Exercises

1	2	3	4	5	6

Surname, First name

____

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

5SSD0 Bayesian machine learning and information processing
5SSD0 Final Exam Q2

a ☒ c d e f → b
a b ☒ d e f → c
● b c ☒ e f → a

Fill in your answer(s) to the multiple-choice questions as shown above (circles = one correct answer).

Particular Ans on paper exam instructions

- Write in a black or blue pen.

Dear student,

You're about to take an exam. Write down your name and your student ID at the appropriate places above. Make sure that you enter your student ID by fully coloring the appropriate boxes. On the examination attendance card, you fill in a document number. You can find the correct number on the top of the first page of your exam (10 numbers).

Please read the following information carefully:

Date exam: 03-02-2022
Start time: 13.30
End time: 16.30 (+30 minutes for time extension students)

Answering style: multiple choice
Number of questions: 6
Maximum number of points: 27
Method of determining the final grade:

The maximal score for the written exam is $27/3 = 9.0$ points. The maximal score for the programming assignment is 1.0 point. These scores are added together for a score x =written + assignment. If (and only if) $5.0 \leq x \leq 5.4$, then I will take a look at your piazza activity. If you have been active and helpful on piazza (subjective measure, not up for discussion afterward), then I will give you a bonus and your score will become $x = 5.5$. The final score will be based on rounding towards whole integers, i.e., final score $= round(x)$. In other words, if $x >= 5.5$, you will pass the class.

Permitted examination aids

- Scrap paper (fully blank)
- Formula sheet
  - The **Gaussian distribution** is given by

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{(2\pi)^{M/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

Important:

- You are only permitted to visit the toilets under supervision
- 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
- You are not permitted to share examination aids or lend them to each other
- Do not communicate with any other person by any means

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

You can start the exam now, good luck!

1. Conceptual Understanding

1p **1a** Below are four statements about the Bayesian approach to machine learning. Which of the below statements is false?

ⓐ The Bayesian approach does not require splitting the data set into a training and test set. All data can be used for training.

ⓑ The Bayesian approach is a fundamentally sound approach to optimal processing of incomplete information sources such as a data set, since it is based on probability theory.

ⓒ The Bayesian approach to machine learning requires upfront to state all model assumptions explicitly in the model specification. Alternative approaches may hide these assumptions in various ways through cost functions, learning rates, etc.

ⓓ The Bayesian approach to machine learning is a fast alternative to the more fundamental maximum likelihood method.

1p **1b** Which of the following statements is NOT a property of the Variational Bayesian (VB) approach to machine learning?.

ⓐ The VB approach transfers Bayesian ML to an optimization problem.

ⓑ VB finds posterior distributions by maximizing Bayesian model evidence.

ⓒ Minimization of Variational Free Energy leads both to (possibly approximate) results for (1) the posterior distribution over the latent variables, and (2) Bayesian model evidence.

ⓓ Global optimization of Variational Free Energy leads to the realization of Bayes rule.

1p **1c** Consider two model specifications $p(x,\theta,m_k) = p(x|\theta,m_k)p(\theta|m_k)p(m_k)$ for $k = \{1,2\}$. After training both models on the same data set $D = \{x_1, x_2, \ldots, x_N\}$, you want to use the Bayes Factor (BF) to select the best model for this data set. The BF can be expressed as follows:

ⓐ $B_{12} = \frac{p(D|m_1)}{p(D|m_2)} = \frac{p(m_1|D)}{p(m_2|D)} \cdot \frac{p(m_1)}{p(m_2)}$

ⓑ $B_{12} = \frac{p(D|m_1)}{p(D|m_2)} = \frac{p(m_1|D)}{p(m_2|D)} \cdot \frac{p(m_2)}{p(m_1)}$

ⓒ $B_{12} = \frac{p(m_1|D)}{p(m_2|D)} = \frac{p(D|m_1)}{p(D|m_2)} \cdot \frac{p(m_2)}{p(m_1)}$

ⓓ $B_{12} = \frac{p(m_1|D)}{p(m_2|D)} = \frac{p(D|m_1)}{p(D|m_2)} \cdot \frac{p(m_1)}{p(m_2)}$

1p **1d** In the Free Energy Principle approach to designing an intelligent agent, which of the following statements describes most accurately how to equip the agent with goal-driven behavior?

ⓐ Specify a cost function of future states and choose actions to minimize future costs.

ⓑ Extend the generative model with target priors for future observations. Then choose actions that minimize Free Energy in the extended model.

