LINKÖPING UNIVERSITY
Division of Statistics and Machine Learning
Department of Computer and Information Science
Mattias Villani and Per Sidén

Bayesian Learning

## Mathematical Exercises 3

Try to solve the problems before class. Don't worry if you fail, the important thing is trying. You should not hand in any solutions.
This part of the course is not obligatory and is not graded.

1. Fill in the blanks

   (a) Show that the full conditional posterior of $I_i$ on Lecture 7, Slide 20, is correct.
   **Solution**: (a) By Bayes' theorem, the full conditional posterior of $I_i$ is

   $$p(I_i = 1|\mathbf{x}, \cdot) \propto p(\mathbf{x}|I_i = 1, I_{-i}, \cdot)p(I_i = 1),$$

   where $\cdot$ is a shorthand for all other model parameters: $\pi, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2$. The symbol $I_{-i}$ denotes all allocation variables except for $I_i$. Now, since the data are iid we have $p(\mathbf{x}|I_i, I_{-i}, \cdot) = \prod_{j=1}^n p(x_j|I_j, \cdot)$. Note that only the factor $p(x_i|I_i, \cdot)$ in the product depends on $I_i$, and the other $n-1$ factors can therefore be moved into the proportionality constant. We therefore have

   $$p(I_i = 1|\mathbf{x}, I_{-i}, \cdot) \propto \phi(x_i|\mu_1, \sigma_1^2)\pi,$$

   where $\phi(x_i|\mu_1, \sigma_1^2)$ denotes the pdf of a $N(\mu_1, \sigma_1^2)$ variable evaluated in the point $x_i$. By exactly the same reasoning

   $$p(I_i = 2|\mathbf{x}, I_{-i}, \cdot) \propto \phi(x_i|\mu_2, \sigma_2^2)(1-\pi).$$

   All that remains to do is to normalize these two probabilities so that they sum to one, and we get the required result

   $$p(I_i = 1|\mathbf{x}, I_{-i}, \cdot) = \frac{\phi(x_i|\mu_1, \sigma_1^2)\pi}{\phi(x_i|\mu_1, \sigma_1^2)\pi + \phi(x_i|\mu_2, \sigma_2^2)(1-\pi)}.$$

   (b) On Lecture 7, Slide 24, I argue that one can simulate from the joint posterior distribution of the regression coefficients, $\beta$, the noise variance $\sigma^2$ and the regularization/shrinkage hyperparameter $\lambda$ using Gibbs sampling. In particular, I claim that the full conditional posterior of $\lambda$ is Gamma distributed. Derive this full conditional posterior of $\lambda$. [Hint: start by writing up the expression for the joint posterior of $\beta$, $\sigma^2$ and $\lambda$ using the Tattoo-version of Bayes Theorem. The full conditional posterior of $\lambda$ is proportional to this expression.]
   **Solution**: Use Bayes' theorem to get

   $$\begin{aligned} p\left(\lambda|\beta, \sigma^2, \mathbf{y}, \mathbf{X}\right) &\propto p\left(\mathbf{y}|\beta, \sigma^2, \lambda, \mathbf{X}\right) p\left(\lambda|\beta, \sigma^2\right) \\ &\propto p\left(\beta, \sigma^2, \lambda\right) \\ &\propto p\left(\beta|\sigma^2, \lambda\right) p\left(\lambda\right) \end{aligned}$$

since the likelihood $p\left(\mathbf{y}|\beta,\sigma^2,\lambda,\mathbf{X}\right)$ does not depend on $\lambda$, neither does the prior for $\sigma^2$. Use the Gamma prior for $\lambda$

$$\lambda \sim \mathsf{Gamma}\left(\frac{\eta_0}{2},\frac{\eta_0}{2\lambda_0}\right).$$

So,

$$\begin{aligned}
p\left(\lambda|\beta,\sigma^2,\mathbf{y},\mathbf{X}\right) &\propto p\left(\beta|\sigma^2,\lambda\right)p\left(\lambda\right)\\
&\propto \prod_{i=1}^{m}\frac{1}{\sqrt{2\pi\sigma^2/\lambda}}\exp\left(-\frac{\beta_i^2}{2\sigma^2/\lambda}\right)\cdot\lambda^{\eta_0/2-1}\exp\left(-\lambda\frac{\eta_0}{2\lambda_0}\right)\\
&\propto \lambda^{m/2}\exp\left(-\frac{\lambda}{2\sigma^2}\sum_{i=1}^{m}\beta_i^2\right)\cdot\lambda^{\eta_0/2-1}\exp\left(-\lambda\frac{\eta_0}{2\lambda_0}\right)\\
&\propto \lambda^{(m+\eta_0)/2-1}\exp\left(-\lambda\left(\frac{\sigma^{-2}\sum_{i=1}^{m}\beta_i^2+\eta_0/\lambda_0}{2}\right)\right).
\end{aligned}$$

