cdot \pi\left(p\right)}{\int_0^1 f\left(x \mid p\right) \cdot \pi\left(p\right) dp}, 0 \leq p \leq 1$$

and zero elsewhere. The marginal distribution of x is

$$m(x) = \int_0^1 f\left(x \mid p\right) \cdot \pi\left(p\right) dp.$$

KTH Matematik

The Posterior Density

The posterior $\pi(p \mid x)$ expresses our updated uncertainty of the 'true' position of W given the data $X = x$.

KTH Matematik

The Posterior Density

One way to get further from here is to use an explicit expression for $\pi(p)$. There are many possible choices (some more systematic choices outlined below), some have straightforward analytical advantages. Laplace assumed that $p \in \mathrm{R}(0,1)$. i.e.,

$$\pi(p) = \begin{cases} 1 & 0 \leq p \leq 1 \\ 0 & \text{elsewhere,} \end{cases}$$

KTH Matematik

The marginal distribution of x: uniform prior

$$m(x) = \int_0^1 f\left(x \mid p\right) \cdot \pi\left(p\right) dp$$

$$= \left(\begin{array}{c} n \\ x \end{array} \right) \int_0^1 p^x \cdot (1-p)^{n-x} \, dp.$$

We use the Beta integral:

The Beta Integral

$$\int_0^1 p^{\alpha-1}(1-p)^{\beta-1}dp = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Recall also that $\Gamma(x+1) = x!$, if x is a positive integer. $\alpha = \beta = 1$ gives the distribution $R(0,1)$. We set

$$B(\alpha,\beta) := \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

KTH VETENSKAP OCH KONST
KTH Matematik

The Beta Density

$$\pi(p) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} & 0 < p < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

is a probability density $\mathcal{B}\text{eta}(\alpha, \beta)$. $\alpha > 0$ and $\beta > 0$ are hyperparameters.

The marginal distribution of x: uniform prior

$$m(x) = \int_0^1 f\left(x \mid p\right) \cdot \pi\left(p\right) dp$$

$$= \left(\begin{array}{c} n \\ x \end{array} \right) \frac{x!(n-x)!}{(n+1)!}$$

KTH Matematik

The marginal distribution of x, $p \in U(0,1)$

$$m(x) = \int_0^1 f(x \mid p) \cdot dp = \left(\begin{array}{c} n \\ x \end{array} \right) \frac{x!(n-x)!}{(n+1)!}$$

$$= \frac{n!}{x!(n-x)!} \frac{x!(n-x)!}{(n+1)!} = \frac{1}{(n+1)}$$

There is an interpretation of Bayes' work claiming that the problem really attacked and solved by Bayes was: What should $\pi(p)$ be so that

$$\int_0^1 f(x \mid p) \cdot \pi(p) dp = \frac{1}{(n+1)}$$

holds for the Billiard Balls.

KTH Matematik

The Posterior Density for n rolls of Bayes' Orange Ball

$$\pi(p \mid x) = \frac{\binom{n}{x} p^x \cdot (1-p)^{n-x}}{m(x)}$$

$$= \begin{cases} \frac{(n+1)!}{x!(n-x)!} \cdot p^k (1-p)^{n-k} & 0 \leq p \leq 1 \\ 0 & \text{elsewhere.} \end{cases}$$

This is again a Beta density, i.e., we have used a conjugate family of priors.