ⓒ Specify a cost function of actions and choose actions that minimize the cost function.

ⓓ Extend the generative model with a posterior distribution for actions and choose the action that maximizes this posterior distribution.

Generative Classification

Consider a two-class classification problem with data set $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, where $x_n \in \mathbb{R}^{2\times 1}$ are observed features (note that x_n is a two-dimensional column vector) and $y_n \in \{(1,0), (0,1)\}$ is a one-hot coded class identifier. We define a data generating distribution by

$$p(x_n|y_{n1} = 1)p(y_{n1} = 1) = \mathcal{N}(x_n|\mu_1, \Sigma_1) \cdot \pi_1$$
$$p(x_n|y_{n2} = 1)p(y_{n2} = 1) = \mathcal{N}(x_n|\mu_2, \Sigma_2) \cdot \pi_2$$

where $\theta = \{\pi_1, \pi_2, \mu_1, \mu_2, \Sigma_1, \Sigma_2\}$ are the model parameters.

1p **2a** What is the expression for the "generative model" $p(x_n, y_n)$?

ⓐ $p(x_n, y_n) = \prod_{k=1}^{2} \pi_k \cdot \mathcal{N}(x_n|\mu_k, \Sigma_k)^{y_{nk}}$

ⓑ $p(x_n, y_n) = \prod_{k=1}^{2} \left(\pi_k \cdot \mathcal{N}(x_n|\mu_k, \Sigma_k)\right)^{y_{nk}}$

ⓒ $p(x_n, y_n) = \prod_{k=1}^{2} \pi_k^{y_{nk}} \cdot \mathcal{N}(x_n|\mu_k, \Sigma_k)$

ⓓ $p(x_n, y_n) = \sum_{k=1}^{2} y_{nk} \log\left(\pi_k \cdot \mathcal{N}(x_n|\mu_k, \Sigma_k)\right)$

1p **2b** The posterior class distribution $p(y_{n1} = 1|x_n)$ for a given input x_n is given by:

ⓐ $p(y_{n1} = 1|x_n) = \frac{\mathcal{N}(x_n|\mu_1,\Sigma_1)}{\mathcal{N}(x_n|\mu_1,\Sigma_1)+\mathcal{N}(x_n|\mu_2,\Sigma_2)}$

ⓑ $p(y_{n1} = 1|x_n) = \frac{\pi_1}{\pi_1+\pi_2}$

ⓒ $p(y_{n1} = 1|x_n) = \frac{\pi_2 \cdot \mathcal{N}(x_n|\mu_2,\Sigma_2)}{\pi_1 \cdot \mathcal{N}(x_n|\mu_1,\Sigma_1)+\pi_2 \cdot \mathcal{N}(x_n|\mu_2,\Sigma_2)}$

ⓓ $p(y_{n1} = 1|x_n) = \frac{\pi_1 \cdot \mathcal{N}(x_n|\mu_1,\Sigma_1)}{\pi_1 \cdot \mathcal{N}(x_n|\mu_1,\Sigma_1)+\pi_2 \cdot \mathcal{N}(x_n|\mu_2,\Sigma_2)}$

1p **2c** The log-likelihood $\log p(D|\theta)$ can be worked out to

ⓐ $\log p(D|\theta) = \sum_k y_{nk} \log \mathcal{N}(x_n|\mu_k, \Sigma_k) + \sum_k y_{nk} \log \pi_k$

ⓑ $\log p(D|\theta) = \sum_n \sum_k y_{nk} \log \mathcal{N}(x_n|\mu_k, \Sigma_k) + \sum_n \sum_k \log \pi_k$

ⓒ $\log p(D|\theta) = \sum_n \sum_k y_{nk} \log \mathcal{N}(x_n|\mu_k, \Sigma_k) + \sum_n \sum_k y_{nk} \log \pi_k$

ⓓ $\log p(D|\theta) = \sum_k y_{nk} \log\left(\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)\right)$

1p **2d** Let $\hat{\mu}_2$ be the maximum likelihood estimate for μ_2. The maximum likelihood estimate $\hat{\Sigma}_2$ for the variance parameter Σ_2 is given by

ⓐ $\hat{\Sigma}_2 = \frac{1}{N} \sum_n (x_n - \hat{\mu}_2)(x_n - \hat{\mu}_2)^T$

ⓑ $\hat{\Sigma}_2 = \frac{1}{N} \sum_n y_{n2}(x_n - \hat{\mu}_2)(x_n - \hat{\mu}_2)^T$