Thus,

$$\lambda|\beta,\sigma^2,\mathbf{y},\mathbf{X} \sim \mathsf{Gamma}\left(\frac{m+\eta_0}{2},\frac{\sigma^{-2}\sum_{i=1}^{m}\beta_i^2+\eta_0/\lambda_0}{2}\right).$$

2. Frequentist meltdown or Bayesian breakdown?

   (a) Let $x_1,...,x_n \overset{iid}{\sim} Uniform(\theta-\frac{1}{2},\theta+\frac{1}{2})$. Let $\hat{\theta}=\bar{x}=\sum_{i=1}^{n}$ be an estimator of $\theta$. Derive an expression for the (repeated) sampling variance of $\hat{\theta}$.

   (b) Derive the posterior distribution for $\theta$ assuming a uniform prior distribution. [Hint: Here it absolutely crucial to think about the support for the data distribution. Once you have observed some data, some $\theta$ values are no longer possible. I strongly suggest that you plot some imaginary data on the real line and plot the data distribution in the same graph for some made up values of $\theta$. Just to make you think in the right direction.]

   (c) Assume that you have observed three data observations: $x_1=1.1, x_2=2.09, x_3=1.4$. What would a frequestist conclude about $\theta$? What would a Bayesian conclude? Discuss.

   **Solution**: See sketched solution in the end of this document.

3. Who doesn't want to be Normal?

   (a) Let $x_1,...,x_n \overset{iid}{\sim} \text{Bern}(\theta)$ and $\theta \sim \text{Beta}(\alpha,\beta)$ a priori. Find the posterior mode of $\theta$.

   (b) Approximate the posterior distribution of $\theta$ by a normal distribution.

   (c) Assume now that you have the data $n=6$ and $s=1$. Plot the true posterior distribution and the normal approximation in the same graph. Assume a uniform prior for $\theta$.

   (d) Redo the previous exercise, but this time with twice the data size: $n=12$ and $s=2$.

   **Solution**: See sketched solution in the end of this document.

4. Naive doctors

(a) Three diseases (A,B and C) have very common symptoms and are therefore hard to distinguish between for a doctor. A medical company has developed two different tests (T1 and T2) to discriminate between the three diseases. A training data from $n = 20$ patients was collected to learn a predictive model that can be used to classify a patient into disease A-C on the basis of the results from both T1 and T2. $n_A = 5$ of the patients had disease A, $n_B = 5$ of the patients had disease B and $n_C = 10$ of the patients had disease C. The table below gives the mean measurement in each patient group for both tests. The test measurements can be assumed to follow a normal distribution with variance $\sigma^2 = 1$ for all patient groups, and for both tests. Develop a Naive Bayes classifier based on this training data. You can assume uniform priors in any place you needs a prior. Make a prediction for a new patient with measurement 1.3 on T1 and 4.2 on T2.

| | $\bar{X}_1$ | $\bar{X}_2$ |
|---|---|---|
| Disease A | 1.2 | 2.1 |
| Disease B | 1.4 | 3.5 |
| Disease C | 0.7 | 4.7 |

**Solution**: Let us first focus on the predictive probability for disease A given the outcomes from the two tests $T_1$ and $T_2$

$$\begin{aligned}\Pr(A|T_1, T_2) &\propto \Pr(T_1, T_2|A)\Pr(A) \\ &= \Pr(T_1|A)\Pr(T_2|A)\Pr(A),\end{aligned}$$

where the latter equality is a result of the simplifying naive Bayes assumption that features (the tests) are independent conditional on the class (the disease). Similar expression hold for $\Pr(B|T_1, T_2)$ and $\Pr(C|T_1, T_2)$. Let us first estimate the class probabilities $P(A)$, $P(B)$ and $P(C)$. This estimation problem is a Multinomial-Dirichlet problem with $n = 20$ trials ending up in the categories: $n_A = 5, n_B = 5$ and $n_C = 10$. We can therefore obtain the posterior distribution for the vector $(\Pr(A), \Pr(B), \Pr(C))$ as a Dirichlet distribution. However, Naive Bayes typically uses just a point estimate of $(\Pr(A), \Pr(B), \Pr(C))$, and one immediate candidate for a point estimator is the posterior mean. The uniform prior distribution here is the Dirichlet$(1, 1, 1)$ distribution, and for the observed data, the posterior mean vector of $(\Pr(A), \Pr(B), \Pr(C))$ is

$$\left(\frac{n_A + 1}{n + 3}, \frac{n_B + 1}{n + 3}, \frac{n_C + 1}{n + 3}\right) = \left(\frac{6}{23}, \frac{6}{23}, \frac{11}{23}\right).$$

