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

2020-10-22
Bayesian Learning, 6 hp
732A91/TDDE07/732A73

Take Home Computer Exam
Bayesian Learning (732A91/TDDE07/732A73), 6 hp

Time:	8-12
Allowable material:	All aids are permitted during the exam with the following exceptions: • You may not communicate with anyone else except the responsible teacher • You may not look at the solutions of any other students.
Teacher:	Per Sidén. Contact via email at per.siden@liu.se
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:
For full information, see the document *Info take home exam.pdf.*
You should submit your solutions via the Submission in LISAM.
Your submitted solutions must contain the following two things:
- A PDF named *ComputerSol.pdf* containing your computer-based solutions
- Files named *PaperSol* containing your hand-written solutions (when asked to answer on **Paper**)

Full score requires clear and well motivated answers.

---

1. CLASSIFIERS

A company has developed two different machine learning based classifiers, C_1 and C_2, for detecting whether an MR image of a brain contains a tumor or not. To evaluate the classifiers, the company has a large test set of images, which are labeled to indicate whether they contain a tumor or not. Let θ_1 denote the probability that C_1 makes a correct classification of a random image in the test set and let θ_2 denote the probability that C_2 makes a correct classification of a random image in the test set. Assume uniform priors for θ_1 and θ_2.

In a first evaluation, each of the classifiers were evaluated on 100 randomly drawn images from the test set. Out of 100 images, C_1 classified 95 correctly, and out of 100 different images, C_2 classfied 87 correctly.

(a) *Credits: 2p.* Compute the posterior probabilty that $\theta_1 > 0.9$ and the posterior probabilty that $\theta_2 > 0.9$.
Solution: See the code in `Exam732A91_201022_Sol.R`.

(b) *Credits: 2p.* Compute the posterior probabilty that $\theta_1 > \theta_2$ and make an interpretation about this probability.
Solution: See the code in `Exam732A91_201022_Sol.R`.

(c) *Credits: 4p.* Compute a 95% Highest Posterior Density interval for the difference $\theta_1 - \theta_2$. You may use `density` to first make a kernel density estimate of the posterior of $\theta_1 - \theta_2$.
Solution: See the code in `Exam732A91_201022_Sol.R`.

(d) *Credits: 2p.* Compute the posterior predictive probability that a new random image from the test set is classified correctly, for each classifier.
Solution: See the code in `Exam732A91_201022_Sol.R`.

2. BINOMIAL MODEL COMPARISON

Let $x_1, \ldots, x_n|\theta \overset{iid}{\sim} \text{Bin}(K, \theta)$ be n independent observations from the binomial distribution, where K is known. This problem should only be solved on **Paper**.

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

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

The likelihood can be defined by

$$p(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} \binom{K}{x_i} \theta^{x_i}(1-\theta)^{K-x_i}.$$

Using Bayes' theorem we get

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

where we can identify the form of a Beta distribution, so we can conclude that the posterior is $\theta|x_1, \ldots, x_n \sim \text{Beta}\left(\sum_{i=1}^n x_i + \alpha, nK - \sum_{i=1}^n x_i + \beta\right)$.

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

$$p(x_1, \ldots, x_n) = \frac{\prod_{i=1}^n \binom{K}{x_i} \cdot \Gamma(\alpha+\beta)\Gamma(S+\alpha)\Gamma(nK-S+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+nK)},$$

where $S = \sum_{i=1}^{n} x_i$.
Solution: The marginal likelihood can be computed using

$$\begin{aligned}
p(x_1, \ldots, x_n) &= \frac{p(x_1, \ldots, x_n|\theta)\, p(\theta)}{p(\theta|x_1, \ldots, x_n)} \\
&= \frac{\prod_{i=1}^{n} \begin{pmatrix} K \\ x_i \end{pmatrix} \theta^{x_i} (1-\theta)^{K-x_i} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}}{\frac{\Gamma(S+\alpha+nK-S+\beta)}{\Gamma(S+\alpha)\Gamma(nK-S+\beta)} \theta^{S+\alpha-1} (1-\theta)^{nK-S+\beta-1}} \\
&= \frac{\prod_{i=1}^{n} \begin{pmatrix} K \\ x_i \end{pmatrix} \cdot \Gamma(\alpha+\beta)\Gamma(S+\alpha)\Gamma(nK-S+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+nK)}.
\end{aligned}$$

(c) *Credits: 4p.* Assume that $n = 3$, $x_1 = 9$, $x_2 = 1$, and $x_3 = 9$. Do a Bayesian model comparison of two different Binomial models, M_1 and M_2, which have the same Beta$(1, 1)$-prior for θ, but where M_1 assumes $K = 10$ and M_2 assumes $K = 20$. Assume that both models are equally probable a priori, that is, that they have the same prior probability. What are your conclusions? [Hint: Any numerical computations can be carried out and reported in the submitted `R`-code. You may want to use the functions `gamma`, `beta` and `choose` in `R`.]
Solution: The Bayesian model comparison is carried out by computing the marginal likelihood for both models, and then using that the posterior model probabilities

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

See the numerical computations in `Exam732A91_201022_Sol.R`. The computed posterior model probabilities are $(0.12, 0.88)$ for the two models, so Model 2 is much more probable. This result is sensible, considering that for $K = 10$, any value for θ will be unlikely to generate both value 1 and 9.