ⓒ $\hat{\Sigma}_2 = \frac{1}{N} \sum_n y_{n2}(x_n - \hat{\mu}_2)^T(x_n - \hat{\mu}_2)$

ⓓ $\hat{\Sigma}_2 = \frac{1}{N} \sum_n y_{n2}(x_n - \hat{\mu}_2)^2$

1p **2e** Aside from degenerative cases, the discrimination boundary between the two classes will be given by a

ⓐ straight line

ⓑ parabola

ⓒ square

ⓓ triangle

Coin Toss Prediction

Consider a biased coin with outcomes

$$x_n = \begin{cases} 0 & \text{(heads)} \\ 1 & \text{(tails)} \end{cases}$$

We assume that the data generating process is governed by a Bernoulli distribution,

$$p(x_n|\mu) = \mu^{x_n}(1-\mu)^{(1-x_n)}$$

and we assume a Beta distribution for the prior on μ:

$$p(\mu) = \mathrm{Beta}(\mu|\alpha = 3, \beta = 2)\,.$$

Note that the Beta distribution is given by $\mathrm{Beta}(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$, where $\Gamma(\cdot)$ is the gamma function. The **mean** of the Beta distribution is given by $\mathrm{E}\left[x\right] = \frac{\alpha}{\alpha+\beta}$.

We throw the coin 7 times and observe outcomes $D = \{0,1,0,0,1,0,0\}$.

1p **3a** Which of the following interpretations of the choice $\alpha = 3$, $\beta = 2$ is most valid?

ⓐ We assume 5 "pseudo" coin tosses with outcomes 2 tails and 3 heads.

ⓑ We assume 3 "pseudo" coin tosses with outcomes 2 tails and 1 heads.

ⓒ We assume that the probability of throwing tails is $2/3$ times the probability of throwing heads.

ⓓ We assume 5 "pseudo" coin tosses with outcomes 3 tails and 2 heads.

1p **3b** Work out the likelihood function $p(D|\mu)$ for μ.

ⓐ $p(D|\mu) = \binom{5}{2} \cdot \mu^5(1-\mu)^2$

ⓑ $p(D|\mu) = \mu^5(1-\mu)^2$

ⓒ $p(D|\mu) = \mu^2(1-\mu)^5$

ⓓ $p(D|\mu) = \mu^1(1-\mu)^4$

1p **3c** Compute the posterior distribution $p(\mu|D)$.

ⓐ $p(\mu|D) = \mathrm{Beta}(\mu|4,6)$
ⓑ $p(\mu|D) = \mu^4(1-\mu)^6$
ⓒ $p(\mu|D) = \mu^5(1-\mu)^7$
ⓓ $p(\mu|D) = \mathrm{Beta}(\mu|5,7)$

2p **3d** Now compute the probability for throwing tails after the data set has been absorbed in the model.

ⓐ $p(x_{n+1} = 1|D) = 4/11$
ⓑ $p(x_{n+1} = 1|D) = \frac{3}{5}$
ⓒ $p(x_{n+1} = 1|D) = \frac{1}{2}$
ⓓ $p(x_{n+1} = 1|D) = 5/12$

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

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

1p **4a** Work out the probability $p(x=1|m_1)$.

ⓐ $1/4$
ⓑ $1/2$
ⓒ $\theta/(1+\theta)$
ⓓ 3/4

1p **4b** Determine the posterior $p(\theta|x=1, m_1)$

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

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

$$p(x|\theta, m_2) = (1-\theta)^x \theta^{(1-x)}$$
$$p(\theta|m_2) = 2\theta$$

1p **4c** Determine the probability $p(x=1|m_2)$.

ⓐ 2/3 ⓑ 1/4 ⓒ 1/2 ⓓ $1/3$

1p **4d** Now assume that the model priors are given by $p(m_1) = 1/3$ and $p(m_2) = 2/3$. Compute the probability $p(x = 1)$ by "Bayesian model averaging", i.e., by weighing the predictions of both models appropriately.