Now, the class conditional feature distributions are $T_1|A \sim N(\bar{x}_{1A}, 1+1/n_A)$ and similarly for disease $B$ and $C$. Note that this is the predictive distribution in the normal model with known variance equal to one, and a uniform prior for mean. [That is, the old result from Lecture 4: $\tilde{y}|\mathbf{y} \sim N(\bar{y}, \sigma^2(1 + 1/n_A))$ with $\sigma = 1$]. So, we have for disease A:

$$\begin{aligned}\Pr(A|T_1, T_2) &\propto \Pr(T_1|A)\Pr(T_2|A)\Pr(A) \\ &= \phi(1.3, \mu = 1.2, \sigma^2 = 1 + 1/5) \cdot \phi(4.2, \mu = 2.1, \sigma^2 = 1 + 1/5) \cdot \frac{6}{23},\end{aligned}$$

where $\phi(x, \mu, \sigma^2)$ is the pdf of a normal distribution with mean $\mu$ and variance $\sigma^2$ evaluated in the point $x$. The predictive distribution for the new patient is obtained by

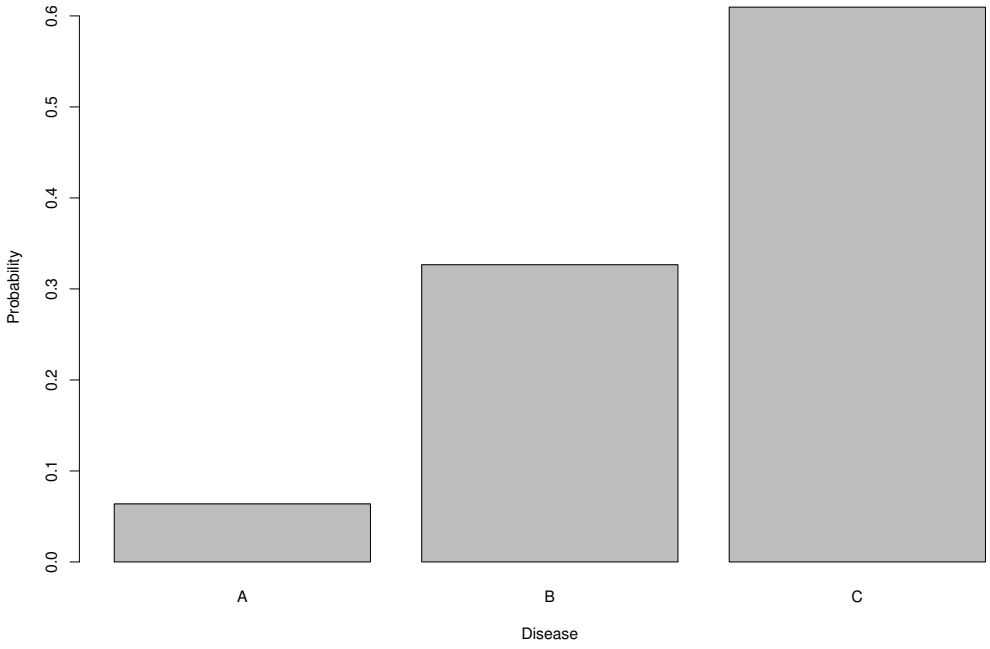



Figure 1: Estimated class conditional distributions for test $T_1$ and the value for $T_1$ for the predicted patient (as green circle).

repeating this type of calculation also for $\Pr(B|T_1, T_2)$ and $\Pr(C|T_1, T_2)$ and normalizing so that the three probabilities sum to one. The predictive distribution for the new patient is plotted in Figure 1. Figure 2 displays the estimated feature distribution for each disease and the observed $T_1$ (left) and $T_2$ (right) for the new patient. Note how $T_1$ gives somewhat more support for disease A and B, but that the observed value for $T_2$ is extremely improbable for a patient with disease A. This explains why disease B gets the highest predictive probability, and A gets a predictive probability close to zero.

Have fun!

