TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

Surname and initials:

Student number:

Examination cover sheet

(to be completed by the examiner)

Course name:Bayesian Machine Learning and Information Processing Course code: 5SSD0

Date: 15-April 2021

Start time: 18:00 End time: 21:00

Number of pages: 10

Number of questions: 6

Maximum number of points/distribution of points over questions: each question 5 points. Total 30 points

Method of determining final grade: grade = (sum of # points)/3 + correction. Correction based on essay as discussed in email/Canvas.

Answering style: formulation, order, argumentation, multiple choice: multiple choice

Exam inspection:

Other remarks: good luck!!

INSTRUCTIONS FOR STUDENTS AND INVIGILATORS (to be indicated by examinor)

Write in black or blue. Pencil only allowed for drawings.

Permitted examination aids (to be supplied by students):

☐ Computer

☐ Calculator Graphic

☐ Calculator

☐ Lecture notes/book

☐ One A4 sheet of annotations

☐ Dictionaries. If yes, please specify:

Important:

- examinees are only permitted to visit the toilets under supervision
- it is not permitted to leave the examination room within 15 minutes of the start and within the final 15 minutes of the examination, unless stated otherwise
- examination scripts (fully completed examination paper, stating name, student number, etc.) must always be handed in
- the house rules must be observed during the examination
- the instructions of subject experts and invigilators must be followed
- keep your work place as clean as possible: put pencil case and breadbox away, limit snacks and drinks
- examinees are not permitted to share examination aids or lend them to each other

During written examinations, the following actions will in any case be deemed to constitute fraud or attempted fraud:

- using another person's proof of identity/campus card (student identity card)
- having a mobile telephone or any other type of media-carrying device on your desk or in your clothes
- using, or attempting to use, unauthorized resources and aids, such as the internet, a mobile telephone, smartwatch, smart glasses etc.
- having any paper at hand other than that provided by TU/e, unless stated otherwise
- copying (in any form)
- visiting the toilet (or going outside) without permission or supervision

The final grade will be announced no later than fifteen working days after this examination took place. Final grades of first-year bachelor study components in Q4 will be announced within 5 working days. Final test grades of bachelor study components in the interim period will be announced no later than 5 working days before the 1st of September.

Associated with the Central Examination Regulations

Exercises

1	2	3	4	5	6

Surname, First name

Bayesian machine learning and information processing (5SSD0)

5SSD0 Bayesian machine learning and information processing Q2 20-21 Resit - Handwritten exam

1	1	1	1	1	1	1
2	2	2	2	2	2	2
3	3	3	3	3	3	3
4	4	4	4	4	4	4
5	5	5	5	5	5	5
6	6	6	6	6	6	6
7	7	7	7	7	7	7
8	8	8	8	8	8	8
9	9	9	9	9	9	9
0	0	0	0	0	0	0

Probabilities of drawing objects from a box

Box 1 contains 4 apples and 8 oranges. Box 2 contains 10 apples and 2 oranges. Boxes are chosen with equal probability. You make one draw.

1p **1a** What is the probability of choosing an apple?

(a) $\frac{5}{12}$
(b) $\frac{6}{12}$
(c) $\frac{4}{12}$
(d) $\frac{7}{12}$

2p **1b** If an apple is chosen, what is the probability that it came from box 1?

(a) $\frac{4}{7}$
(b) $\frac{2}{7}$
(c) $\frac{5}{14}$
(d) $\frac{1}{3}$

Instead of one draw, we now take two draws without replacement. Again, we can draw from either box at each draw, and boxes are choses with equal probabilities.

2p **1c** If you know that your second draw will be an orange from box 2, what is now the probability of drawing an apple at the first draw?

(a) $7/12$
(b) $41/66$
(c) $7/11$
(d) The correct answer is not shown.

Model comparison

A model m_1 is described by a single parameter θ, with $0 \leq \theta \leq 1$. The system can produce data $x \in \{0, 1\}$.