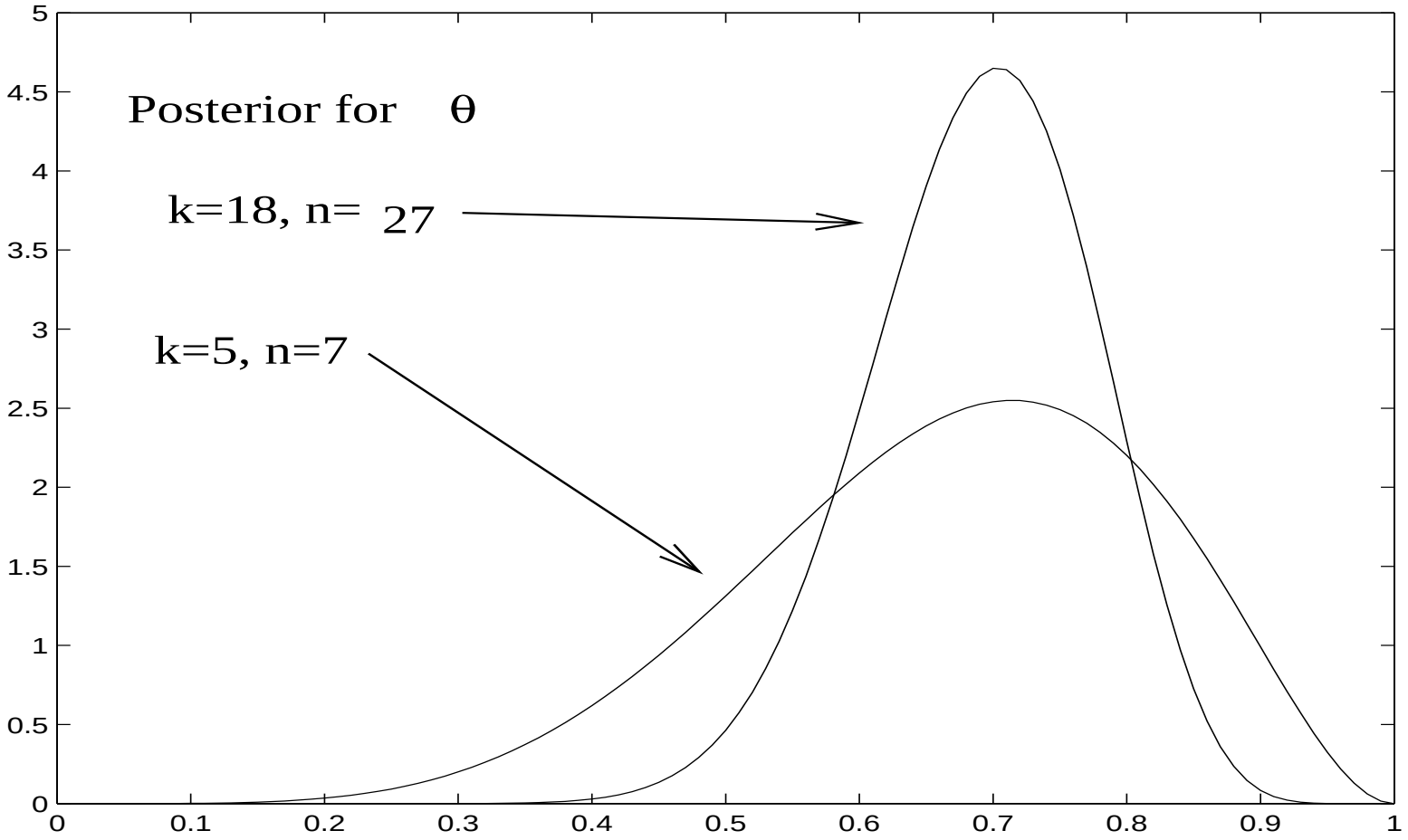
KTH Matematik

The Posterior Density for n rolls of Bayes' Orange Ball

$$\frac{(n+1)!}{x!(n-x)!} = \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} = \frac{1}{B(x+1, n-x+1)}.$$

KTH Matematik

The Posterior Density for n rolls of Bayes' Orange Ball



The densities in the picture do not have the proper constants.

KTH Matematik

The Posterior Density with n rolls of Bayes Ball, $p \in \mathcal{B}\text{eta}(\alpha, \beta)$

$$\pi(p \mid x) = \begin{cases} \frac{1}{B(x+\alpha, n-x+\beta)} \cdot p^{x+\alpha-1}(1-p)^{\beta+n-x-1} & 0 \leq p \leq 1 \\ 0 & \text{elsewhere.} \end{cases}$$

This is the Beta density $\mathcal{B}\text{eta}(\alpha + x, \beta + n - x)$.

End of Example

The example of Bayes' Billiard Balls W and O is discontinued for the moment, but will be reconsidered later.