- Mattias

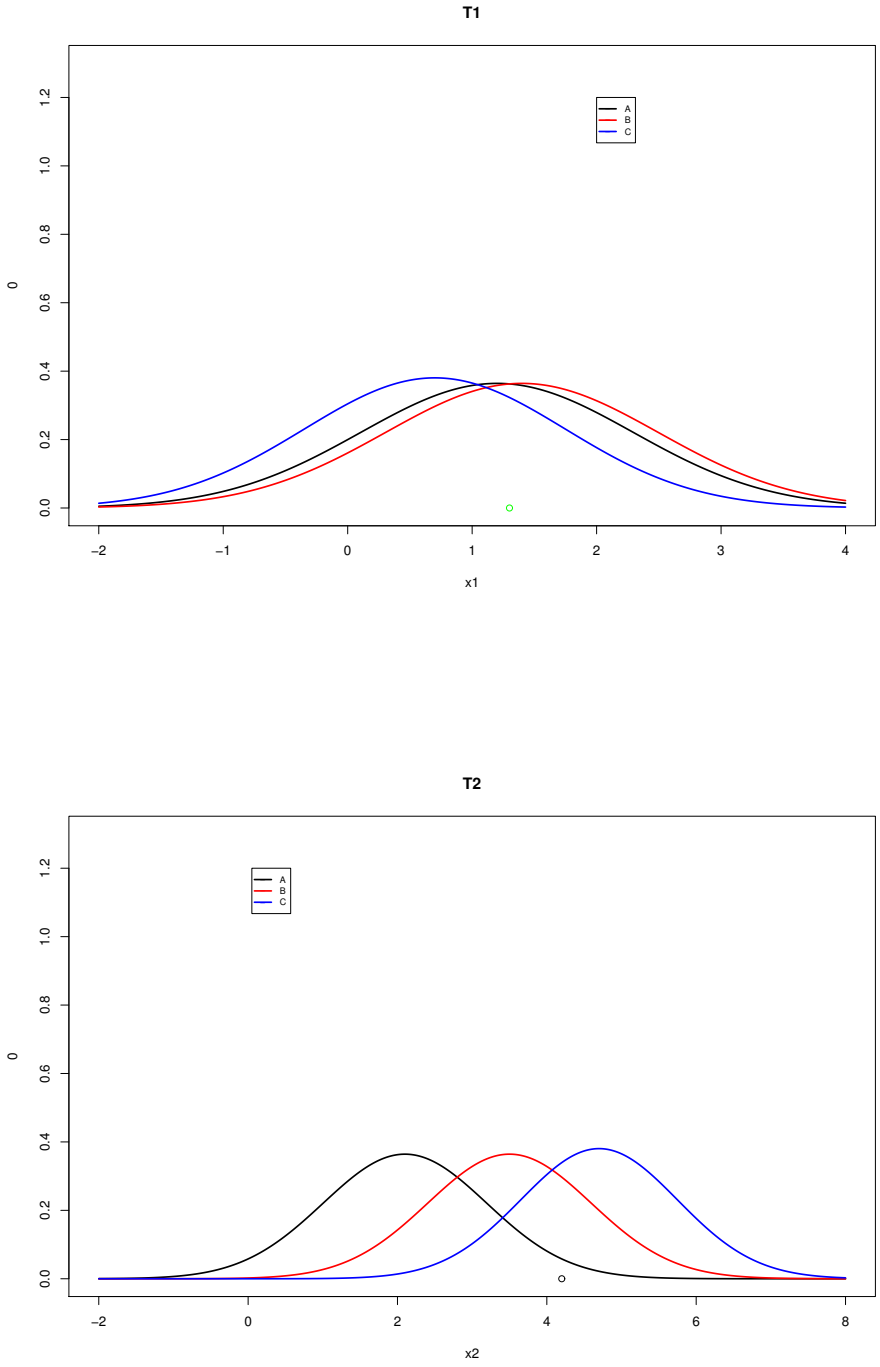
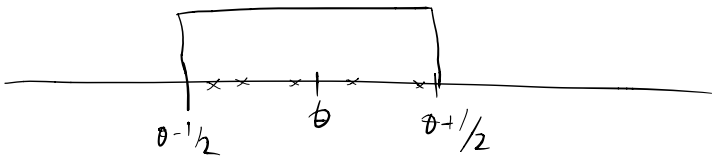



Figure 2: Estimated class conditional distributions for test $T_1$ (left) and $T_2$ (right). The value for $T_1$ and $T_2$ for the predicted patient are plotted as a green circles in respective graphs.

