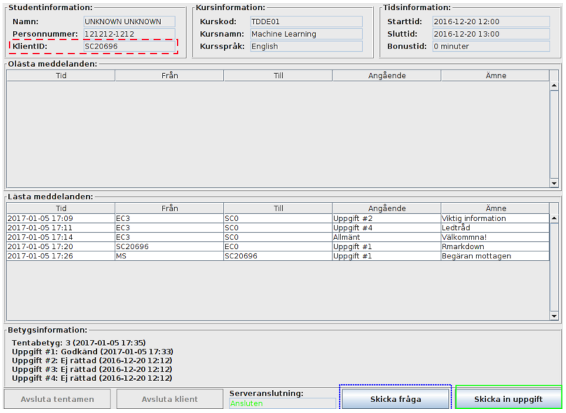
LINKÖPING UNIVERSITY
Dept. of Computer and Information Science
Division of Statistics and Machine Learning
Mattias Villani

2017-05-30
Bayesian Learning, 6 hp
732A91/TDDE07

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

Time:	8-12 AM
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*
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 credit requires clear and well motivated answers.



Studentinformation:
Namn: UNKNOWN UNKNOWN
Personnummer: 121212-1212
KlientID: SC20696

Kursinformation:
Kurskod: TDDE01
Kursnamn: Machine Learning
Kursspråk: English

Tidsinformation:
Starttid: 2016-12-20 12:00
Sluttid: 2016-12-20 13:00
Bonustid: 0 minuter

Olästa meddelanden:

Tid	Från	Till	Angående	Ämne

Lästa meddelanden:

Tid	Från	Till	Angående	Ämne
2017-01-05 17:09	EC3	SC0	Uppgift #2	Viktig information
2017-01-05 17:11	EC3	SC0	Uppgift #4	Ledtråd
2017-01-05 17:14	EC3	SC0	Allmänt	Välkommna!
2017-01-05 17:20	SC20696	EC0	Uppgift #1	Rmarkdown
2017-01-05 17:26	MS	SC20696	Uppgift #1	Begäran mottagen

Betygsinformation:
Tentabetyg: 3 (2017-01-05 17:35)
Uppgift #1: Godkänd (2017-01-05 17:33)
Uppgift #2: Ej rättad (2016-12-20 12:12)
Uppgift #3: Ej rättad (2016-12-20 12:12)
Uppgift #4: Ej rättad (2016-12-20 12:12)

Serveranslutning: Ansluten

1. Bayesian inference for the rice distribution

A commonly occuring distribution for positive data is the *Rice distribution*, which we denote by $\text{Rice}(\theta, \psi)$. The PDF for a Rice distribution is of the form

$$p(x|\theta,\psi) = \frac{x}{\psi}\exp\left(\frac{-\left(x^2+\theta^2\right)}{2\psi}\right)\cdot I_0\left(\frac{x\theta}{\psi}\right) \quad \text{for } x > 0.$$

where $\theta \geq 0$ is the location parameter and $\psi > 0$ is related to the variance. $I_0(\cdot)$ is the modified Bessel function of the first kind and order zero, which is implemented in R as `BesselI`. We will assume for simplicity that $\psi = 1$.

(a) *Credits: 4p.* Write a function in R that computes the log posterior distribution of θ based on iid observations $\mathbf{x} = (x_1, ..., x_n)$ from $\text{Rice}(\theta, \psi = 1)$. Use that function to plot the posterior distribution of θ for the $n = 10$ observations in the data vector `riceData` in the supplied file `ExamData.R`.
Solution: See the code in `Exam732A91_170530_Sol.R`.

(b) *Credits: 3p.* Use numerical optimization to obtain a normal approximation of the posterior distribution of θ based on the data in `riceData`. Use the `lines` command in R to plot this approximate posterior in the same graph as the posterior obtained in 1a. [Hints: use the argument `lower` in `optim`, and `method=c("L-BFGS-B")`]. Is the approximation accurate?
Solution: See the code in `Exam732A91_170530_Sol.R`.

(c) *Credits: 3p.* Explain on **Paper** how the predictive distribution for a new observation x_{n+1} is computed by integration. You don't need to actually compute the integral, just give a general formula for the predictive distribution. Now, compute the predictive distribution for a new observation x_{n+1} by simulation. You can use the normal approximation of the posterior from 1b). A simulator (`rRice`) for the Rice distribution is provided in the file `ExamData.R`.
Solution: The predictive distribution is obtained by averaging the data distribution with respect to the posterior distribution. Since the data are iid

$$p(x_{n+1}|x_1, ..., x_n) = \int p(x_{n+1}|\theta, \psi = 1)p(\theta|x_1, ..., x_n)d\theta,$$

which can be simulated by iterating many time between:

i. simulating $\theta^{(i)}$ from the normal approximate posterior $p(\theta|x_1, ..., x_n)$

ii. simulating a data observation from the Rice distribution $p(x_{n+1}|\theta^{(i)}, \psi = 1)$. Note that we have $\theta = \theta^{(i)}$ here.

