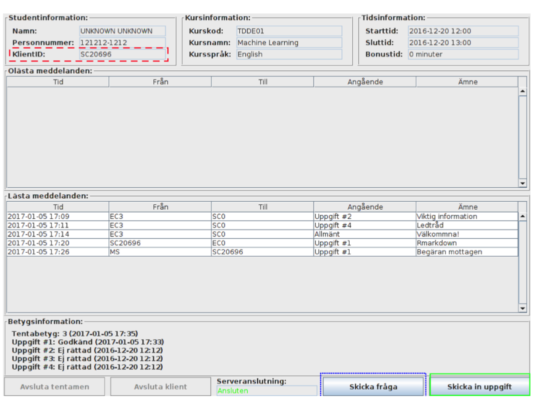
LINKÖPING UNIVERSITY
Dept. of Computer and Information Science
Division of Statistics and Machine Learning
Per Sidén

2018-11-01
Bayesian Learning, 6 hp
732A91/TDDE07

Computer Exam - Bayesian Learning (732A91/TDDE07), 6 hp

Time:	8-12
Allowable material:	- The allowed material in the folder given_files in the exam system. - Calculator with erased memory.
Teacher:	Mattias Villani. Phone: 070 – 0895205 and through the Communication client.
Exam scores:	Maximum number of credits on the exam: 40. Maximum number of credits on each exam question: 10.
Grades (732A91):	A: 36 points B: 32 points C: 24 points D: 20 points E: 16 points F: <16 points
Grades (TDDE07):	5: 34 points 4: 26 points 3: 18 points U: <18 points

INSTRUCTIONS:
When asked to give a solution on **Paper**, give that answer on physical papers supplied with the exam.
Each submitted sheet of paper should be marked with your *Client ID* from the *Communication Client*.
The client ID is the code in the `red` dashed rectangle in figure below.
All other answers should be submitted in a single PDF file using the *Communication Client.*
Include important code needed to grade the exam (inline or at the end of the PDF).
Submission starts by clicking the button in the `green` solid rectangle in figure below.
The submitted PDF file should be named *BayesExam.pdf*
Questions can be asked through the Communication client (`blue` dotted rectangle in figure below).
Full score requires clear and well motivated answers.



1. Prediction and decision

   A shopping mall considers replacing a staircase with an escalator. Data has been collected on the maximal weight y (in kg) on the staircase during four different weeks: $y_1 = 1690$, $y_2 = 1790$, $y_3 = 1760$ and $y_4 = 1750$. Assume the following model for these measurements: $y_1, ..., y_4|\theta \sim N(\theta, \sigma^2)$, where σ^2 is assumed known at $\sigma^2 = 50^2$. Assume a non-informative prior.

   (a) *Credits: 3p.* Simulate 1000 draws from the predictive distribution of the maximal weight in a given future week, and plot them.
   Solution: The predictive distribution is $N(\bar{y}, \sigma^2/4 + \sigma^2) = N(1747.5, 55.9^2)$, according to the slides of Lecture 4. See `Exam732A91_181101_Sol.R`.

   (b) *Credits: 2p.* Use simulation to approximate the expected number of weeks out of the coming 52 weeks in which the maximal weight will exceed 1850 kg, based on the predictive distribution.
   Solution: See `Exam732A91_181101_Sol.R`.

   (c) *Credits: 5p.* The weight that the escalator can hold at any given time is given by

   $$1000 \log(a),$$

   where a is the build cost. If the weight is exceeded the escalator breaks and has to be repaired. The loss function for the shopping mall is

   $$L(a, \theta) = a + n(a, \theta),$$

   where $n(a, \theta)$ is the number of weeks out of the coming 52 in which the escalator breaks. Compute the optimal build cost (a) using a Bayesian approach.
   Solution: We choose a to minimize the posterior expected loss

   $$E[L(a, \theta)] = a + E(n(a, \theta)).$$

   See `Exam732A91_181101_Sol.R`.

2. Regression

   The file `fish` which is loaded by the code in `ExamData.R` contains experimental data on 44 different fish. For each fish we have observed the length (mm), the age (days) and the temperature (temp) of the water tank in which the fish has grown (degrees Celsius). The dataframe also contains a column intercept with ones to get an intercept in the model. Now, use `BayesLinReg.R` to sample from the joint posterior distribution in the Gaussian linear regression

   $$\text{length} = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{temp} + \beta_3 \cdot \text{age}^2 + \beta_4 \cdot \text{temp}^2 + \beta_5 \cdot \text{age} \cdot \text{temp} + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

   Analyze the dataset by simulating 5000 draws from the joint posterior. Use the prior with $\boldsymbol{\mu}_0 = (0, 0, 0, 0, 0, 0)$, $\Omega_0 = 0.01 \cdot I_6$, $\nu_0 = 1$ and $\sigma_0^2 = 10000$.

