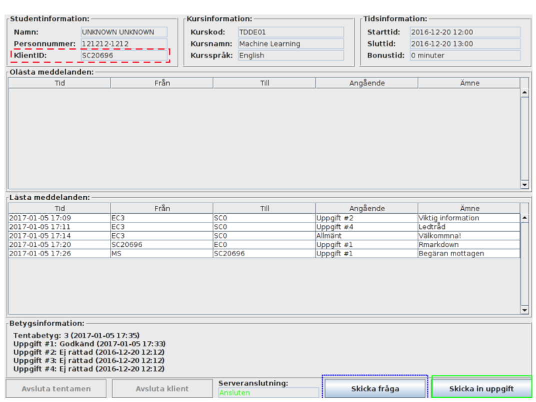
LINKÖPING UNIVERSITY
Dept. of Computer and Information Science
Division of Statistics and Machine Learning
Per Sidén

2018-08-22
Bayesian Learning, 6 hp
732A91/TDDE07

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

Time:	14-18
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. The log-normal model

The file `lions` which is loaded by the code in `ExamData.R` contains the reported weights in kg of 92 captured male lions. Let $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ denote the weights and assume a log-normal model $x_1, \ldots, x_n|\mu, \sigma^2 \overset{iid}{\sim} \text{lognormal}\left(\mu, \sigma^2\right)$.

(a) *Credits: 4p.* Assume for now that we know that $\sigma^2 = 0.04$. Plot the posterior distribution of μ based on the data in the supplied file `lions.RData`. Use the prior $\mu \sim N(5, 1)$. Make sure that the plotted posterior is a valid (normalized) density function. [Hint: The lognormal density function is available in `R` as `dlnorm`.]
Solution: See `Exam732A91_180822_Sol.R`.

(b) *Credits: 4p.* Now assume that also σ^2 is unknown and that $\sigma^2 \sim \textit{Inv-}\chi^2(\nu_0, \sigma_0^2)$ a priori independently from μ, with $\nu_0 = 5$ and $\sigma_0^2 = 0.04$. Implement Stan-code that produces at least 2000 samples from the posterior of μ and σ^2. Use 500 samples for burnin. Based on the samples compute the posterior mean and standard deviation of μ and σ^2 and plot the joint posterior of μ and σ^2. [Hint: The lognormal sampling statement in Stan is `y~lognormal(mu,sigma);`].
Solution: See `Exam732A91_180822_Sol.R`.

(c) *Credits: 2p.* It is of interest to compute an estimate of the average weight of male lions. Give an estimate and a 95% credible interval of the average weight of male lions based on the posterior computed in (b). [Hint: The mean of the lognormal distribution is $\exp\left(\mu + \frac{1}{2}\sigma^2\right)$.].
Solution: See `Exam732A91_180822_Sol.R`.

2. Logistic regression

The file `titanic` which is loaded by the code in `ExamData.R` contains the information about 1316 passengers on the ship Titanic that sunk in the Atlantic ocean in April 1912. This was a big disaster. Descriptions of the variables presented in the table below. A passenger could travel in first, second or third class.

Variable	**Data type**	**Meaning**	**Role**
Survived	Binary	Whether or not the passenger survived	Response
Intercept	1	Constant to the intercept	Feature
Adult	Binary	Whether or not the passenger was an adult	Feature
Man	Binary	Whether or not the passenger was a man	Feature
Class1	Binary	Whether or not the passenger travelled in first class	Feature
Class2	Binary	Whether or not the passenger travelled in second class	Feature

(a) *Credits: 4p.* Consider the logistic regression

$$\Pr(y = 1|\mathbf{x}) = \frac{\exp\left(\mathbf{x}^T\boldsymbol{\beta}\right)}{1 + \exp\left(\mathbf{x}^T\boldsymbol{\beta}\right)},$$

where y is the binary variable with $y = 1$ if the passenger survived and $y = 0$ otherwise. $\mathbf{x}$ is a 5-dimensional vector containing the five features (including a one for the constant term that models the intercept) and $\boldsymbol{\beta}$ is the corresponding vector of coefficients. Use numerical optimization to obtain a normal approximation of the *joint* posterior distribution of $\boldsymbol{\beta}$. Use the prior $\boldsymbol{\beta} \sim \mathcal{N}(0, \tau^2 I)$, with $\tau = 50$. Plot the marginal posterior distribution of each variable except the intercept, using your approximation. For full points, use the probability density function (pdf) to construct the plots.
Solution: See `Exam732A91_180822_Sol.R`.

(b) *Credits: 2p.* Compute the posterior probability that the β_i corresponding to the adult feature is smaller than 0. Interpret this probability.
Solution: See `Exam732A91_180822_Sol.R`.

(c) *Credits: 4p.* A first class adult woman and a third class adult man are together during the disaster. Compute the predictive probability that the woman survives but the man dies.
Solution: See `Exam732A91_180822_Sol.R`.

3. Geometric model comparison

Let $x_1, ..., x_n|\theta \overset{iid}{\sim}$ Geometric(θ). The Geometric distribution has probability function

$$p(x|\theta) = (1-\theta)^x \theta, \text{ for } x = 0, 1, 2, ...,$$

and zero otherwise.

