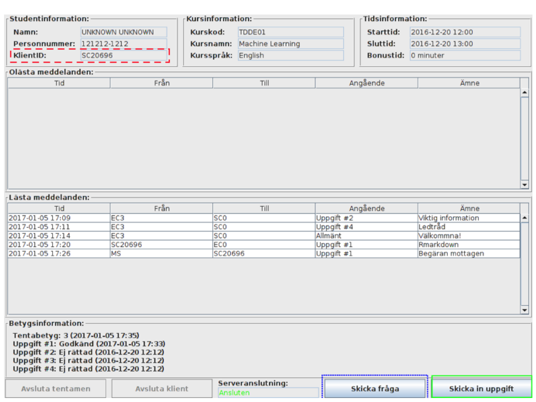
LINKÖPING UNIVERSITY
Dept. of Computer and Information Science
Division of Statistics and Machine Learning
Per Sidén

2018-06-01
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:	Per Sidén. Phone: $070 - 4977175$ 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. Poisson predictions

Let $x_1, \ldots x_n | \mu \overset{iid}{\sim} \text{Poisson}(\mu)$ be observations from the Poisson distribution. Assume that $n = 50$ and that the average of the observations is $\bar{x} = 10$.

(a) *Credits: 4p.* Assume a Gamma (α, β) prior for μ with $\beta = 2$ and with α such that the prior mean equals the posterior mean. Draw 1000 samples from the prior distribution and 1000 samples from the posterior distribution. Plot the prior and posterior distributions using both the samples and their analytical expressions.
Solution: We know from the lecture slides that the conjugate prior is $\mu \sim \text{Gamma}(\alpha, \beta)$ giving the posterior $\mu | x_1, ..., x_n \sim \text{Gamma}(\alpha + \sum_{i=1}^n x_i, \beta + n)$, and the data is: $n = 50$ and $\sum_{i=1}^n x_i = n \cdot \bar{x} = 500$. We need to determine α in the prior. From the properties of the Gamma distribution

$$E(\mu) = \frac{\alpha}{\beta} = \frac{\alpha}{2}$$
$$E(\mu|x) = \frac{\alpha + 500}{\beta + 50} = \frac{\alpha + 500}{52}.$$

Setting the expectations equal to each other gives $\alpha = 20$. The posterior is therefore $\mu | x_1, ..., x_n \sim \text{Gamma}(520, 52)$. See also `Exam732A91_180601_Sol.R`.

(b) *Credits: 2p.* Simulate 1000 draws from the predictive distribution of a new observation, x_{51}, and plot the distribution using the samples.
Solution: We can simulate draws from the posterior predictive distribution of x_{51} by repeatedly simulating i) from the posterior $\mu | x_1, ..., x_n$ followed by ii) simulation from the model $x_{51} | \mu \sim \text{Poisson}(\mu)$. See also `Exam732A91_180601_Sol.R`.

(c) *Credits: 2p.* What is the probability that $x_{51} = 10$, based on the posterior predictive distribution?
Solution: See `Exam732A91_180601_Sol.R`.

(d) *Credits: 2p.* Explain on **Paper** the main difference between making predictions for this model in the Bayesian way using the posterior predictive distribution with the classical approach of using $x_{51} \sim \text{Poisson}(\hat{\mu})$, where $\hat{\mu}$ is the maximum likelihood estimate. Also compare the predictive variance in the two cases. Be brief but mathematically detailed in your answer.
Solution: The main difference is that the posterior predictive distribution averages over all possible values of μ according to the posterior distribution through

$$p(x_{51}|x_1, ..., x_n) = \int p(x_{51}|\mu) \, p(\mu|x_1, ..., x_n) d\mu$$

while the classical approach just uses a point estimate in $p(x_{51}|\hat{\mu})$. The variance of posterior predictive distribution will be larger since the classical approach does not account for the uncertainty about the parameter values. In this particular model this can be seen by comparing

$$Var(x_{51}|\hat{\mu}) = \hat{\mu} = \bar{x} = 10$$

with

$$\begin{aligned} Var(x_{51}|x_1, \ldots, x_n) &= E_{\mu|x_1,\ldots,x_n}(Var[x_{51}|\mu]) + Var_{\mu|x_1,\ldots,x_n}(E[x_{51}|\mu]) \\ &= E_{\mu|x_1,\ldots,x_n}(\mu) + Var_{\mu|x_1,\ldots,x_n}(\mu) \\ &= \frac{\alpha + n\bar{x}}{\beta + n} + \frac{\alpha + n\bar{x}}{(\beta + n)^2} = \frac{520}{52} + \frac{520}{52^2} = 10.2. \end{aligned}$$

See also `Exam732A91_180601_Sol.R`.