3. Listeners

The dataset `listeners`, which is loaded by the code in `ExamData.R`, contains the number of listeners (in millions) to the different songs of a music artist on an online music streaming service. The data are modeled as independent from a two-component mixture of normals (MN(2)) model, so the number of listeners x_i to song i is distributed as

$$p(x_i) = \pi \cdot \phi(x_i|\mu_1, \sigma_1^2) + (1-\pi) \cdot \phi(x_i|\mu_2, \sigma_2^2),$$

where $\phi(x|\mu, \sigma^2)$ is the normal probability density function (pdf), and π, μ, σ_1^2, μ_2 and σ_2^2 are parameters of the MN(2) model.

(a) *Credits: 3p.* Use the supplied function `GibbsMixNormal` to do Gibbs sampling for the MN(2) model. This function automatically sets non-informative priors for all parameters. Set the seed using `set.seed(100)` and run the Gibbs sampling algorithm with `nIter=1000` draws. Plot trajectories over the iterations for the μ_1 and μ_2 parameters, based on the produced samples, which can be found in the output of the function (`muSample`). Based on the plots, suggest a sufficient number of initial burn-in iterations, after which the chain has reached its stationary distribution.
Solution: See the code in `Exam732A91_201022_Sol.R`.

(b) *Credits: 4p.* Compute the posterior mean of all the parameters of the MN(2) model, and denote these $\hat{\pi}$, $\hat{\mu}$, $\hat{\sigma}_1^2$, $\hat{\mu}_2$ and $\hat{\sigma}_2^2$. Use the computed means to plot the following in a single figure: 1) a histogram of the data, 2) MN(2) model density (`MixDensMean` in the output of the function), 3) The scaled density of the first component $\hat{\pi} \cdot \phi(x_i|\hat{\mu}_1, \hat{\sigma}_1^2)$, 4) The scaled density of the second component $(1-\hat{\pi}) \cdot \phi(x_i|\hat{\mu}_2, \hat{\sigma}_2^2)$.
Solution: See the code in `Exam732A91_201022_Sol.R`.

(c) *Credits: 3p.* The MN(2) model is used by the manager of the artist to analyze the artists' ability to produce hit songs (very popular songs). When seen as a generative model, datapoints that come from the mixture component of the MN(2) model with the largest μ value are seen as hit songs. Based on this interpretation, and the estimated model, what is the posterior predictive probability that a new song from the artist becomes a hit song? Given that a song is a hit song, what is the posterior predictive probability that the song gets more than 60 million listeners?
Solution: See the code in `Exam732A91_201022_Sol.R`.

4. Truncated zinc

The concentration of zinc in $\mu g/L$ in blood samples from 390 human subjects was measured and the data can be found in the file `zinc` which is loaded by the code in `ExamData.R`. The measurements are assumed to be independent and follow a truncated normal distribution with density

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

where $\phi(x)$ is the standard normal probability density function (pdf) and $\Phi\left(x\right)$ is the standard normal cumulative distribution function (cdf). $a = 400$ is the lower truncation point. The priors for μ and σ are both assumed to be normal with mean 1000 and standard deviation 100 and independent of each other.

(a) *Credits: 5p.* Use numerical optimization to obtain a normal approximation of the *joint* posterior distribution of μ and σ. Print the posterior mean and covariance matrix. [Hints: use the argument `method=c("L-BFGS-B")` in `optim`, and `control=list(fnscale=-1)`].
Solution: See the code in `Exam732A91_201022_Sol.R`.

(b) *Credits: 5p.* Simulate from the actual posterior using the Metropolis-Hastings algorithm. As proposal density, use independent uniform distributions according to

$$\mu_p|\mu^* \sim \text{Uniform}(\mu^* - 100, \mu^* + 100),$$
$$\sigma_p|\sigma^* \sim \text{Uniform}(\sigma^* - 100, \sigma^* + 100),$$

where and μ^* and σ^* are the previous samples. Use $\mu = 1000$, and $\sigma = 1000$ as starting values, 1000 iterations burn-in and thereafter draw 3000 samples from the posterior. Evaluate the convergence of the sampler. Is the sampling algorithm efficient? Plot the univariate posterior distribution of μ together with your approximation in (a), and do the same for σ.
Solution: See the code in `Exam732A91_201022_Sol.R`.

Good luck! Mattias and Per