Joint probability

The probability distribution $\phi(x, y)$ is a joint distribution of (x, y).
We assume a parametric model $(x, y, \theta) \in \phi(x, y, \theta)$ so that:

$$\phi(x, y) = \int \phi(x, y, \theta)\, d\theta.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Predictive Distribution (1) (General)

The predictive distribution $g(y|x)$ of y conditional on x, denoted by $g\left(y|x\right)$ is

$$g\left(y|x\right) = \frac{\phi(x,y)}{m(x)} = \frac{\int \phi\left(x,y,\theta\right) d\theta}{m(x)}$$

$$\frac{\phi(x,y)}{m(x)} = \frac{\int \phi\left(x,y,\theta\right) d\theta}{m(x)} = \frac{\int g\left(y|x,\theta\right) \phi\left(x,\theta\right) d\theta}{m(x)}$$

$$= \frac{\int g\left(y|x,\theta\right) f\left(x|\theta\right) \pi\left(\theta\right) d\theta}{m(x)}$$

Predictive Distribution (2)

i.e.,

$$g(y|x) = \frac{\int g(y|x,\theta) f(x|\theta) \pi(\theta) d\theta}{m(x)}$$

$$= \frac{\int g(y|x,\theta) f(x|\theta) \pi(\theta) d\theta}{\int f(x \mid \theta) \cdot \pi(\theta) d\theta}$$

$$= \int g(y|x,\theta) \frac{f(x|\theta) \pi(\theta)}{\int f(x \mid \theta) \cdot \pi(\theta) d\theta} d\theta$$

KTH Matematik

Predictive Distribution (3)

The predictive distribution $g(y|x)$ is thus

$$g(y|x) = \int g(y|x,\theta)\,\pi(\theta \mid x)\,d\theta.$$

KTH Matematik

Predictive Distribution (4)

If y and x are conditionally independent given θ, then

$$g\left(y|x,\theta\right) = g\left(y|\theta\right)$$

and

$$g\left(y|x\right) = \int g\left(y|\theta\right)\pi\left(\theta \mid x\right)d\theta.$$

KTH Matematik

Predictive Distribution (5)

Introduce order (e.g., in time). $x = x^{(n)} = (x_1, \ldots, x_n)$ $y = x_{n+1}$.

$$g\left(x_{n+1}|x^{(n)}\right) = \int g\left(x_{n+1}|x^{(n)}, \theta\right) \pi\left(\theta \mid x^{(n)}\right) d\theta.$$

We assume conditional independence

$$\phi\left(x, y|\theta\right) = \phi\left(x^{(n)}, x_{n+1}|\theta\right) = \left(f\left(x_1|\theta\right) \cdots f\left(x_n|\theta\right)\right) \cdot f\left(x_{n+1}|\theta\right)$$

$$= \phi\left(x^{(n)}|\theta\right) \cdot f\left(x_{n+1}|\theta\right)$$

KTH Matematik

Predictive Distribution (6)

$$\phi\left(x^{(n)}, x_{n+1} \mid \theta\right) = \phi\left(x^{(n)}|\theta\right) \cdot f\left(x_{n+1}|\theta\right)$$

Then

$$g\left(x_{n+1}|x^{(n)}, \theta\right) = \frac{\phi\left(x^{(n)}, x_{n+1} \mid \theta\right)}{\phi\left(x^{(n)}|\theta\right)} = f\left(x_{n+1}|\theta\right)$$

and

$$g\left(x_{n+1}|x^{(n)}\right) = \int f\left(x_{n+1}|\theta\right) \pi\left(\theta \mid x^{(n)}\right) d\theta$$

KTH VETENSKAP OCH KONST
KTH Matematik

Predictive Distribution (7)

$$g\left(x_{n+1}|x^{(n)}\right) = \int f\left(x_{n+1}|\theta\right)\pi\left(\theta \mid x^{(n)}\right)d\theta$$

We often need to update $g\left(x_{n+1}|x^{(n)}\right)$ for new observations. We need clearly only to update $\pi\left(\theta \mid x^{(n)}\right)$.