2a)




$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$ $\quad \sigma^2 = \text{Var}(X_i) \quad X_i \sim U(\theta - 1/2, \theta + 1/2)$

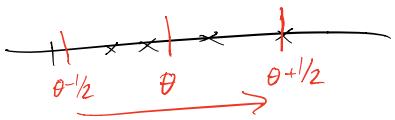
$X \sim U(0,1) \quad \text{Var}(X) = \frac{1}{12}$

So $\text{Var}(\bar{x}) = \frac{1}{12n}$

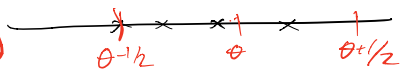
2b) $P(\theta | X_1, \ldots, X_n) \propto P(X_1 \ldots X_n | \theta) P(\theta)$

$= \prod_{i=1}^{n} P(X_i | \theta) P(\theta)$

$= \prod_{i=1}^{n} \mathbb{I}\left(\theta - \frac{1}{2} \le X_i \le \theta + 1/2\right) \cdot 1$




$\theta + 1/2 \ge X_{max}$
$\theta - 1/2 \le X_{min}$
$\Rightarrow \theta \in [X_{max} - 1/2, X_{min} + 1/2)$




$P(\theta | X_1 \ldots X_n) \propto 1 \quad \text{for} \quad \theta \in [X_{max} - 1/2, X_{min} + 1/2]$

$= 0 \text{ otherwise.}$

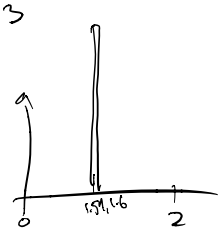
$\theta | X_1 \ldots X_n \sim U(X_{max} - 1/2, X_{min} + 1/2)$

2c) Frequentist : $\hat{\theta} = \bar{X} = 1.53$

$Var(\hat{\theta}) = \frac{1}{12n} = \frac{1}{12 \cdot 3} = 0.027777$

$SD(\hat{\theta}) = 0.1666$

Bayesian : $\theta | X_1 X_2 X_3 \sim U(1.5, 1.6)$



3 a) $\theta | x_1 \dots x_n \sim Beta(\alpha+s, \beta+f)$

$p(\theta|x) \propto \theta^{\alpha+s-1}(1-\theta)^{\beta+f-1} \Rightarrow \ln p(\theta|x) \propto (\alpha+s-1)\ln\theta + (\beta+f-1)\ln(1-\theta)$

$$\frac{\partial \ln p(\theta|x)}{\partial \theta} = \frac{\alpha+s-1}{\theta} + \frac{\beta+f-1}{1-\theta}(-1)$$

$$\frac{\partial \ln p(\theta|x)}{\partial \theta} = 0 \Rightarrow \frac{\alpha+s-1}{\theta} = \frac{\beta+f-1}{1-\theta}$$

$$\Rightarrow \hat{\theta} = \frac{\alpha+s-1}{\alpha+\beta+n-2}$$

3 b) $\theta|x_1 \dots x_n \overset{approx}{\sim} N(\hat{\theta}, -\mathcal{I}_{\theta,x}^{-1})$ $1-\hat{\theta} = \frac{\beta+f-1}{\alpha+\beta+n-2}$

$$\frac{\partial^2 \ln p(x|\theta)}{\partial \theta^2} = -\frac{\alpha+s-1}{\theta^2} + \frac{\beta+f-1}{(1-\theta)^2}(-1)$$

$$\left.\frac{\partial^2 \ln p(x|\theta)}{\partial \theta^2}\right|_{\theta=\hat{\theta}} = -\left(\frac{\alpha+s-1}{\left(\frac{\alpha+s-1}{\alpha+\beta+n-2}\right)^2} + \frac{\beta-f-1}{\left(\frac{\beta+f-1}{\alpha+\beta+n-2}\right)^2}\right)$$

$$= -(\alpha+\beta+n-2)^2\left(\frac{1}{\alpha+s-1} + \frac{1}{\beta+f-1}\right)$$

$$= -(\alpha+\beta+n-2)^2\left(\frac{\alpha+\beta+n-2}{(\alpha+s-1)(\beta+f-1)}\right)$$

$$= -\frac{(\alpha+\beta+n-2)^3}{(\alpha+s-1)(\beta+f-1)}$$

$$\theta | x_1 \dots x_n \overset{\text{approx}}{\sim} N\left( \hat{\theta} = \frac{\alpha+s-1}{\alpha+\beta+n-2}, \; \hat{J}^{-1}_{\hat{\theta},x} = \frac{(\alpha+s-1)(\beta+f-1)}{(\alpha+\beta+n-2)^3} \right)$$

Check: $\theta \sim \text{Beta}(\alpha+s, \beta+f)$

$$\text{Var}(\theta) = \frac{(\alpha+s)(\beta+f)}{(\alpha+s+\beta+f)^2(\alpha+s+\beta+f+1)}$$

$$= \frac{(\alpha+s)(\beta+f)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}$$

3 c) Se R-code

3 d) ——"——