STK4021 Problem set 1

Erik Alexander Sandvik

Nils collection 1. Prior to posterior updating with Poisson data

We say that $Z \sim \text{Gamma}(a, b)$ if its density is

$$g(z) = \frac{b^a}{\Gamma(a)} z^{a-1} e^{-bz}, \;\; z \in \langle 0, \infty \rangle, \;\; a, b > 0.$$

a)

The expectation value of z is

$$E(z) = \int_0^\infty z g(z) dz = \frac{b^a}{\Gamma(a)} \int_0^\infty z^a e^{-bz} dz.$$

Making a change of variable $u = bz \rightarrow z = u \,/\, b \rightarrow dz = du \,/\, b$ we have

$$E(z) = \frac{b^a}{\Gamma(a)} \int_0^\infty \left(\frac{u}{b}\right)^a e^{-u} \frac{du}{b} = \frac{1}{b} \frac{1}{\Gamma(a)} \int_0^\infty u^a e^{-u} du.$$

The Gamma function is defined as

$$\Gamma(a) = \int_0^\infty u^{a-1} e^{-u} du,$$

so we further have

$$E(z) = \frac{1}{b} \frac{\Gamma(a+1)}{\Gamma(a)} = \frac{1}{b} \frac{a\Gamma(a)}{\Gamma(a)} = \frac{a}{b}.$$

To find the variance of z we can first find the second moment:

$$E\big(z^2\big) = \int_0^\infty z^2 g(z) dz = \frac{b^a}{\Gamma(a)} \int_0^\infty z^{a+1} e^{-bz} dz$$

$$= \frac{b^a}{\Gamma(a)} \int_0^\infty \left(\frac{u}{b}\right)^{a+1} e^{-u} \frac{du}{b} = \frac{1}{b^2 \Gamma(a)} \int_0^\infty u^{a+1} e^{-u} du$$

$$= \frac{\Gamma(a+2)}{b^2 \Gamma(a)} = \frac{\Gamma(a+1+1)}{b^2 \Gamma(a)} = \frac{(a+1)\Gamma(a+1)}{b^2 \Gamma(a)}$$

$$= \frac{(a+1)a\Gamma(a)}{b^2 \Gamma(a)} = \frac{(a+1)a}{b^2}.$$

Then the variance is

$$\mathrm{var}(u) = E\big(z^2\big) - E(z)^2$$

$$= \frac{(a+1)a}{b^2} - \frac{a^2}{b^2} = \frac{a^2 + a - a^2}{b^2}$$

$$= \frac{a}{b^2} = \frac{E(z)}{b}.$$

b)

The Poisson distribution is given by

$$p(k|\lambda) = \frac{\lambda^k}{k!} e^{-\lambda}, \; k = 0, 1, 2, \ldots.$$

In this problem we assume that $y|\theta$ is Poisson distributed and that θ has prior distribution Gamma(a, b). Then the posterior distribution is given by

$$p(\theta|y) \propto p(y|\theta) p(\theta)$$

$$= \frac{\theta^y}{y!} e^{-\theta} \cdot \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta}.$$

Dropping all factors independent of θ:

$$p(\theta|y) \propto \theta^y e^{-\theta} \cdot \theta^{a-1} e^{-b\theta}$$

$$= \theta^{y+a-1} e^{-(b+1)\theta}$$

We recognize this as the unnormalized Gamma distribution $\theta \sim$ Gamma$(y+a,\ b+1)$.

c)