KTH Matematik

Up-date of Posterior Distribution (1)

$$\pi\left(\theta \mid x^{(n+1)}\right) = \frac{f\left(x^{(n+1)}|\theta\right)\pi\left(\theta\right)}{\int f\left(x^{(n+1)} \mid \theta\right)\cdot\pi\left(\theta\right)d\theta}$$

$$= \frac{f\left(x_{n+1}|\theta\right)f\left(x_1|\theta\right)\cdots f\left(x_n|\theta\right)\pi\left(\theta\right)}{\int f\left(x_{n+1}|\theta\right)f\left(x_1|\theta\right)\cdots f\left(x_n|\theta\right)\cdot\pi\left(\theta\right)d\theta}$$

$$= \frac{f\left(x_{n+1}|\theta\right)\frac{f(x_1|\theta)\cdots f(x_n|\theta)\pi(\theta)}{m\left(x^{(n)}\right)}}{\int f\left(x_{n+1}|\theta\right)\frac{f(x_1|\theta)\cdots f(x_n|\theta)\pi(\theta)}{m\left(x^{(n)}\right)}d\theta}$$

$$= \frac{f\left(x_{n+1}|\theta\right)\pi\left(\theta|x^{(n)}\right)}{\int f\left(x_{n+1}|\theta\right)\pi\left(\theta|x^{(n)}\right)d\theta}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Up-date of Posterior Distribution (2)

$$\pi\left(\theta \mid x^{(n+1)}\right) = \frac{f\left(x_{n+1}|\theta\right)\pi\left(\theta|x^{(n)}\right)}{\int f\left(x_{n+1}|\theta\right)\pi\left(\theta|x^{(n)}\right)d\theta}$$

Hence, under the assumptions made, we can update posterior distribution in a sequential manner. Or, we can use the posterior of $\pi\left(\theta|x^{(n)}\right)$ as a new prior for computing $\pi\left(\theta \mid x^{(n+1)}\right)$.

KTH Matematik

Fundamental Task of Statistical Inference

The fundamental task of statistical inference is to pass from one set of observations x to express an opinion about another, as yet unobserved set y.
We have above accomplished this in terms of the predictive distribution

$$g(y|x) = \int f(y|\theta)\,\pi(\theta \mid x)\,d\theta,$$

which shows the role of learning about parameters in accomplishing this task

KTH Matematik

An Example $f(x \mid \theta) \leftrightarrow \mathrm{N}\left(\mu, \sigma^2\right)$

$f(x \mid \theta) \leftrightarrow \mathrm{N}\left(\mu, \sigma^2\right)(x)$, the mean μ and variance σ^2 are unknown, i.e., $\theta = (\mu, \sigma^2)$. The (improper) prior density is taken as

$$\pi(\theta) \propto d\mu \frac{1}{\sigma} d\sigma.$$

Let $x = \underline{x}_n = \left(x^{(1)}, \ldots, x^{(n)}\right)$, $x^{(i)} \in \mathrm{N}\left(\mu, \sigma^2\right)$. We have the estimates $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$, and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x^{(i)} - \widehat{\mu}\right)^2$.

KTH VETENSKAP OCH KONST
KTH Matematik

An Example of Bayesian prediction $f(x \mid \theta) \leftrightarrow \mathrm{N}\left(\mu, \sigma^2\right)$ $\pi(\theta) \propto d\mu \frac{1}{\sigma} d\sigma$

The predictive density is

$$g\left(x^{(n+1)} | \underline{x}_n\right) =$$

$$\sqrt{\frac{n}{(n^2-1)\pi}} \cdot \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{1}{2}(n-1)s\right)} \cdot \left(1 + \frac{n\left(x^{(n+1)} - \widehat{\mu}\right)^2}{(n^2-1)s^2}\right)^{-n/2}$$

KTH Matematik

a t-like distribution

The predictive density

$$\sqrt{\frac{n}{(n^2-1)\pi}} \cdot \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{1}{2}(n-1)s\right)} \cdot \left(1 + \frac{n\left(x^{(n+1)} - \widehat{\mu}\right)^2}{(n^2-1)s^2}\right)^{-n/2}$$

is known in Bayesian statistics as 't-like distribution'[1].

