Assignment 1

Thijs Quast (thiqu264)

10/25/2019

Contents

Contributions

This report was written individually. However, during a meeting I have discussed solutions and collaborated with Julius Kittler, Faton Rekathati, Gustav Lundberg and Alexander Karlsson. During this meeting everyone contributed with ideas, remarks and thoughts on how to solve the assignments.

1

a

If the GP uses a very simplistic way of inference, the doctor would argue the patient has disease A1, as this is the disease with the highest frequency of causing symptoms. This approach is however very simplistic as it does not apply Bayes' Theorem.

b

i

We now have enough information to apply Bayes' Theorem:

$$p(A_i|S) = \frac{p(S|A_i) * p(A_i)}{p(S)}$$

Where:

$$p(A_1) = 0.002$$
$$p(A_2) = 0.003$$
$$p(A_3) = 0.01$$
$$p(A_{no}) = 1 - 0.002 - 0.003 - 0.01 = 0.985$$

Where:

$$p(S) = p(A_1)p(S|A_1) + p(A_2)p(S|A_2) + p(A_3)p(S|A_3) + p(A_{no})p(S|A_{no})$$

$$p(S) = (0.002 * 0.2) + (0.003 * 0.1) + (0.01 * 0.02) + (0.985 * 0.001)$$

$$p(S) = 0.0004 + 0.0003 + 0.0002 + 0.00098 \approx 0.0018$$

Now:

$$p(A_1|S) = \frac{p(S|A_1) * p(A_1)}{p(S)} = \frac{0.002 * 0.2}{0.0018} \approx 0.2122$$
$$p(A_2|S) = \frac{p(S|A_2) * p(A_2)}{p(S)} = \frac{0.003 * 0.1}{0.0018} \approx 0.1591$$
$$p(A_3|S) = \frac{p(S|A_3) * p(A_3)}{p(S)} = \frac{0.01 * 0.002}{0.0018} \approx 0.1061$$

$$p(A_{no}|S) = \frac{p(S|A_{no}) * p(A_{no})}{p(S)} = \frac{0.985 * 0.001}{0.0018} \approx 0.5225$$

```r
# Prior probabilities
p_A1 <- 0.002
p_A2 <- 0.003
p_A3 <- 0.01
p_A_No <- 0.985
priors <- c(p_A1, p_A2, p_A3, p_A_No)


# Conditional probabilities
p_S_A1 <- 0.2
p_S_A2 <- 0.1
p_S_A3 <- 0.02
p_S_ANo <- 0.001
conditionals <- c(p_S_A1, p_S_A2, p_S_A3, p_S_ANo)


# Dataframe
df <- as.data.frame(cbind(priors, conditionals))
rownames(df) <- c("disease A1", "disease A2", "disease A3", "No disease")
```

```r
# Proportional Posteriors
prop_posteriors <- df$priors * df$conditionals
df$prop_posteriors <- prop_posteriors

# Probability of S
prob_S <- sum(df$prop_posteriors)

# Posterior probabilities
df$posteriors <- df$prop_posteriors/prob_S
```

```r
library(knitr)
kable(df)
```

	priors	conditionals	prop_posteriors	posteriors
disease A1	0.002	0.200	0.000400	0.2122016
disease A2	0.003	0.100	0.000300	0.1591512
disease A3	0.010	0.020	0.000200	0.1061008
No disease	0.985	0.001	0.000985	0.5225464

ii

Given that we now have the posterior probabilities, the GP now comes to a different conclusion, namely that the patient most likely has no disease. Therefore in this case I do not think inference to the best explanation is a good way to come to conclusions for a GP and perhaps it's better to apply Bayes' Theorem, conditional on the priors and data that we have.

2

i

Binomial sampling:

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

$$p(r|\theta) \propto \theta^r(1-\theta)^{1-r}$$

Pascal sampling:

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

$$p(n|\theta) \propto \theta^r(1-\theta)^{1-r}$$

As one can see, the proportional likelihoods for both types of sampling are the same. Therefore when applying Bayes' Theorem to arrive at the proportional posterior distribution, under a similar prior, both sampling methods will result in the same proportional posterior distribution.

ii

This sampling until the experimenter is tired is a special way of Pascal sampling, not until a number of successes is achieved but until the experimenter is tired. This will still have the same proportional likelihood, thus the same proportional posterior. Resulting from this, I conclude that the stopping rule is noninformative.

3

Looking at page 80 of the book we can conclude this experiment follows a Poisson distribution:

$$p(\tilde{r} = r|t, \lambda) = \frac{e^{-\lambda t}(\lambda t)^r}{r!}$$

Using the R software we can easily compute these densities.

To arrive at the appropriate values for lambda multiplied by time, we multiply the 5 workers, times their potential speed times 2 hours.