2. Regression

The file `fish` which is loaded by the code in `ExamData1.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

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

Analyze the dataset by simulating 5000 draws from the joint posterior. Use the prior with $\boldsymbol{\mu}_0 = (0,0,0)$, $\Omega_0 = 0.01 \cdot I_3$, $\nu_0 = 1$ and $\sigma_0^2 = 10000$.

(a) *Credits: 2p.* Plot the marginal posterior distribution of each parameter.
Solution: See the code in `Exam732A91_180601_Sol.R`.

(b) *Credits: 2p.* Construct 90% equal tail probability interval for β_1 and interpret it.
Solution: See the code in `Exam732A91_180601_Sol.R`.

(c) *Credits: 2p.* Assume that an earlier research article shows strong evidence that the water temperature has no effect on the length of fish. Give a brief but mathematically detailed discussion on **Paper** about how you can change the model to include this information.
Solution: The Bayesian way to include this information through the prior for β_2. The original prior is $\beta_2|\sigma^2 \sim N(0, 100\sigma^2)$, but we could change this e.g. to $\beta_2|\sigma^2 \sim N(0, 0.001\sigma^2)$ by setting $\Omega_0(3,3) = 1000$.

(d) *Credits: 4p.* In a new experiment, fish have been grown in water tank with water temperature 30 degrees Celsius. Newborn fish have been inserted into the tank at two time points, 30 days ago and 100 days ago. Assume that the tank is populated by an equal amount of fish of the two different ages. You pick up a fish randomly from the water tank. Do a Bayesian analysis (using simulation methods) to determine the predictive distribution of the length of the picked up fish.
Solution: See the code in `Exam732A91_180601_Sol.R`.

3. Binomial model comparison

Let $x|n, p \sim \text{Bin}(n, p)$ be an observation from the binomial distribution, where n is known. This problem should only be solved on **Paper**, except for perhaps any numerical computations in (c).

(a) *Credits: 3p.* Compute the posterior distribution for p when the prior $p \sim \text{Beta}(\alpha, \beta)$ is used.
Solution: Let $p \sim \text{Beta}(\alpha, \beta)$ with density

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

We have the likelihood defined as

$$p(x|p) = \begin{pmatrix} n \\ x \end{pmatrix} p^x (1-p)^{n-x}.$$

Using Bayes' theorem we get

$$\begin{aligned} p(p|x) &\propto p^x (1-p)^{n-x} p^{\alpha-1} (1-p)^{\beta-1} \\ &\propto p^{x+\alpha-1} (1-p)^{n-x+\beta-1}, \end{aligned}$$

where we can identify the form of a Beta distribution, so we can conclude that the posterior is $p|x \sim \text{Beta}(x+\alpha, n-x+\beta)$.

(b) *Credits: 3p.* Show that the marginal likelihood of the data for the binomial model with Beta prior can be expressed as

$$p(x) = \frac{\begin{pmatrix} n \\ x \end{pmatrix} \Gamma(\alpha+\beta)\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n)}.$$

Solution: The marginal likelihood can be computed using

$$
\begin{aligned}
p(x) &= \frac{p(x|p)p(p)}{p(p|x)} \\
&= \frac{\binom{n}{x} p^x (1-p)^{n-x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}}{\frac{\Gamma(x+\alpha+n-x+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} p^{x+\alpha-1}(1-p)^{n-x+\beta-1}} \\
&= \frac{\binom{n}{x} \Gamma(\alpha+\beta)\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n)}.
\end{aligned}
$$