2. Modeling count data

The data set `bids` which is loaded by the code in `ExamData.R` contains data on the number of bids in 1000 eBay auctions for collectors coins. Let $x_1, ..., x_n$, for $n = 1000$, denote the data points.

(a) *Credits: 2p.* Assume the Poisson model $x_1, ..., x_n|\theta \overset{iid}{\sim} \text{Pois}(\theta)$ for the data, and use the prior $\theta \sim \text{Gamma}(1, 1)$. Compute the posterior distribution for θ and plot it.
Solution: See the code in `Exam732A91_170530_Sol.R`.

(b) *Credits: 2p.* Use graphical methods to investigate if the Poisson model fits the data well.
Solution: See the code in `Exam732A91_170530_Sol.R`.

(c) *Credits: 2p.* Use the supplied function `GibbsMixPois.R` in the file `ExamData.R` to do Gibbs sampling for a mixture of Poissons model

$$p(x) = \sum_{k=1}^{K} w_k \cdot \text{Pois}(x|\theta_k),$$

where $w_1, ..., w_K$ are the weights (probabilities) of the mixture components (sometimes also called denoted $\pi_1, ..., \pi_K$). $\text{Pois}(x|\theta_i)$ is here used as a shorthand for the probability function (density for

a discrete variables) of a Poisson distribution with mean θ_k in the kth mixture component. Use the same $\theta \sim \text{Gamma}(1,1)$ prior for all the K components, and a uniform prior on the weights $w_1, ..., w_K$. Estimate the mixture of Poissons both with $K = 2$ and $K = 3$. Use `nIter=500` draws, and no burn-in.
Solution: See the code in `Exam732A91_170530_Sol.R`.

(d) *Credits: 2p.* Use graphical methods to investigate if the mixture of Poissons with $K = 2$ fits the data well. Note that `GibbsMixPois.R` returns the posterior mean of the mixture density (`GibbsResults$mixDensMean`). Is $K = 2$ enough, or would you recommend $K = 3$?
Solution: See the code in `Exam732A91_170530_Sol.R`.

(e) *Credits: 2p.* The number of mixture components, K, is usually unknown. Discuss on **Paper** how a Bayesian could do inference for K. You do not need to compute anything here, just discuss the principles.
Solution: **Bayesian model comparison** can be used to compute the posterior distribution over K (which are different models)

$$\Pr(K = k|\mathbf{x}) \propto p(\mathbf{x}|K = k) \cdot \Pr(K = k),$$

where

$$p(\mathbf{x}|K = k) = \int p(\mathbf{x}|\theta_k, K = k)p(\theta_k|K = k)d\theta$$

is the **marginal likelihood** for the model with k mixture components and θ_k contains all the parameters in this model. $p(\mathbf{x}|\theta_k, K = k)$ is the usual likelihood for mixture model with k components, and $p(\theta_k|K = k)$ is just a prior on the parameters in the model with k components. Exactly how to compute these integrals is complicated, but can be done with more sophisticated MCMC.

3. Regression

`BayesLinReg.R` samples from the joint posterior of $\boldsymbol{\beta}$ and σ^2 in the Gaussian linear regression with conjugate prior

$$\begin{aligned}\boldsymbol{\beta}|\sigma^2 &\sim N\left(\boldsymbol{\mu}_0, \sigma^2\Omega_0^{-1}\right)\\ \sigma^2 &\sim \text{Inv}-\chi^2(\nu_0, \sigma_0^2).\end{aligned}$$

(a) The file `cars` which is loaded by the code in `ExamData.R` contains data on 32 cars. For each car we have observations on how many miles that car can travel on a gallon of gasoline (mpg), the weight of the car (weight) and two dummy variables that indicates if the car's engine has four cylinders (sixcyl=0 and eightcyl=0) six cylinders (sixcyl=1 and eightcyl=0) or eigth cylinders (sixcyl=0 and eightcyl=1). 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

$$\text{mpg} = \beta_0 + \beta_1 \cdot \text{weight} + \beta_2 \cdot \text{sixcyl} + \beta_3 \cdot \text{eightcyl} + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