The sampling distribution $p(x|\theta, m_1)$ and prior $p(\theta|m_1)$ are given by

$$p(\theta|m_1) = 6\theta(1-\theta)$$
$$p(x|\theta, m_1) = \theta^x(1-\theta)^{1-x}$$

1p **2a** Determine the evidence $p(x = 1|m_1)$

(a) $\frac{2}{3}$
(b) $\frac{1}{6}$
(c) $\frac{1}{3}$
(d) $\frac{1}{2}$

2p **2b** Determine the posterior probability $p(\theta|x = 1, m_1)$.

(a) $6\frac{\theta^2}{1-\theta}$
(b) $12\theta^2(1-\theta)$
(c) $6\theta^2(1-\theta)$
(d) $12\frac{\theta^2}{1-\theta}$

Consider a second model m_2 with the following sampling distribution and prior on $0 \leq \theta \leq 1$:

$$p(\theta|m_2) = 1$$
$$p(x|\theta, m_2) = (1-\theta)^x \theta^{1-x}$$

The model priors are given by $p(m_1) = 2/3$ and $p(m_2) = 1/3$.

2p **2c** Determine the ratio of posterior model probabilities $\frac{p(m_1|x=1)}{p(m_2|x=1)}$.

(a) $\frac{1}{3}$
(b) $\frac{1}{2}$
(c) $\frac{2}{3}$
(d) 2

Recursive Bayesian Filtering
We observe a process

$x_t = \theta + \epsilon_t$ with $\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$.

We are interested in recursively updating estimates for θ from observations $x_1, x_2, \ldots$. The estimate for θ after k observations $D_k = \{x_1, x_2, \ldots, x_k\}$ is written as

$p(\theta|D_k) = \mathcal{N}(\theta|\mu_k, \sigma_k^2)$.

The prior for θ (after zero observations) is given by $p(\theta) = p(\theta|D_0) = \mathcal{N}(\theta|\mu_0, \sigma_0^2)$.

1p **3a** Which is a correct expression for the likelihood $p(x_k|\theta)$?

(a) $p(x_k|\theta) = \mathcal{N}(x_t|\mu_k, \sigma_\epsilon^2 + \sigma_\theta^2)$
(b) $p(x_k|\theta) = \mathcal{N}(x_t|\theta, \sigma_\epsilon^2)$
(c) $p(x_k|\theta) = \mathcal{N}(x_t|0, \sigma_\epsilon^2 + \sigma_\theta^2)$
(d) $p(x_k|\theta) = \mathcal{N}(x_t|\theta, \sigma_\theta^2)$

3p **3b** Next, you are asked to work out the recursive update formula for posterior $p(\theta|D_k) = \mathcal{N}(\theta|\mu_k, \sigma_k^2)$. Basically this is a derivation of a very simple Kalman filter. In this derivation you may want to use the formula for Gaussian multiplication:

$\mathcal{N}(x|\mu_a, \sigma_a^2)\mathcal{N}(x|\mu_b, \sigma_b^2) \propto \mathcal{N}(x|\mu_c, \sigma_c^2)$ with $\frac{1}{\sigma_c^2} = \frac{1}{\sigma_a^2} + \frac{1}{\sigma_b^2}$ and $\frac{1}{\sigma_c^2}\mu_c = \frac{1}{\sigma_a^2}\mu_a + \frac{1}{\sigma_b^2}\mu_b$.

I will start the derivation here:

$$\begin{aligned} p(\theta|D_k) &= p(\theta|x_k, D_{k-1}) \\ &\propto p(\theta, x_k|D_{k-1}) \\ &= p(x_k|\theta)p(\theta|D_{k-1}) \\ &= \dots \end{aligned}$$

Now you complete this and derive expressions for μ_k and σ_k^2.