[1] J.M. Dickey (1968): Three Multidimensional-Integral Identities with Bayesian Applications. *The Annals of Mathematical Statistics*, 39, pp. 1615–1628.

KTH VETENSKAP OCH KONST
KTH Matematik

Predictive Distribution for the next m rolls of Bayes' Orange Ball

The predictive distribution of y positions of O left of W in m additional rolls is

$$g(y|x) = \int_0^1 f(y|p)\,\pi(p \mid x)\,dp$$

$$= \binom{m}{y} \int_0^1 p^y \cdot (1-p)^{m-y}\,\pi(p \mid x)\,dp$$

$y = 0, 1, \ldots, m$

KTH VETENSKAP OCH KONST
KTH Matematik

Predictive Distribution for the next m rolls of Bayes' Ball

$$\int_0^1 p^y \cdot (1-p)^{m-y} \, \pi(p \mid x) \, dp =$$

$$\int_0^1 p^y \cdot (1-p)^{m-y} \, \frac{1}{B(x+1, n-x+1)} \cdot p^x (1-p)^{n-x} \, dp =$$

$$= \frac{1}{B(x+1, n-x+1)} \int_0^1 p^{y+x} \cdot (1-p)^{m+n-y-x} \, dp =$$

$$= \frac{B(y+x+1, m+n-x-y+1)}{B(x+1, n-x+1)}.$$

by the Beta integral.

KTH Matematik

Predictive Distribution for the next m rolls of Bayes' Ball

$$g(y|x;m) = \binom{m}{y} \int_0^1 p^y \cdot (1-p)^{m-y} \, \pi(p \mid x) \, dp$$

$$= \binom{m}{y} \frac{B(y+x+1, m+n-x-y+1)}{B(x+1, n-x+1)}.$$

$$= \frac{m!}{(m-y)!y!} \frac{\Gamma(n+2)\Gamma(y+x+1)\Gamma(m+n-x-y+1)}{\Gamma(x+1)\Gamma(n-x+1)\Gamma(m+n+2)}.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Predictive Distribution for $y = 1$ in the next $m = 1$ roll of Bayes' Ball

$$g(1|x;1) = \frac{\Gamma(x+2)\Gamma(n-x+1)\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)\Gamma(n+3)}$$

$$= \frac{(x+1)!(n-x)!(n+1)!}{x!(n-x)!(n+2)!} = \frac{x+1}{n+2}$$

A famous predictive probability, known as *Laplace's rule of succession*,

$$\frac{x+1}{n+2} = \frac{x+1}{(n+1)+1}$$

The prior knowledge can be translated into equivalent sample information.

KTH Matematik

The Maximum Likelihood Estimate of p in Bayes' Billiard

$$\widehat{p}_{\mathrm{ML}} = \mathrm{argmin}_{0\leq p\leq 1} ll\left(p \mid x\right)$$

$$= \mathrm{argmin}_{0\leq p\leq 1}\left[-\log\left(\begin{array}{c} n \\ x \end{array}\right) - x\log p - (n-x)\log\left(1-p\right)\right].$$

$$= \mathrm{argmin}_{0\leq p\leq 1}\left(-x\log p - (n-x)\log\left(1-p\right)\right).$$

$$\Rightarrow$$

$$\widehat{p}_{\mathrm{ML}} = \frac{x}{n}$$

If you observed $x = 0$, would you belive in the estimate $\widehat{p} = 0$ for all future purposes ?

KTH VETENSKAP OCH KONST
KTH Matematik

MLE and Predictive Probability in Bayes' Billiard

The predictive probability found above

$$\frac{x+1}{n+2} = \frac{x+1}{(n+1)+1}$$

is a maximum likelihood estimate of p when $n+1$ rolls of the ball O and the first roll of the ball W are included in the data.

KTH Matematik

Q: How do we choose $\pi(\theta)$?

- Assessment (by Questionnaries)
- Conjugate prior
- Non-informative or reference prior
  - Laplace's prior
  - Jeffreys' prior
- Maximum entropy prior

KTH Matematik

Assessment of prior knowledge

(One form of) Bayesian statistics relies upon a **personalistic theory of probability** for quantification of prior knowledge. In such a theory

- probability measures the confidence that a particular individual (assessor) has in the truth of a particular proposition
- no attempt is made to specify which assessments are correct
- personal probabilities should satisfy certain postulates of coherence.