   (a) *Credits: 2p.* Compute the posterior mean and 95% equal tail credible intervals for all β-parameters.
   Solution: See the code in `Exam732A91_181101_Sol.R`.

   (b) *Credits: 1p.* Compute the posterior mean and posterior median of the noise standard deviation σ.
   Solution: See the code in `Exam732A91_181101_Sol.R`.

   (c) *Credits: 4p.* The first eleven data points all come from a water tank with temperature 25 degrees Celsius. Produce a scatter plot of these data points with length and age on the two axes. Overlay a curve for the posterior mean of the regression function with respect to age

   $$f(\text{age}) = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{temp} + \beta_3 \cdot \text{age}^2 + \beta_4 \cdot \text{temp}^2 + \beta_5 \cdot \text{age} \cdot \text{temp},$$

   with temp = 25. The plot should also include a 95% (point-wise) equal tail credible interval for $f(\text{age})$.
   Solution: See the code in `Exam732A91_181101_Sol.R`.

(d) *Credits: 3p.* Assume that you want to make predictions for fish in a new water tank with a temperature of 15 degrees Celsius, which is lower than any of the temperatures in the original data set. Discuss on **Paper** what could be a potential problem with using the estimated model for prediction and how the prior could be changed to control this problem. Be brief and concise in your answer.
Solution: The given second order model is flexible and could easily overfit. Using it for covariate values far from the values used to fit the model could therefore lead to bad predictions. One way to reduce the risk of overfitting is to use a stronger prior that estimated β should be small. This is accomplished in this model by increasing the diagonal values of Ω_0 apart from element $(1,1)$ which corresponds to the intercept.

3. THE INVERSE GAMMA DISTRIBUTION

Let $x_1, ..., x_n|\alpha, \beta \overset{iid}{\sim} \text{Inv} - \text{gamma}(\alpha, \beta)$ be Inverse-gamma distributed data. This problem should only be solved on **Paper**.

(a) *Credits: 4p.* Consider the case when α is known and show that $\beta \sim \text{Gamma}(\gamma, \delta)$ is the conjugate prior for β for independent Inverse-gamma distributed data and that the posterior is a $\text{Gamma}(n\alpha + \gamma, \sum_{i=1}^n \frac{1}{x_i} + \delta)$.
Solution: Let $\beta \sim \text{Gamma}(\gamma, \delta)$ with density

$$p(\beta) = \frac{\delta^\gamma}{\Gamma(\gamma)} \beta^{\gamma-1} e^{-\delta\beta}.$$

We have the likelihood of one datapoint defined as

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

By Bayes' theorem and independence

$$\begin{aligned} p(\beta|x_1, ..., x_n) &\propto p(x_1, ..., x_n|\beta)p(\beta) \\ &\propto \left(\prod_{i=1}^n \beta^\alpha e^{-\beta/x_i} \right) \beta^{\gamma-1} e^{-\delta\beta} \\ &\propto \beta^{n\alpha+\gamma-1} e^{-\beta\left(\sum_{i=1}^n \frac{1}{x_i} + \delta\right)}, \end{aligned}$$

which is proportional to the $\text{Gamma}(n\alpha + \gamma, \sum_{i=1}^n \frac{1}{x_i} + \delta)$ density. Since the posterior belongs to the same distributional family (Gamma) as the prior, the prior is indeed conjugate to the Inverse-gamma data model (likelihood).

(b) *Credits: 4p.* Show that the marginal likelihood of the data for the Inverse-gamma model with Gamma prior can be written as

$$\frac{\Gamma(n\alpha + \gamma)\delta^\gamma \prod_{i=1}^n x_i^{-(\alpha+1)}}{\Gamma(\alpha)^n \Gamma(\gamma) \left(\sum_{i=1}^n \frac{1}{x_i} + \delta \right)^{n\alpha+\gamma}}$$

when α is known.