Analyze the dataset by simulating 1000 draws from the joint posterior. Use the prior with $\boldsymbol{\mu}_0 = (0, 0, 0, 0)$, $\Omega_0 = 0.01 \cdot I_4$, $\nu_0 = 1$ and $\sigma_0^2 = 36$ (which is the data variance).

i. *Credits: 2p.* Plot the marginal distributions of each parameter.
Solution: See the code in `Exam732A91_170530_Sol.R`.

ii. *Credits: 2p.* Compute point estimates for each parameter assuming the linear loss function $L(\beta_k, a) = |\beta_k - a|$, where β_k is the kth regression coefficient.
Solution: See the code in `Exam732A91_170530_Sol.R`.

iii. *Credits: 2p.* Construct 95% equal tail probability intervals for each parameter and interpret them.
Solution: See the code in `Exam732A91_170530_Sol.R`.

(b) *Credits: 2p.* Investigate if the effect on mpg is different in cars with six cylinders compared to cars with 8 cylinders.
Solution: See the code in `Exam732A91_170530_Sol.R`.

(c) *Credits: 2p.* Compute by simulation the predictive distribution for a new 4 cylinder car with weight = 3.5.
Solution: See the code in `Exam732A91_170530_Sol.R`.

4. Geometric data and decisions

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($n+\alpha, \sum_{i=1}^n x_i + \beta$) density.

(b) *Credits: 3p.* Show on **Paper** that the predictive distribution for a new observation x_{n+1} is of the form

$$p(x_{n+1}|x_1,...,x_n) \propto \frac{\Gamma(x_{n+1}+\sum_{i=1}^n x_i+\beta)}{\Gamma(x_{n+1}+\sum_{i=1}^n x_i+n+\alpha+\beta+1)}.$$

Solution:

$$\begin{aligned} p(x_{n+1}|x_1,...,x_n) &= \int_0^1 p(x_{n+1}|\theta)p(\theta|x_1,...,x_n)d\theta \\ &\propto \int_0^1 (1-\theta)^{x_{n+1}}\theta\theta^{n+\alpha-1}(1-\theta)^{\sum_{i=1}^n x_i+\beta-1}d\theta \\ &= \int_0^1 (1-\theta)^{x_{n+1}+\sum_{i=1}^n x_i+\beta-1}\theta^{n+\alpha}d\theta. \end{aligned}$$

This integral can be computed by realizing that the integrand is proportional to a Beta($n+\alpha+1, x_{n+1}+\sum_{i=1}^n x_i+\beta$) density, or by using the definition of the Beta function from the collection of statistical and mathematical results. Either way, we obtain

$$p(x_{n+1}|x_1,...,x_n) \propto \frac{\Gamma(x_{n+1}+\sum_{i=1}^n x_i+\beta)\Gamma(n+\alpha+1)}{\Gamma(x_{n+1}+\sum_{i=1}^n x_i+n+\alpha+\beta+1)} \propto \frac{\Gamma(x_{n+1}+\sum_{i=1}^n x_i+\beta)}{\Gamma(x_{n+1}+\sum_{i=1}^n x_i+n+\alpha+\beta+1)}.$$

(c) *Credits: 4p.* Your favorite sports team has had following result in its first $n = 10$ games of the season (W=won, L=lost): $W, L, L, W, W, L, L, L, W, W$. Assume that the games are independent and that the team has the same chance of winning in every game. Your local bookie has introduced a new game where you win $2^k - 1$ dollars if your team loses the k subsequent games and then wins the $(k+1)$th game. The game costs \$2 dollars to play. Should you play it? Use a uniform prior wherever needed. [Hint: one way to solve this problem uses the results from 4b) above.]
Solution: (also in the code in `Exam732A91_170530_Sol.R.`). This is a decision problem with two actions (a): Play and NoPlay. The NoPlay action gives utility (money) regardless of the future games for your team. The Play action gives utility $(2^k - 1) - 2$ which depends on the unknown quantity k. The Bayesian solution to a decision problem is to choose the action that maximizes the posterior expected utility. To compute this expectation we need the posterior (predictive) distribution for k given the observed team performance in the last 10 games. Now, the geometric

distribution is the distribution for the number of failed Bernoulli trials before the first success appears, so that is a good model here. We need to turn the sequence $W, L, L, W, W, L, L, L, W, W$ into Geometric data by counting the number of lost games before each win. This gives the data: $x_1 = 0, x_2 = 2, x_3 = 0, x_4 = 3, x_5 = 0$. The uniform prior is the Beta($\alpha = 1, \beta = 1$) and $\sum_{i=1}^{n} x_i = 5$ and $n = 5$, so the predictive distribution for $k = x_8$ is