(a)
$$\begin{cases} K_k &= \frac{\sigma_\epsilon^2}{\sigma_{k-1}^2+\sigma_\epsilon^2} \\ \mu_k &= \mu_{k-1} + K_k \cdot (x_k - \mu_{k-1}) \\ \sigma_k^2 &= (1 - K_k) \cdot \sigma_{k-1}^2 + \sigma_\epsilon^2 \end{cases}$$

(b)
$$\begin{cases} K_k &= \frac{\sigma_\epsilon^2}{\sigma_{k-1}^2+\sigma_\epsilon^2} \\ \mu_k &= \mu_{k-1} + K_k \cdot (x_k - \mu_{k-1}) \\ \sigma_k^2 &= (1 - K_k) \cdot \sigma_{k-1}^2 \end{cases}$$

(c)
$$\begin{cases} K_k &= \frac{\sigma_{k-1}^2}{\sigma_{k-1}^2+\sigma_\epsilon^2} \\ \mu_k &= \mu_{k-1} + K_k \cdot (x_k - \mu_{k-1}) \\ \sigma_k^2 &= \sigma_{k-1}^2 + K_k \cdot (\sigma_\epsilon^2 - \sigma_{k-1}^2) \end{cases}$$

(d)
$$\begin{cases} K_k &= \frac{\sigma_{k-1}^2}{\sigma_{k-1}^2+\sigma_\epsilon^2} \\ \mu_k &= \frac{1}{2}\mu_{k-1} + \frac{1}{2}K_k \cdot (x_k - \mu_{k-1}) \\ \sigma_k^2 &= (1 - K_k) \cdot \sigma_{k-1}^2 + \sigma_\epsilon^2 \end{cases}$$

(e)
$$\begin{cases} K_k &= \frac{\sigma_{k-1}^2}{\sigma_{k-1}^2+\sigma_\epsilon^2} \\ \mu_k &= \mu_{k-1} + K_k \cdot (x_k - \mu_{k-1}) \\ \sigma_k^2 &= (1 - K_k) \cdot \sigma_{k-1}^2 \end{cases}$$

1p **3c** What happens for $k \to \infty$ (# observations goes to infinity).

(a) $\sigma_k^2 \to \sigma_{k-1}^2$

(b) $\mu_k \approx \mu_{k-1}$ (stationarity)

(c) $\sigma_k^2 \to \sigma_\epsilon^2$

(d) $K_k \to 1$

Comprehension

1p **4a** Which of the following statements is **most** consistent with Friston's Free Energy Principle?

(a) Actions aim to minimize the *free energy of future states* of the world.

(b) Actions aim to minimize the *complexity of future states* of the world.

(c) Intelligent decision making requires minimization of a *functional of beliefs* about future states of the world.

(d) Intelligent decision making requires minimization of a *cost function of future states* of the world.

1p **4b** Which of the following statements is most accurate about Bayesian vs Maximum Likelihood-based estimation?

(a) The ML estimate tends to become a better approximation to the Bayesian estimate as the data size grows, since the prior distribution in Bayesian estimation tends to become wider with more data.

(b) The ML estimate tends to become a worse approximation to the Bayesian estimate as the data size grows, since both the likelihood function and prior distribution tend to become wider with more data.

(c) The ML estimate tends to become a better approximation to the Bayesian estimate as the data size grows, since the likelihood function tends to become wider with more data while the prior distribution in Bayesian estimation does not depend on the data set size.

(d) The ML estimate tends to become a better approximation to the Bayesian estimate as the data size grows, since the likelihood function tends to become narrower with more data while the prior distribution in Bayesian estimation does not depend on the data set size.

1p **4c** A Bayesian sticks a microphone in the air and records a signal $x = (x_1, x_2, \ldots, x_T)$. She is interested in retrieving a speech signal $s = (s_1, s_2, \ldots, s_T)$ from the recording. How might she proceed?

(a) She describes the recorded signal by a model $p(x, s, z) = p(s|x, z)p(x, z)$ where z is a set of latent states and model parameters. Then, she will proceed to infer her beliefs about s by computing $p(s|x) \propto \int p(s|x, z)\mathrm{d}z$.

