UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in: STK4021 — Applied Bayesian statistics - Solutions

Day of examination: November 30 - 2022

Examination hours: 15.00 – 19.00.

This problem set consists of 4 pages.

Appendices: None

Permitted aids: Approved calculator

Please make sure that your copy of the problem set is
complete before you attempt to answer anything.

Problem 1

(a) We have that

$$\Pr(C = c|y) \propto b_c^a e^{-b_c y}$$

and using that the probabilities needs to sum to one, we get

$$\Pr(C = c|y) = \frac{b_c^a e^{-b_c y}}{\sum_{c'=1}^3 b_{c'}^a e^{-b_{c'} y}}$$

(b) We have that

$$E[L(c, \hat{c})|y] = \Pr(c \neq \hat{c}|y) = 1 - \Pr(c = \hat{c}|y)$$

giving that we minimize loss by putting $\hat{c}$ to the class with highest probability.

We get

$$\begin{aligned} b_2^a e^{-b_2 y} >& b_1^a e^{-b_1 y} \\ &\Updownarrow \\ e^{(b_1 - b_2)y} >& (b_1/b_2)^a \\ &\Updownarrow \\ (b_1 - b_2)y >& a\log(b_1/b_2) \\ &\Updownarrow \\ y >& (a\log(b_1/b_2))/(b_1 - b_2) = 6\log(2) = 4.159 \end{aligned}$$

(Continued on page 2.)

Similarly, we get

$$\begin{aligned} b_3^a e^{-b_3 y} >& b_2^a e^{-b_2 y} \\ y >& (a\log(b_2/b_3))/(b_2-b_3) = 12\log(2) = 8.318 \end{aligned}$$

which gives

$$\hat{c} = \begin{cases} 1 & \text{if } y < 4.159 \\ 2 & \text{if } 4.159 \leq y < 8.318 \\ 3 & \text{if } y \geq 8.318 \end{cases}$$

(c) We now have

$$\begin{aligned} E[L(C,1)|y] =& L(1,1)\Pr(C=1|y) + L(2,1)\Pr(C=2|y) + L(3,1)\Pr(C=3|y) \\ =& 1.5\Pr(C=2|y) + \Pr(C=3|y) \\ E[L(C,2)|y] =& L(1,2)\Pr(C=1|y) + L(2,2)\Pr(C=2|y) + L(3,2)\Pr(C=3|y) \\ =& \Pr(C=1|y) + \Pr(C=3|y) \\ E[L(C,3)|y] =& L(1,3)\Pr(C=1|y) + L(2,3)\Pr(C=2|y) + L(3,3)\Pr(C=3|y) \\ =& \Pr(C=1|y) + 1.5\Pr(C=2|y) \end{aligned}$$

We then have

$$\begin{gathered} E[L(C,1)|y] < E[L(C,2)|y] \\ \Updownarrow \\ 1.5\Pr(C=2|y) + \Pr(C=3|y) < \Pr(C=1|y) + \Pr(C=3|y) \\ \Updownarrow \\ 1.5\Pr(C=2|y) < \Pr(C=1|y) \end{gathered}$$

Similarly

$$\begin{gathered} E[L(C,1)|y] < E[L(C,3)|y] \\ \Updownarrow \\ 1.5\Pr(C=2|y) + \Pr(C=3|y) < \Pr(C=1|y) + 1.5\Pr(C=2|y) \\ \Updownarrow \\ \Pr(C=3|y) < \Pr(C=1|y) \end{gathered}$$

and

$$\begin{gathered} E[L(C,2)|y] < E[L(C,3)|y] \\ \Updownarrow \\ \Pr(C=1|y) + \Pr(C=3|y) < \Pr(C=1|y) + 1.5\Pr(C=2|y) \\ \Updownarrow \\ \Pr(C=3|y) < 1.5\Pr(C=2|y) \end{gathered}$$

which put together gives the decision rule specified.

(d) We end up with a larger region for classification to class 2 which is reasonable since the cost of ”missing” this class is higher.

(Continued on page 3.)