$$p(x_6|x_1, ..., x_5) \propto \frac{\Gamma(x_6 + 6)}{\Gamma(x_6 + 13)}$$

which is depicted in Figure 1. Note that we need to normalize this expression by dividing by the sum of all terms. Now, let $U(x_6, a = Play)$ denote the utility of choosing to play when the number of consequtive losses turn out to be x_6. The posterior expected utlity is then

$$EU_{Play} = \sum_{k=1}^{\infty} U(x_6 = k, Play) p(x_6 = k|x_1, ..., x_5) \approx \sum_{k=1}^{k_{max}} \left((2^k - 1) - 2\right) p(x_6 = k|x_1, ..., x_5),$$

for some upper truncation point on the number of terms in the sum. Looking at the predictive distribution for x_6, it seems that $k_{max} = 10$ would be enough number of terms (since $p(x_6 > 10|x_1, ..., x_5)$ is tiny). With $k_{max} = 10$ we get $EU_{play} = 7.416$ (see my solution code) which is larger than the expected utility for the NoPlay action (which is zero). So you should play the game. Note however that using $k_{max} = 20$ gives $EU_{play} = 355.241$. In fact, we can make EU_{play} as large as we want by increasing k_{max}. It turns out that EU_{Play} is actually infinite! I didn't think about this when constructing the problem (it should have been obvious to me, as I will explain now), and you will get full points for any value of k_{max} that is not too small ($p(x_6 > k_{max}|x_1, ..., x_5)$ should be very small).
The reason why $EU_{play} = \infty$ is related to the *St Petersburg paradox* (https://en.wikipedia.org/wiki/St._Petersburg_paradox). We can show this by doing additional work on the predictive distribution by making repeated use of Formula 1.18 for Gamma function in the 'Useful statistical and mathematical results' papers distributed at the exam. We can then rewrite the predictive distribution as

$$p(x_6 = k|x_1, ..., x_5) \propto \frac{\Gamma(k + 6)}{\Gamma(k + 13)} = \frac{1}{(k + 6)(k + 7)(k + 8)(k + 9)(k + 10)(k + 11)(k + 12)}.$$

For large k the predictive distribution behaves as (only the largest power will matter for large k):

$$p(x_6 = k|x_1, ..., x_5) \propto \frac{1}{k^7}.$$

This decays rapidly with k, and was the reason why I thought that $k_{max} = 10$ or $k_{max} = 20$ would be enough for EU_{Play}. But I should have realized [It was late at night, ok? :)] that the utility $(2^k - 1) - 2$ grows exponentially fast, and so will dominate. That is, the terms in EU_{Play} for large k will behave as $\frac{(2^k-1)-2}{k^7} \approx \frac{2^k}{k^7}$, and we have the well known result (at least if you know about it!)

$$\lim_{k \to \infty} \frac{2^k}{k^a} = \infty$$

for any a. This result says that 'Exponentials grow faster than any polynomial'. Since all terms in EU_{Play} are positive, and the large ones grow to infinity as k grows, the whole sum naturally grows to infinity. Again, you don't have to do all this to get full points for this problem as it was not my intention to make this extra complication.

Good luck!

Mattias

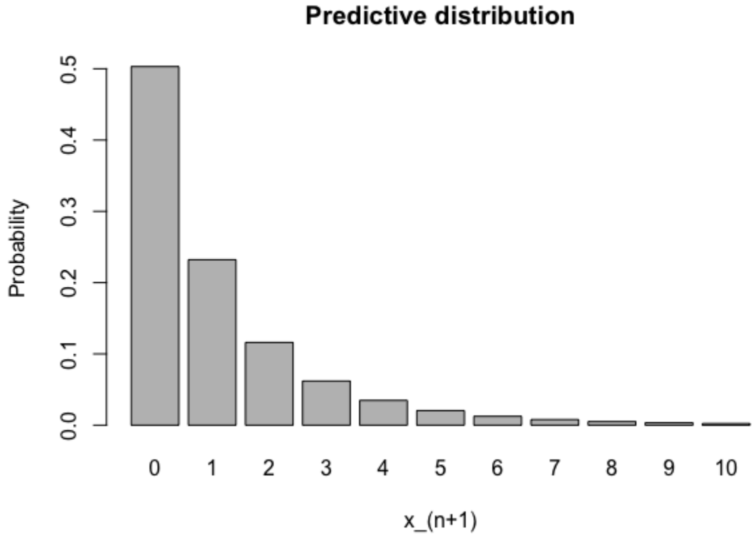


Figure 1: Predictive distribution in Problem 4c.