(a) *Credits: 3p.* Derive the posterior distribution $p(\theta|x_1, ..., x_n)$ on **Paper** using the conjugate Beta(α, β) prior.
Solution:

$$\begin{aligned} p(\theta|x_1, ..., x_n) &\propto p(x_1, ..., x_n|\theta)p(\theta) \\ &\propto (1-\theta)^{\sum_{i=1}^n x_i} \theta^n \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{n+\alpha-1}(1-\theta)^{\sum_{i=1}^n x_i + \beta - 1} \end{aligned}$$

which is proportional to the Beta($n+\alpha, n\bar{x}+\beta$) density, where $\bar{x}$ is the data mean.

(b) *Credits: 3p.* Derive the marginal likelihood of the data for the geometric model with Beta prior on **Paper** .
Solution: The marginal likelihood can be computed using

$$\begin{aligned} p(x_1, \ldots, x_n) &= \frac{p(x_1, \ldots, x_n|\theta)p(\theta)}{p(\theta|x_1, \ldots, x_n)} \\ &= \frac{(1-\theta)^{n\bar{x}} \theta^n \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}}{\frac{\Gamma(n+\alpha+n\bar{x}+\beta)}{\Gamma(n+\alpha)\Gamma(n\bar{x}+\beta)} \theta^{n+\alpha-1}(1-\theta)^{n\bar{x}+\beta-1}} \\ &= \frac{\Gamma(\alpha+\beta)\Gamma(n+\alpha)\Gamma(n\bar{x}+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(n+n\bar{x}+\alpha+\beta)}. \end{aligned}$$

(c) *Credits: 4p.* Assume that we have $n=3$ observations, $x_1 = 2$, $x_2 = 1$ and $x_3 = 12$. Do a Bayesian model comparison of the following two models, which both have the geometric likelihood, and where the first uses the Beta prior with $\alpha = 0.5$ and $\beta = 0.5$ and the second is a null model which assumes that θ is known with $\theta = 0.5$. Assume the prior probabilities $p(M_1) = \frac{1}{10}$ and $p(M_2) = \frac{9}{10}$. State your conclusions.
Solution: The Bayesian model comparison is carried out by computing the marginal likelihood for the two models, and then using the posterior model probabilities

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

Model 2 has no unknown parameters so the marginal likelihood reduces to the likelihood as defined in (a), while the marginal likelihood for Model 1 can be computed using the result in (b) (this can also be found in the Lecture 10 slides if (b) was not solved). See the numerical computations in `Exam732A91_180822_Sol.R`. The computed posterior model probabilities are $(0.62, 0.38)$ for the two models, so Model 1 is the most probable given the data.

4. Election

The day before a political election, an opinion poll asked 400 randomly selected registered voters about which party they were going to vote for. There exists only three parties, and the answers were distributed in the following way (party A: 184, party B: 67, party C: 149).

(a) *Credits: 3p.* Assume a multinomial model for the data and a non-informative Dirichlet prior. On **Paper**, derive the posterior distribution of the underlying share of voters for each party.
Solution: The multinomial model is

$$p(y|\theta) \propto \prod_{k=1}^{K} \theta_k^{y_k}, \text{ where } \sum_{k=1}^{K} \theta_k = 1,$$

where y_k is the number of answers for party k and θ_k is the underlying share of voters for party k. The non-informative Dirichlet prior is

$$p(\theta) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1},$$

where we set $\alpha_1 = \alpha_2 = \alpha_3 = 1$ to make this prior non-informative. The posterior is

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ &\propto \prod_{k=1}^{K} \theta_k^{y_k} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \\ &\propto \prod_{k=1}^{K} \theta_k^{y_k + \alpha_k - 1}. \end{aligned}$$

So, $\theta|y \sim \text{Dirichlet}(\alpha_1 + y_1, \alpha_2 + y_2, \alpha_3 + y_3) = \text{Dirichlet}(185, 68, 150)$.

(b) *Credits: 1p.* Compute the probability that party A gets a majority of the votes (more than 50%) in the election. Assume that everyone in the population is voting.
Solution: See `Exam732A91_180822_Sol.R`.

(c) *Credits: 2p.* Compute the probability that party A becomes the largest party. Assume that everyone in the population is voting.
Solution: See `Exam732A91_180822_Sol.R`.

(d) *Credits: 2p.* Assume that the probability in (c) was estimated using Monte Carlo with 10000 independent posterior samples. If (c) was not solved, assume that the answer was 0.9. Compute a 95% confidence interval for the estimated probability in (c) with respect to the error from the Monte Carlo estimation.
Solution: The Monte Carlo estimate can be seen as an average of 10000 Bernoulli random variables and the central limit theorem gives that this average is approximately normal. So the standard deviation of the estimate is

$$\sqrt{\frac{1}{10000}\,(1-0.9)\,0.9} = 0.003.$$

So a 95% confidence interval is $0.9 \pm 1.96 \cdot 0.003 = (0.894, 0.906)$. Also see `Exam732A91_180822_Sol.R`.

(e) *Credits: 2p.* How many additional samples would be needed to reduce the width of the interval in (d) by half?
Solution: According to the slides of Lecture 8, the variance of a Monte Carlo estimate with N iid samples is σ^2/N, where σ^2 is the variance of each sample. The width of the interval in (d) is $2\sqrt{\sigma^2/10000}$, so to get the number of samples needed to reduce the width by half we solve

$$\begin{aligned} \sqrt{\sigma^2/10000} &= 2\sqrt{\sigma^2/N} \\ &\Rightarrow N = 40000. \end{aligned}$$

So we need 30000 additional samples.

Good luck!

Mattias and Per