ⓐ $7/18$

ⓑ 1/2

ⓒ 4/9

ⓓ 8/18

1p **4e** Compute the fraction of posterior model probabilities $\frac{p(m_1|x=1)}{p(m_2|x=1)}$

ⓐ $3/4$ ⓑ $4/9$ ⓒ $5/9$ ⓓ 2/3

Miscellaneous

1p **5a** For a state space model with given process model $p(z_t|z_{t-1})$ and observation model $p(x_t|z_t)$, write out how to recursively update the latent state estimate $p(z_t|x_{1:t})$

ⓐ $p(z_t|x_{1:t}) = p(x_t|z_t) \sum_{z_{t-1}} p(z_t|z_{t-1}) p(z_{t-1}|x_{1:t-1})$

ⓑ $p(z_t|x_{1:t}) \propto p(x_t|z_t) \sum_{z_{t-1}} p(z_t|z_{t-1}) p(z_{t-1}|x_{1:t-1})$

ⓒ $p(z_t|x_{1:t}) = \sum_{x_t} p(x_t|z_t) p(z_t|x_{1:t-1})$

ⓓ $p(z_t|x_{1:t}) \propto p(z_t, x_{1:t})$

1p **5b** Which of the following statements are consistent with the Free Energy Principle:
(a) An active inference agent holds a generative model for its sensory inputs
(b) Actions are inferred from differences between the predicted and desired future observations.
(c) Actions are inferred from differences between the predicted and actual future observations.
(d) An active inference agent focuses on explorative behavior only.

ⓐ (a) and (b)

ⓑ (a)

ⓒ (b) and (d)

ⓓ (c) and (d)

2p **5c** Given is a model

$$p(z) = \mathcal{N}(z|0, I)$$
$$p(x|z) = \mathcal{N}(x|Wz, \Sigma)$$

Work out an expression for the marginal $p(x)$.

ⓐ $p(x) = \mathcal{N}(x|0, W^T W + \Sigma)$

ⓑ $p(x) = \mathcal{N}(x|Wz, \Sigma)$

ⓒ $p(x) = \mathcal{N}(x|0, W\Sigma W^T)$

ⓓ $p(x) = \mathcal{N}(x|0, WW^T + \Sigma)$

1p **5d** Which of the following statements are true?
(a) If X and Y are independent Gaussian-distributed variables, then $Z = XY$ is also Gaussian distributed.
(b) If X and Y are independent Gaussian-distributed variables, then $Z = 3X - Y$ is also Gaussian distributed.
(c) The sum of two Gaussian distributions is always also a Gaussian distribution.
(d) Discriminative classification is more similar to regression than to density estimation.

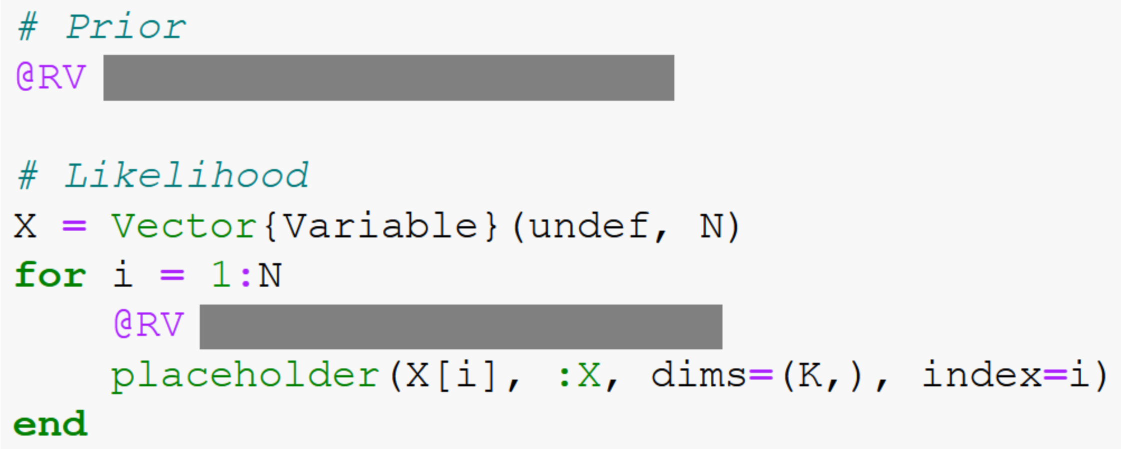
- a (b) and (c)
- b (a) and (d)
- c (b) and (d)
- d (b) and (c)

Probabilistic Programming

1p **6a** Suppose we have the following data set:

$$X = \begin{bmatrix} 0 & 1 & 1 & 0 & 2 \end{bmatrix}$$

with $N = 5$ samples and $K = 3$ possible discrete outcomes. Consider the following model specification code:



```
# Prior
@RV [redacted]

# Likelihood
X = Vector{Variable}(undef, N)
for i = 1:N
    @RV [redacted]
    placeholder(X[i], :X, dims=(K,), index=i)
end
```