Problem 2

(a) We have (not assumed to be derived)

$$\begin{aligned}L(\boldsymbol{\theta}) =&\prod_{c=1}^{3}\prod_{i=1}^{n_i}\frac{b_c^a}{\Gamma(a)}y_{c,i}^{a-1}e^{-b_cy_{c,i}}\\ =&\frac{\prod_{c=1}^3 b_c^{an_i}}{\Gamma(a)^n}(\prod_{c=1}^{3}\prod_{i=1}^{n_i}y_{c,i})^{a-1}\prod_{c=1}^{3}e^{-b_c\sum_{i=1}^{n_i}y_{c,i}}\end{aligned}$$

The sufficient statistics are $\bar{y}_c=\sum_{i=1}^{n_i}y_{c,i}/n_i$ and $\prod_{c=1}^3\prod_{i=1}^{n_i}y_{c,i}$ (one might also include $n_i, i=1,2,3$). The benefit is that both maximum likelihood estimates and the posterior distribution only depend on the data through the sufficient statistics, so we do not need to derive these.

(b) We have

$$\begin{aligned}p(b_1,b_2,b_3|a,\boldsymbol{y})\propto&\prod_{c=1}^{3}b_c^{\alpha-1}e^{-\beta b_c}b_c^{an_i}e^{-b_cn\bar{y}_c}\\ \propto&\prod_{c=1}^{3}b_c^{\alpha+an_i-1}e^{-b_c[beta+n\bar{y}_c}\\ \propto&\prod_{c=1}^{3}\text{Gamma}(b_c;\alpha+an_i,\beta+n\bar{y}_c)\end{aligned}$$

(c) We have that

$$\begin{aligned}p(b_1,b_2,b_3|a,\boldsymbol{y}) =&\frac{p(a,b_1,b_2,b_3,\boldsymbol{y})}{p(a,\boldsymbol{y})}\\ =&\frac{p(a)p(b_1,b_2,b_3)p(y|a,b_1,b_2,b_3)}{p(a|\boldsymbol{y})p(\boldsymbol{y}))}\end{aligned}$$

which gives

$$p(a|\boldsymbol{y})\propto\frac{p(a)p(b_1,b_2,b_3)p(y|a,b_1,b_2,b_3)}{p(b_1,b_2,b_3|a,\boldsymbol{y})}$$

where each terms involved can be easily computed.

Note that in particular we get

$$\begin{aligned}p(a|\boldsymbol{y})\propto&p(a)\frac{\prod_{c=1}^3\text{Gamma}(b_c,\alpha,\beta)\prod_{c=1}^3\prod_{i=1}^{n_i}\text{Gamma}(y_{c,i};a,b_c)}{\prod_{c=1}^3\text{Gamma}(b_c,\alpha+an_i,\beta+n\bar{y}_c)}\\ =&\end{aligned}$$

One can then either use a MCMC algorithm to sample from $p(a|\boldsymbol{y})$ or discretize $p(a|\boldsymbol{y})$ and then simulate directly from this approximate distribution.

Given a, one can then simulate the b_c's.

(Continued on page 4.)

(d) The effective sample size is based on that since we have correlated data the effective number of samples will be reduced. Here the reduction is modest, giving a reasonable number of samples.

The Rhat parameter is a measure of convergence comparing within chain variation with between chain variance. The value should be close to 1, preferable lower than 1.01 which is satisfied here. Note however that the log-likelhood has a value of 1.01, indicating that we might need some more iterations.

The plot shows that the different chains ae well mixed, but again we might need more iterations.

(e) We can include the unknown classes as part of the simulation procedure, that is use $\Pr(C_i = c|y_i, \boldsymbol{\theta})$ as derived in Problem 1 to simulate these.

Problem 3

(a) Defining the two Gamma distributions by p_1 and p_2, we get

$$\begin{aligned}
p(y_t) =& \int_{\gamma_t} = \int_{\lambda_t} [\pi p_1(y) + (1-\pi)p_2(y)]d\lambda_t\\
=&\pi \int_{\gamma_t} p_1(y)d\lambda_t + (1-\pi)\int_{\gamma_t} p_2(y)d\lambda_t\\
=&\pi\text{Neg-Binom}(y;\alpha_1,\beta_1) + (1-\pi)\text{Neg-Binom}(y;\alpha_2,\beta_2)
\end{aligned}$$

(b) In general LOO-CV is build up by components $f(y_i|\boldsymbol{y}_{-i})$ which in principle can be calculated by fitting n models. However, huge computational benefit can be obtained by only fitting one model based on all data and then use importance sampling to correct for the observation to be delected. Some extra improvements can be obtained through the PSIS approach.

The plot shows that π is almost equal to 1, giving no weight on the second mixture. This is also reflected in the large uncertainty about α_2 and β_2. This is confirmed through the LOO values prefering model 1.

(c) The test statistic measures the correlations between the observations. The model assumes no correlation but the results indicate that there is some correlation in the data, even when corrected for individual λ's.