For working speed equal to 30, this will result in the following equation:

$$p(\tilde{r} = 380|t = 2, \lambda = 5 * 30) = \frac{e^{-300}(300)^{380}}{380!}$$

Using R I solve for three possibilities of the working speed.

```r
lambda <- c(30, 40, 50)
time <- c(2,2,2)
lambda_t <- c(300, 400, 500)
priors <- c(0.25, 0.50, 0.25)

df <- as.data.frame(cbind(lambda, time, lambda_t, priors))

ll_30 <- dpois(x = 380, lambda = 300)
```

```
ll_40 <- dpois(x = 380, lambda = 400)
ll_50 <- dpois(x = 380, lambda = 500)
likelihoods <- c(ll_30, ll_40, ll_50)

df$likelihoods <- likelihoods

df$prop_posteriors <- df$priors * df$likelihoods
df$posteriors <- df$prop_posteriors/sum(df$prop_posteriors)

kable(df)
```

lambda	time	lambda_t	priors	likelihoods	prop_posteriors	posteriors
30	2	300	0.25	0.0000011	0.0000003	0.0000448
40	2	400	0.50	0.0123045	0.0061522	0.9999550
50	2	500	0.25	0.0000000	0.0000000	0.0000001

4

$$f(x|\theta) = e^{\sum_{j=1}^{k} A_j(\theta)B_j(x)+C(x)+D(\theta)}$$

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

$$p(x) = B(\alpha,\beta)x^{\alpha-1}(1-x)^{\beta-1}$$

$$e^{ln(p(x))} = e^{lnB(\alpha,\beta)+(\alpha-1)ln(x)+(\beta-1)ln(1-x)+0*f(x)}$$

Where:

$$D(\theta) = lnB(\alpha,\beta)$$
$$A_1(\theta) = (\alpha-1)$$

$$B_1(x) = ln(x)$$
$$A_2(\theta) = (\beta-1)$$
$$B_2(x) = ln(1-x)$$
$$C(x) = 0*f(x)$$

i

Generalization of prior of exponential family:

$$p(\alpha,\beta|\alpha_1,...,\alpha_k,\alpha_{k+1}) = e^{\sum_{j=1}^{k} A_j(\alpha,\beta)\alpha_j+\alpha_{k+1}D(\alpha,\beta)+k(\alpha_1,...,\alpha_k,\alpha_{k+1})}$$

We can drop the function k() under the proportionality constant:

$$p(\alpha,\beta|\alpha_1,...,\alpha_k,\alpha_{k+1}) = e^{\sum_{j=1}^{k} A_j(\alpha,\beta)\alpha_j+\alpha_{k+1}D(\alpha,\beta)}$$

ii

From the slides of meeting 5 we know the following generalization:

PDF of sample point distribution:

$$f(x|\theta) = e^{\sum_{j=1}^{k} A_j(\theta)B_j(x_i)+C(x_i)+D(\theta)}$$

Resulting in the following likelihood:

$$\prod_{i=1}^{n} f(x|\theta) = \prod_{i=1}^{n} e^{\sum_{j=1}^{k} A_j(\theta)B_j(x_i)+C(x_i)+D(\theta)}$$

$$= e^{\sum_{i=1}^{n}(\sum_{j=1}^{k} A_j(\theta)B_j(x_i)+C(x_i)+D(\theta))}$$

$$= e^{\sum_{i=1}^{n}(\sum_{j=1}^{k} A_j(\theta)B_j(x_i)+C(x_i)+D(\theta))}$$

$$= e^{\sum_{j=1}^{k} A_j(\theta)\sum_{i=1}^{n} B_j(x_i)+\sum_{i=1}^{n} C(x_i)+n*D(\theta)}$$

Where

$$\theta = (\alpha, \beta)$$

Now the posterior becomes:

$$p(\alpha, \beta|x) \propto e^{\sum_{j=1}^{k} A_j(\alpha,\beta)\alpha_j+\alpha_{k+1}D(\alpha,\beta)} * e^{\sum_{j=1}^{k} A_j(\alpha,\beta)\sum_{i=1}^{n} B_j(x_i)+\sum_{i=1}^{n} C(x_i)+n*D(\alpha,\beta)}$$

$$p(\alpha, \beta|x) \propto e^{\sum_{j=1}^{k} A(\alpha,\beta)(\sum_{i=1}^{n} B_j(x_i)+\alpha_j)+(n+\alpha_{k+1})*D(\alpha,\beta)}$$

Where:

$$D(\theta) = lnB(\alpha, \beta)$$

$$A_1(\theta) = (\alpha - 1)$$

$$B_1(x) = ln(x)$$

$$A_2(\theta) = (\beta - 1)$$

$$B_2(x) = ln(1 - x)$$

$$C(x) = 0 * f(x)$$

5 From this it follows that:

$$p(\alpha, \beta|x) \propto e^{(\alpha-1)(\sum_{i=1}^{5} ln(x_i)+\alpha_1)+(\beta-1)(\sum_{i=1}^{5} ln(1-x_i)+\alpha_2)+(5+\alpha_3)ln(B(\alpha,\beta))}$$

When looking at the difference between the beta prior and posterior, we see the difference in the exponential, this difference is equivalent to the effect of the data on the posterior distribution.