Interpreting $p(\theta|y_1)$ as the new prior and $p(y_2|\theta)$ as the likelihood we have for the posterior

$$p(\theta|y_1, y_2) \propto p(y_2|\theta) p(\theta|y_1)$$

$$= \frac{\theta^{y_2}}{y_2!} e^{-\theta} \cdot \theta^{y_1+a-1} e^{-(b+1)\theta} \propto \theta^{y_1+y_2+a-1} e^{-(b+2)\theta}$$

$$\sim \text{Gamma}(a + y_1 + y_2, b + 2).$$

Thus in general, if we have a data set $\{y_i\}_{i=1}^n$ the posterior distribution becomes

$$\theta|\{y_i\}_{i=1}^n \sim \text{Gamma}(a + y_1 + ... + y_n, b + n).$$

Since the data set consists of i.i.d. elements, we have that

$$p\Big(\theta|\{y_i\}_{i=1}^n\Big) \propto \prod_{i=1}^n p(y_i|\theta) p(\theta)$$

where the order of the factors doesn't matter and any factor may be absorbed into the prior. The conclusion is that the order of observations doesn't matter. Observing y_i before y_j and vice versa does not affect the final posterior distribution.

Nils collection 2. The Master Recipe for finding the Bayes solution

Consider the general framework:

We have some data $y \in \mathcal{Y}$ where $\mathcal{Y}$ is the space of all possible data, with distribution $p(y|\theta)$. $\theta \in \Omega$ is an unknown parameter value which belongs to the space Ω of all possible parameters values. θ is assumed to have the prior distribution $p(\theta)$.

We have a statistical decision function $\widehat{a} : \mathcal{Y} \to \mathcal{A}$, which from data y yields the action or decision $a = \widehat{a}(y)$. Note that a is a decision, while $\widehat{a}$ is a decision function.

The loss function $L(\theta, a)$ is a measure of how much you messed up if you took action a while the real parameter value happened to be θ.

The risk function $R(\theta, \widehat{a})$ is the expectation value of the loss function over all possible data, given the parameter value θ. That is,

$$R(\theta, \widehat{a}) = \mathrm{E}_{y|\theta}[L(\theta, \widehat{a})] = \int_{\mathcal{Y}} dy L(\theta, \widehat{a}) p(y|\theta).$$

Note that $R(\theta, \widehat{a})$ is a function of θ, and a functional of $\widehat{a}$.

The Bayes risk $\mathrm{BR}(p, \widehat{a})$, which is a functional of the prior $p(\theta)$ and the decision function $\widehat{a}$, is the expectation value of the risk function over all possible parameter values,

$$\mathrm{BR}(p, \widehat{a}) = \mathrm{E}_{\theta}[R(\theta, \widehat{a})] = \int_{\Omega} d\theta R(\theta, \widehat{a}) p(\theta).$$

The minimum Bayes risk is the smallest possible bayes risk over all action functions $\widehat{a}$,

$$\mathrm{MBR}(p) = \min_{\widehat{a}} \mathrm{BR}(p, \widehat{a}),$$

which is a functional of the prior $p(\theta)$.

The Bayes solution of the problem is the decision function $\widehat{a}_B$ which succeeds in minimizing the Bayes risk,

$$\widehat{a}_B = \underset{\widehat{a}}{\text{argmin}}\ \text{BR}(p, \widehat{a}),$$

and of course we may express the minimum Bayes risk as

$$\text{MBR}(p) = \text{BR}(p, \widehat{a}_B).$$

The master theorem about the Bayes procedure is that there is a recipé for finding the optimal Bayes solution $\widehat{a}_B$, given the (limited) data y.

a) & b)

Suppose we have two continuous random variables a and b, with joint probability density $p(a, b)$. We define the marginal probability density for a as

$$p(a) = \int p(a, b) db.$$

The conditional probability density for a is defined as

$$p(a|b) = \frac{p(a, b)}{p(b)},$$

where $p(b)$ is the marginal probability density for b. The conditional probability density for b is of course

$$p(b|a) = \frac{p(a, b)}{p(a)}.$$

The latter two equations give us two ways of expressing $p(a, b)$ in terms of conditional and marginal distributions, which we can equate to get Bayes theorem,

$$p(a|b)p(b) = p(b|a)p(a).$$

The posterior distribution for θ given the data y is thus

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}.$$

We require the posterior distribution to be normalized to one,

$$1 = \int d\theta \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{1}{p(y)} \int d\theta p(y|\theta)p(\theta).$$