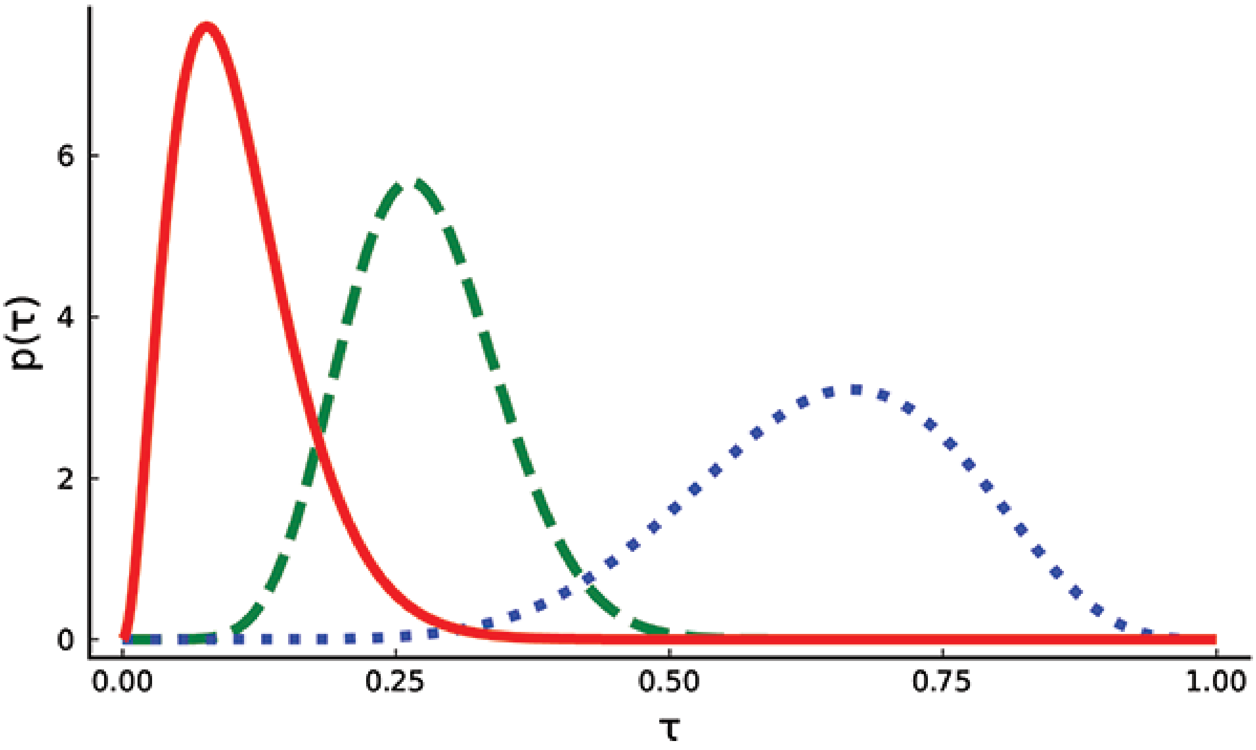
Which of the options below allows for sum-product message passing?

- a $\theta \sim$ Dirichlet$(1, K)$ and $X[i] \sim$ Bernoulli(θ)
- b $\theta \sim$ Beta(K) and $X[i] \sim$ Categorical(θ, N)
- c $\theta \sim$ Dirichlet(ones(K)) and $X[i] \sim$ Categorical(θ)
- d $\theta \sim$ Categorical(K) and $X[i] \sim$ Dirichlet(θ)

1p **6b** Which of the following is a correct factor node specification?

- ⓐ @RV γ ~ Beta$($ones$(9))$
- ⓑ @RV γ ~ Gamma(3)
- ⓒ @RV γ ~ GaussianMeanVariance$(-1,-1)$
- ⓓ @RV γ ~ Bernoulli(0.8)

1p **6c** Below are plotted three probability density functions over a parameter τ:



Two of these functions correspond to messages from factor nodes and one of them is a marginal distribution based on the product of the two messages.

Which one corresponds to the marginal distribution?

- ⓐ The solid red line.
- ⓑ The dashed green line.
- ⓒ The dotted blue line.
- ⓓ None of the lines could be a marginal distribution computed from the other distributions.

This page is left blank intentionally