KTH Matematik

Choice of prior distributions: questionnaires

R.L.Winkler (work published in

- Robert L. Winkler: The Assessment of Prior Distributions in Bayesian Analysis
  Journal of the American Statistical Association, Vol. 62, No. 319. (Sep., 1967), pp. 776-800.)

devises questionnaires (or interviews) to elicit information to write down a prior distribution. Students of Univ. of Chigago were asked to, e.g., assess the uncertainty about the probability of a randomly chosen student of Univ. of Chigago being Roman Catholic using a probability distribution. The assessment was done by four different methods, like giving fractiles, making bets, assessing impact of additional data, drawing graphs. One interesting finding is that the assessments by the same person using different methods may be conflicting.

KTH Matematik

Diffuse/Non-diffuse prior distributions by assessment

The priors in Winkler's study are not diffuse: the students of Univ. of Chigago have, since they have been around, an idea about the number of Roman Catholics at the campus of of Univ. of Chigago.

KTH Matematik

Assessing Priors: Conjugate Prior

The interviews by Winkler were mathematically speaking all concerned with assessing the prior of θ in a Bernoulli $\mathrm{Be}(\theta)$ – I.I.D. process. Winkler claims a sensitivity analysis (loc.cit p. 791) showing that the prior distributions assessed by the interviews yielded posterior distributions that were 'only little' different (by a test of goodness-of-fit) from those obtained from Beta densities on θ.

KTH Matematik

Choice of prior distributions by elicitation

- A. O′Hagan: Eliciting Expert Beliefs in Substantial Practical Applications. *The Statistician* , 47, pp. 21−35, 1998.

Not only priors are elicited in

- R.L. Keeney & D. von Winterfeldt: Eliciting Probabilities in Complex Technical Problems. *IEEE Transactions on Engineering Management*, 38, pp.191−201, 1991.

KTH Matematik

Choice of prior distributions by assessment

But this line of study can evolve rapidly to a topic of research in psychology or (economic) behaviour, c.f.,

- C-A. S. Stael von Holstein: *Assessment and Evaluation of Subjective Probability Distributions*. 1970, Stockholm School of Economics.
- A. G. Wilson: *Cognitive factors affecting subjective probability assessments*.
  ISDS Discussion Paper # 94-02,
  `http://www.isds.duke.edu/`

so it feels safe to leave the matter at rest here.

KTH Matematik

Parametric Statistical Model, n I.I.D. $|\theta$ rv's

$$x_i|\theta \in f\left(x|\theta\right), \text{ I.I.D. },$$

or **independent, identically, distributed conditional on** θ

$$x^{(n)} = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$$

$f\left(x|\theta\right)$ is a probability density on R^p. $f\left(x|\theta\right)$ is a known function of x and θ. θ is an unknown parameter $\in \Theta =$ a vector space of finite dimension.

KTH VETENSKAP OCH KONST
KTH Matematik

Asymptotic Shape of the Posterior

Let us assume that $f(x|\theta)$ is a density with a scalar parameter (for simplicity of notation) , and that $f(x|\theta)$ is some $k \geq 2$ times differentiable in θ. We let $\widehat{\theta}_{ML}$ be the maximum likelihood estimate of θ. We expand the log likelihood function around $\widehat{\theta}_{ML}$

$$\log f\left(x^{(n)}|\theta\right) =$$

$$\log f\left(x^{(n)}|\widehat{\theta}_{ML}\right) + \left(\theta - \widehat{\theta}_{ML}\right) \frac{d}{d\theta} \log f\left(x^{(n)}|\widehat{\theta}_{ML}\right)$$

$$+\frac{1}{2}\left(\theta - \widehat{\theta}_{ML}\right)^2 \frac{d^2}{d\theta^2} \log f\left(x^{(n)}|\widehat{\theta}_{ML}\right) + R_n(\theta)$$