The marginal distribution of y thus playes the role of a normalization constant,

$$p(y) = \int d\theta p(y|\theta)p(\theta).$$

c)

We can of course write the Bayes risk in terms of expectation values of the loss function with respect to both θ and y:

$$\mathrm{BR}(p, \widehat{a}) = \int_\Omega d\theta R(\theta, \widehat{a})p(\theta)$$

$$= \int_\Omega d\theta \left\{ \int_y dy L(\theta, \widehat{a})p(y|\theta) \right\} p(\theta) = \mathrm{E}_\theta[\mathrm{E}_{y|\theta}[L(\theta, \widehat{a})]].$$

d)

We can bring $p(\theta)$ inside the y-integral since it is just a constant,

$$\mathrm{BR}(p, \widehat{a}) = \int_\Omega d\theta \int_y dy L(\theta, \widehat{a})p(y|\theta)p(\theta).$$

Now using Bayes theorem,

$$\mathrm{BR}(p,\widehat{a}) = \int_{\Omega} d\theta \int_{y} dy L(\theta,\widehat{a}) p(\theta|y) p(y),$$

and rearranging the integrals,

$$\mathrm{BR}(p,\widehat{a}) = \int_{y} dy \int_{\Omega} d\theta L(\theta,\widehat{a}) p(\theta|y) p(y)$$

$$= \int_{y} dy \left\{ \int_{\Omega} d\theta L(\theta,\widehat{a}) p(\theta|y) \right\} p(y) = \mathrm{E}_{y}[\mathrm{E}_{\theta|y}[L(\theta,\widehat{a})]].$$

In order to minimize the Bayes risk and get the optimal Bayes solution $\widehat{a}_B$ it is sufficient to minimize the inner integral,

$$\widehat{a}_B = \underset{\widehat{a}}{\operatorname{argmin}} \int_{\Omega} d\theta L(\theta,\widehat{a}) p(\theta|y)$$

$$= \underset{\widehat{a}}{\operatorname{argmin}}\ \mathrm{E}_{\theta|y}[L(\theta,\widehat{a})],$$

i.e the optimal Bayes solution $\widehat{a}_B$ is the minimum of the posterior expectation value of the loss function.

Nils collection 12. Alarm or not?

We assume that $y|\theta, n$ is binomially distributed,

$$p(y|\theta) = \binom{n}{y} \theta^{y} (1-\theta)^{n-y},$$

and that the action space is $\mathcal{A} = \Big\{ \text{alarm},\ \text{no alarm} \Big\}$ with loss function

$$L\Big(\theta, \text{no alarm}\Big) = \begin{cases} 5000 & \text{if } \theta > 0.15 \\ 0 & \text{if } \theta < 0.15 \end{cases},$$

$$L\big(\theta, \text{alarm}\big) = \begin{cases} 0 & \text{if } \theta > 0.15 \\ 1000 & \text{if } \theta < 0.15 \end{cases}.$$

We want to find out for which values of y that the correct decision is 'alarm' for $n = 50$ and for some prior distribution $p(\theta)$.

a) $\theta \sim \text{Uni}(0, 1)$

When θ is uniformly distributed, the posterior distribution is simply

$$p(\theta|y) \propto p(y|\theta)p(\theta))$$

$$\propto \theta^y(1-\theta)^{n-y}.$$

The posterior expectation of the loss function if we don't sound the alarm is

$$\text{E}_{\theta|y}\Big[L\big(\theta, \text{no alarm}\big)\Big] = \int_0^1 L\big(\theta, \text{no alarm}\big)p(\theta|y)d\theta$$

$$\propto \int_{0.15}^1 5000\theta^y(1-\theta)^{n-y}d\theta.$$

If we do sound the alarm, we have

$$\text{E}_{\theta|y}\Big[L\big(\theta, \text{alarm}\big)\Big] = \int_0^1 L\big(\theta, \text{alarm}\big)p(\theta|y)d\theta$$