(b) She describes the recorded signal by a model $p(x, s, z) = p(s|x, z)p(x, z)$ where z is a set of latent states and model parameters. Then, she will proceed to infer her beliefs about s by computing $p(s|x) \propto \int p(s|x, z)p(x, z)\mathrm{d}z$.

(c) She describes the recorded signal by a joint model $p(x, s, z) = p(x|s, z)p(s, z)$ where z is a set of latent states and model parameters. Then, she will proceed to infer her beliefs about s by computing $p(s|x, z) = \frac{p(x|s,z)p(s,z)}{p(x,z)}$.

(d) She describes the recorded signal by a joint model $p(x, s, z) = p(x|s, z)p(s, z)$ where z is a set of latent states and model parameters. Then, she will proceed to infer her beliefs about s by computing $p(s|x) \propto \int p(x|s, z)p(s, z)\mathrm{d}z$.

1p **4d** Consider a data set $\{x_n | n = 1, 2, \ldots, N\}$ with $x_n \in \mathbb{R}^M$ and a set of latent one-hot coded variables $z_n = (z_{n1}, z_{n2}, \ldots, z_{nK})$, i.e., $z_{nk} \in \{0, 1\}$ and $\sum_{k=1}^K z_{nk} = 1$. Which of the following is a correct specification for a Gaussian Mixture Model?

(a) $p(x_n, z_n) = \prod_{k=1}^K (\pi_k \cdot \mathcal{N}(x_n | \mu_k, \Sigma_k))^{z_n}$

(b) $p(x_n, z_n) = \prod_{k=1}^K \pi_k \cdot \mathcal{N}(x_n | \mu_k, \Sigma_k)^{z_{nk}}$

(c) $p(x_n, z_n) = \prod_{k=1}^K \pi_k \cdot \mathcal{N}(x_n | \mu_k, \Sigma_k)$

(d) $p(x_n, z_n) = \prod_{k=1}^K (\pi_k \cdot \mathcal{N}(x_n | \mu_k, \Sigma_k))^{z_{nk}}$

1p **4e** Consider a generative model $p(x, z)$ where x has been observed and z are latent states. We introduce a posterior distribution $q(z)$ for the latent states and define a Free Energy functional $F[q] = \int q(z) \log \frac{q(z)}{p(x,z)} \mathrm{d}z$. Which of the following statements is true?

(a) $F[q] = -\log p(x)$ if $q(z) = 0$.

(b) $F[q] \leq -\log p(x)$ for any choice of $q(z)$.

(c) $F[q] \geq -\log p(x)$ for any choice of $q(z)$.

(d) $F[q] = -\log p(x)$ if $q(z) = p(z)$.

Classifier Orange-ness

You have a machine that measures property x, the "orangeness" of liquids. You wish to discriminate between C_1 = 'Fanta' and C_2 = 'Orangina'. It is known that

$$p(x|C_1) = \begin{cases} 6x(1-x) & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

$$p(x|C_2) = \begin{cases} 2x & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

The probability that x falls outside the interval $[0.0, 1.0]$ is zero. The prior class probabilities $p(C_1) = 0.6$ and $p(C_2) = 0.4$ are also known from experience.

1p **5a** We want to develop a Bayesian classifier. The discrimination boundary on the interval $x \in [0.0, 1.0]$ is given by

(a) $\frac{1}{2} = \frac{p(C_1|x)}{p(C_2|x)} = \frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)} = \frac{6x(1-x)\cdot 0.6}{2x \cdot 0.4} \quad \Rightarrow x = 8/9$

(b) $1 = \frac{p(x|C_1)}{p(x|C_2)} = \frac{6x(1-x)}{2x} \quad \Rightarrow x = 2/3$

(c) $1 = \frac{p(C_1|x)}{p(C_2|x)} = \frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)} = \frac{6x(1-x)\cdot 0.6}{2x \cdot 0.4} \quad \Rightarrow x = 7/9$

2p **5b** Compute $p(C_1|x = 0.5)$.