(c) *Credits: 4p.* Assume that $x = 3$ and $n = 10$. Do a Bayesian model comparison of the following three models, which all have the binomial likelihood, and where the first and second use the Beta prior and the third is a null model which assumes that p is known with $p = 0.5$. Assume that the prior probability for each model is the same, that is, $p(M_1) = p(M_2) = p(M_3) = \frac{1}{3}$. State your conclusions. [Hint: Any numerical computations can be carried out and reported in the submitted `R`-code. See the formulas for the Gamma- and Beta-function in the statistics and math results (provided with the exam) and the functions `gamma`, `beta` and `choose` in `R`.]

$$
\begin{aligned}
&M_1\colon\ p \sim \text{Beta}(1,1) \\
&M_2\colon\ p \sim \text{Beta}(4,4) \\
&M_3\colon\ p = 0.5.
\end{aligned}
$$

.

Solution: The Bayesian model comparison is carried out by computing the marginal likelihood for the three models, and then using the posterior model probabilities

$$p(M_i|x) \propto p(x|M_i)\, p(M_i).$$

Model 3 has no unknown parameters so the marginal likelihood reduces to the likelihood as defined in (a), while the marginal likelihood for Model 1 and 2 can be computed using the result in (b). See the numerical computations in `Exam732A91_180601_Sol.R`. The computed posterior model probabilities are $(0.27, 0.37, 0.35)$ for the three models, so Model 2 is the most probable given the data.

4. Censored normal data

The emissions of sulfur dioxide in mg/Nm^3 from a power plant is measured every day for a month and the data can be found in the file `sulfur` which is loaded by the code in `ExamData1.R`. The measurement device can only register values above 200, otherwise the value is recorded as 200.

(a) *Credits: 4p.* Consider only the data points for which the recorded value was larger than 200 by discarding all data points with value exactly 200. The remaining data points are assumed to be independent and follow a truncated normal distribution with density

$$p(x|\mu,\sigma) = \frac{\phi\left(\frac{x-\mu}{\sigma}\right)}{\sigma\left(1-\Phi\left(\frac{L-\mu}{\sigma}\right)\right)} \quad \text{for } x > L,$$

where $\phi(x)$ is the standard normal probability density function (pdf) and $\Phi(x)$ is the standard normal cumulative distribution function (cdf). $L = 200$ is the lower truncation point. Write a function in R that computes the (unnormalized) log posterior distribution of μ based on iid observations when σ has known value $\sigma = 100$. Assume a constant prior for μ. Use the function to plot the posterior distribution of μ for the observations greater than 200 in the data vector `sulfur`. For the plot, use a grid constructed in `R` with `seq(100,400,1)`.
Solution: See the code in `Exam732A91_180601_Sol.R`.

(b) *Credits: 3p.* Now consider all the data points in `sulfur` and assume they are iid normal observations

$$x_i \overset{iid}{\sim} N\left(\mu, \sigma^2\right),$$

but with the values below 200 being censored and set to 200, and where σ is now considered unknown. The supplied stan model `censModel` can be used to simulate from the posterior of μ and σ, and also from the posterior of the true values of the 8 censored data points. Run the stan program by calling `stan(model_code=censModel, data=censData, ...`. Evaluate the convergence of the sampler using graphical methods. Plot the posterior of μ and σ using the simulated samples.
Solution: See the code in `Exam732A91_180601_Sol.R`.

(c) *Credits: 3p.* In this part, instead consider the time series model

$$\begin{aligned} x_i|z_i &\overset{iid}{\sim} N\left(z_i, 20^2\right), \\ z_t &= \mu + \phi\left(z_{t-1} - \mu\right) + \varepsilon_t, \\ \varepsilon_t &\overset{iid}{\sim} N\left(0, \sigma^2\right), \end{aligned}$$

that is, assume that the observations follows an independent normal distribution when conditioned on a latent AR(1)-process z, but with the values of x_i below 200 being censored and set to 200. Modify the stan code in order to do inference for this model instead. Also put a normal prior on $\mu \sim N(300, 100^2)$. Plot the posterior of ϕ. Also produce a plot that contains both the data and the posterior mean and 95% credible intervals for the latent intensity z over time.
Solution: See the code in `Exam732A91_180601_Sol.R`.

Good luck!

Mattias and Per