$$\propto \int_0^{0.15} 1000\theta^y(1-\theta)^{n-y}d\theta.$$

We thus sound the alarm when the quotient

$$Q(y) \equiv \frac{\text{E}_{\theta|y}\Big[L\big(\theta, \text{alarm}\big)\Big]}{\text{E}_{\theta|y}\Big[L\big(\theta, \text{no alarm}\big)\Big]}$$

$$= \frac{\int_0^{0.15} 1000\theta^y(1-\theta)^{n-y}d\theta}{\int_{0.15}^1 5000\theta^y(1-\theta)^{n-y}d\theta} = \frac{1}{5}\frac{\int_0^{0.15} \theta^y(1-\theta)^{n-y}d\theta}{\int_{0.15}^1 \theta^y(1-\theta)^{n-y}d\theta}$$

is less than one.

b) & c)

The same procedure is repeated for the prior being a beta distribution

$$\theta \sim \text{Beta}(\alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1},$$

for $\theta \sim \text{Beta}(2, 8)$ and $\theta \sim \text{Beta}(2, 8) + \text{Beta}(8, 2)$. This is automated by alarm_no_alarm.py and the plot of the quotient for the respective priors is found in quotient.pdf.

Problem 2.10: A cable car in San Francisco

Suppose there are N cable cars in San Francisco numbered sequentially from 1 to N. We happen to see cable car $\#203$, and we want to estimate the number N of cable cars. The prior distribution is taken out of a hat:

$$P(N) = \left(\frac{1}{100}\right)\left(\frac{99}{100}\right)^{N-1},\ N = 1, 2,\ ...$$

The likelihood is assumed to be uniform:

$$P(y|N) = \frac{1}{N} \cdot I(1 \leq y \leq N)$$

where $I(\,\cdot\,)$ is the indicator function; 1 if $\,\cdot\,$ is true and 0 if $\,\cdot\,$ is false.

a)

The posterior distribution is found from Bayes theorem,

$$P(N|y) = \frac{P(y|N)P(N)}{P(y)},$$

where

$$P(y) = \sum_{N=1}^{\infty} P(y|N)P(N)$$

$$= \sum_{N=1}^{\infty} \frac{1}{N} I(1 \leq y \leq N) \left(\frac{1}{100}\right) \left(\frac{99}{100}\right)^{N-1}.$$

The full normalized distribution is

$$P(N|y) = \frac{\frac{1}{N} I(1 \leq y \leq N) \left(\frac{1}{100}\right) \left(\frac{99}{100}\right)^{N-1}}{\sum_{N'=1}^{\infty} \frac{1}{N'} I(1 \leq y \leq N') \left(\frac{1}{100}\right) \left(\frac{99}{100}\right)^{N'-1}}.$$

We happened to see the cable car numbered $y = 203$, so we have

$$P(N|203) = \frac{\frac{1}{N} \left(\frac{1}{100}\right) \left(\frac{99}{100}\right)^{N-1}}{\sum_{N'=203}^{\infty} \frac{1}{N'} \left(\frac{1}{100}\right) \left(\frac{99}{100}\right)^{N'-1}} I(N \geq 203)$$

$$= \frac{\frac{1}{N} \left(\frac{99}{100}\right)^{N-1}}{\sum_{N'=203}^{\infty} \frac{1}{N'} \left(\frac{99}{100}\right)^{N'-1}} I(N \geq 203).$$

b)

The first two moments are, of course

$$\mathrm{E}[N] = \sum_{N=203}^{\infty} N P(N|203),$$

$$\mathrm{E}\left[N^2\right] = \sum_{N=203}^{\infty} N^2 P(N|203),$$

and the variance is, as always

$$\mathrm{var}(N) = \mathrm{E}\left[N^2\right] - \mathrm{E}[N]^2.$$

The mean and standard deviation are 280 and 80 respectively (solved numerically).