(a) $p(C_1|x = 0.5) = \frac{p(x=0.5|C_1)p(C_1)}{p(x=0.5|C_2)p(C_2)} = \frac{\frac{6}{4} \cdot \frac{6}{10}}{1 \cdot \frac{4}{10}}$

(b) $p(C_1|x = 0.5) = \frac{p(x=0.5|C_1)p(C_1)}{p(x=0.5)} = \frac{\frac{6}{4} \cdot \frac{6}{10}}{\frac{6}{4} \cdot \frac{6}{10} + 1 \cdot \frac{4}{10}}$

(c) $p(C_1|x = 0.5) = p(x = 0.5|C_1)p(C_1) = \frac{6}{4} \cdot \frac{6}{10}$

2p **5c** Now assume that the prior probabilities $p(C_1)$ and $p(C_2)$ are not known and you have a data set of correctly labeled samples $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$. How would a Bayesian approach proceed to classify a new sample $x_\bullet$?

(a) Assume parameterized prior class probabilities, e.g., $p(C_1|\theta) = \theta$ and $p(C_2|\theta) = 1 - \theta$, with a Gaussian prior on θ. Then absorb the data in the model by Bayes rule to get $p(\theta|D)$ and classify $x_\bullet$ by $p(C_1|x_\bullet, D) \propto \int p(x_\bullet|C_1, \theta)p(\theta|D)\mathrm{d}\theta$.

(b) Assume $p(C_1) = p(C_2) = 0.5$ and use Bayes rule to compute $p(C_1|x_\bullet)$.

(c) Assume parameterized prior class probabilities, e.g., $p(C_1|\theta) = \theta$ and $p(C_2|\theta) = 1 - \theta$, with a Beta prior on θ. Then absorb the data in the model by Bayes rule to get $p(\theta|D)$ and classify $x_\bullet$ by $p(C_1|x_\bullet, D) \propto \int p(x_\bullet|C_1)p(C_1|\theta)p(\theta|D)\mathrm{d}\theta$.

Probabilistic Programming

1p **6a** Suppose someone gives you the following data set:

```
X = [1  1  0  0  0  0  2  0  1  0  1  1  2  1  1]
```

What would be an appropriate choice for a likelihood function?

(a) @RV X ~ Dirichlet(.)

(b) @RV X ~ Bernoulli(.)

(c) @RV X ~ Gaussian(.)

(d) @RV X ~ Categorical(.)

2p **6b** What is missing from the following mixture model specification?

```
graph = FactorGraph()

@RV ϕ ~ Dirichlet(a0)

for k = 1:num_components

    @RV Λ[k] ~ Wishart(V0[:,:,k], n0)
    @RV μ[k] ~ GaussianMeanPrecision(m0[:,k], Λ[k])

    push!(θ, μ[k], Λ[k])
end

for i = 1:num_samples

    [redacted]
    @RV X[i] ~ GaussianMixture(z[i], θ...)

    placeholder(X[i], :X, dims=(num_features,), index=i)
end
```

(a) @RV z[i] ~ Beta(ϕ)
(b) @RV z[i] ~ Categorical(ϕ)
(c) @RV z[i] ~ GaussianMeanVariance(ϕ)
(d) @RV z[i] ~ Bernoulli(ϕ)

2p **6c** What is wrong with the following dynamic model specification?

```
graph = FactorGraph()

@RV σ_x ~ Gamma(placeholder(:a_x), placeholder(:b_x))

@RV x_kmin1 ~ GaussianMeanVariance(placeholder(:m), placeholder(:v))

@RV x_k ~ GaussianMeanPrecision(x_kmin1, σ_x)

@RV y_k ~ GaussianMeanVariance(x_k, 0.1)
placeholder(y_k, :y_k)
```

(a) There is nothing wrong with this model specification.
(b) x_k should be distributed according to a GaussianMeanVariance distribution.
(c) σ_x should be distributed according to a Wishart distribution.
(d) The placeholder for x_kmin1 is missing.