KTH Matematik

Asymptotic Shape of the Posterior

But here $\widehat{\theta}_{ML}$ is a solution of the equation

$$\frac{d}{d\theta} \log f\left(x^{(n)}|\widehat{\theta}_{ML}\right) = 0$$

Hence

$$\log f\left(x^{(n)}|\theta\right) =$$

$$\log f\left(x^{(n)}|\widehat{\theta}_{ML}\right) + \frac{1}{2}\left(\theta - \widehat{\theta}_{ML}\right)^2 \frac{d^2}{d\theta^2} \log f\left(x^{(n)}|\widehat{\theta}_{ML}\right) + R_n\left(\theta\right)$$

KTH Matematik

Asymptotic Shape of the Posterior: Law of Large Numbers

We have by assumption of I.I.D. data

$$\frac{d^2}{d\theta^2}\log f\left(x^{(n)}|\theta\right) = \sum_{l=1}^{n}\frac{d^2}{d\theta^2}\log f\left(x_l|\theta\right)$$

We set $Y_l = \frac{d^2}{d\theta^2}\log f\left(x_l|\theta\right)$. Then the Law of Large Numbers says that

$$\frac{1}{n}\sum_{l=1}^{n} Y_l \to E\left[Y\right], n \to \infty$$

where

$$E\left[Y\right] = \int_{\mathcal{X}} \frac{d^2}{d\theta^2}\log f\left(x|\theta\right) f\left(x \mid \theta\right) dx$$

KTH Matematik

Asymptotic Shape of the Posterior: Fisher Information

The integral

$$I(\theta) = -\int_{\mathcal{X}} \frac{d^2}{d\theta^2} \log f(x|\theta)\, f(x \mid \theta)\, dx$$

is called *Fisher information*.

KTH Matematik

Asymptotic Shape of the Posterior: Fisher Information

Then we may feel inclined to believe that

$$\frac{d^2}{d\theta^2} \log f\left(x^{(n)}|\widehat{\theta}_{ML}\right) = \sum_{l=1}^{n} \frac{d^2}{d\theta^2} \log f\left(x_l|\widehat{\theta}_{ML}\right) \approx -n \cdot I\left(\widehat{\theta}_{ML}\right)$$

Note that even $\widehat{\theta}_{ML}$ depends on n.

KTH Matematik

Asymptotic Shape of the Posterior

This gives

$$\log f\left(x^{(n)}|\theta\right) \approx$$

$$\log f\left(x^{(n)}|\widehat{\theta}_{ML}\right) - \frac{1}{2}\left(\theta - \widehat{\theta}_{ML}\right)^2 n \cdot I\left(\widehat{\theta}_{ML}\right)$$

The first term does not involve θ.

KTH VETENSKAP OCH KONST
KTH Matematik

Asymptotic Shape of the Posterior

Then

$$f\left(x^{(n)}|\theta\right) \approx e^{-\frac{n}{2}\left(\theta-\widehat{\theta}_{ML}\right)^2 \cdot I\left(\widehat{\theta}_{ML}\right)}$$

The interpretation of the relation is that the likelihood function can be for large n be approximated by a normal density for which the mean is $\widehat{\theta}_{ML}$ and the variance is $\frac{1}{nI\left(\widehat{\theta}_{ML}\right)}$.

KTH VETENSKAP OCH KONST
KTH Matematik

Finally: Jeffreys' prior

Let $I(\theta)$ be the Fisher information of a parametric model. Take the prior density as

$$\pi(\theta) \stackrel{def}{=} \frac{\sqrt{I(\theta)}}{\int \sqrt{I(\theta)} d\theta},$$

assuming the integral exists. This choice of prior is known as Jeffreys' prior. This prior is invariant to monotonous transformations of θ.

KTH Matematik

Finally: Jeffreys' prior

It turns out that Jeffreys' prior for a binomial likelihood is obtained by $\alpha = 1/2$ and $\beta = 1/2$ in

$$\pi(p) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} & 0 < p < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

and with $\Gamma(1/2+1/2) = 1$, $\Gamma(1/2) = \sqrt{\pi}$,

$$\pi_{Jeffreys}(p) = \begin{cases} \frac{1}{\pi} p^{-1/2}(1-p)^{-1/2} & 0 < p < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

KTH Matematik