Solution: The marginal likelihood can be computed using

$$\begin{aligned}
p\left(x_1,...,x_n\right) &= \frac{p(x_1,...,x_n|\beta)p\left(\beta\right)}{p\left(\beta|x_1,...,x_n\right)} \\
&= \frac{\prod_{i=1}^n \frac{\beta^\alpha}{\Gamma(\alpha)} x_i^{-(\alpha+1)} e^{-\beta/x_i} \cdot \frac{\delta^\gamma}{\Gamma(\gamma)} \beta^{\gamma-1} e^{-\delta\beta}}{\frac{\left(\sum_{i=1}^n \frac{1}{x_i}+\delta\right)^{n\alpha+\gamma}}{\Gamma(n\alpha+\gamma)} \beta^{n\alpha+\gamma-1} e^{-\beta\left(\sum_{i=1}^n \frac{1}{x_i}+\delta\right)}} \\
&= \frac{\frac{1}{\Gamma(\alpha)^n} \frac{\delta^\gamma}{\Gamma(\gamma)} \prod_{i=1}^n x_i^{-(\alpha+1)}}{\frac{\left(\sum_{i=1}^n \frac{1}{x_i}+\delta\right)^{n\alpha+\gamma}}{\Gamma(n\alpha+\gamma)}} \\
&= \frac{\Gamma(n\alpha+\gamma)\delta^\gamma \prod_{i=1}^n x_i^{-(\alpha+1)}}{\Gamma(\alpha)^n \Gamma(\gamma) \left(\sum_{i=1}^n \frac{1}{x_i}+\delta\right)^{n\alpha+\gamma}}.
\end{aligned}$$

(c) *Credits: 2p.* Discuss one way to numerically approximate the marginal likelihood of the data for the Inverse-gamma model when α is known if the prior is not conjugate, for example when $\beta \sim \text{lognormal}(\mu, \sigma^2)$.
Solution: A simple approach is to use the Laplace approximation. That is, compute the posterior mode $\hat{\beta}$ and observed information $J_{\hat{\beta},\boldsymbol{x}}$ numerically by optimizing the unnormalized log-posterior. The Laplace approximation is then

$$\log \hat{p}\left(\boldsymbol{x}\right) = \log\left(\boldsymbol{x}|\hat{\beta}\right) + \log\left(\hat{\beta}\right) + \frac{1}{2}\log\left|J_{\hat{\beta},\boldsymbol{x}}^{-1}\right| + \frac{p}{2}\log 2\pi,$$

where the first term is the Inverse-gamma log-likelihood function and the second term is the log-prior.

4. Metropolis for Weibull

Let $x_1,...,x_n|\alpha,\beta \overset{iid}{\sim} \text{Weibull}(\alpha,\beta)$ be Weibull distributed data. The file `weibull` which is loaded by the code in `ExamData.R` contains 50 observations from the Weibull distribution with unknown α and β. Assume the following prior for α and β

$$p\left(\alpha,\beta\right) \propto \left(\frac{1}{\alpha\beta}\right)^2.$$

(a) *Credits: 5p.* Use numerical optimization to obtain a normal approximation of the *joint* posterior distribution of α and β. You don't need to plot the distribution, just provide its mean and covariance matrix. [Hints: use the argument `lower` in `optim`, and `method=c("L-BFGS-B")`].
Solution: See the code in `Exam732A91_181101_Sol.R`.

(b) *Credits: 5p.* Simulate from the actual posterior using the Metropolis algorithm. Denote $\theta = (\alpha,\beta)^T$ and use as proposal density the multivariate normal density (random walk Metropolis):

$$\theta_p|\theta_c \sim N\left(\theta_c, \tilde{c}\cdot\Sigma\right),$$

where Σ is the normal approximation covariance matrix obtained in (a) (if (a) was not solved you can use $\Sigma = \text{diag}(0.01, 0.1)$), and θ_c is the current draw. Use $\alpha = 1$, $\beta = 1$ as starting values, 500 iterations burn-in and thereafter draw 2000 samples from the posterior. Run the algorithm for three different values of $\tilde{c}$: 0.1, 4 and 100, and use the draws from the best choice of $\tilde{c}$; motivate your choice of $\tilde{c}$. Compute the posterior mean and variance for the two parameters based on your samples. [Hint: The proposal distribution can be truncated to avoid proposals of $\alpha \leq 0$ or $\beta \leq 0$ which can lead to numerical errors, use e.g. `theta[theta<=0]=1e-6`.]
Solution: See the code in `Exam732A91_181101_Sol.R`.

Good luck!

Mattias and Per