UNIVERSITY OF AMSTERDAM

Faculty of Science

Exam

Machine Learning 2

Final Exam
Date: May 29, 2019
Time: 13:00-16:00

Number of pages: 6 (including front page)
Number of questions: 4
Maximum number of points to earn: 89
At each question is indicated how many points it is worth.

BEFORE YOU START

- Please **wait** until you are instructed to open the booklet.
- Check if your version of the exam is complete.
- Write down **your name, student ID number**, and if applicable the **version number** on **each sheet** that you hand in. Also **number the pages**.
- Your **mobile phone** has to be switched off and in the coat or bag. Your **coat and bag** must be under your table.
- **Tools allowed**: 1 handwritten double-sided A4-size cheat sheet, pen.

PRACTICAL MATTERS

- The first 30 minutes and the last 15 minutes you are not allowed to leave the room, not even to visit the toilet.
- You are obliged to identify yourself at the request of the examiner (or his representative) with a proof of your enrollment or a valid ID.
- During the examination it is not permitted to visit the toilet, unless the proctor gives permission to do so.
- 15 minutes before the end, you will be warned that the time to hand in is approaching.
- If applicable, please fill out the evaluation form at the end of the exam.

Good luck!

1 Information Theory **/16**

a) We define the *higher interaction information* between three random variables X, Y, Z as follows:

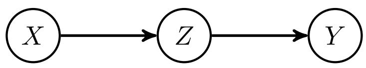
$$I(X;Y;Z) := I(X;Y) - I(X;Y|Z).$$

Show that we have the symmetry: **/6**

$$I(X;Y;Z) = I(X;Z;Y).$$

b) Sketch an "information diagram" for three random variables X, Y, Z (three circles intersecting each other) and indicate which part of the diagram corresponds to which information theoretic quantity. Indicate at least the (marginal) entropies, the joint entropy, a conditional entropy, a conditional mutual information and an (unconditional) mutual information with the correct variables. Where would you put $I(X;Y;Z)$ in this diagram? **/6**

c) Consider the following Markov chain:



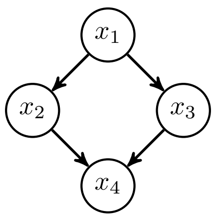
Prove the following *information processing inequality*: **/4**

$$I(X;Z) \geq I(X;Y).$$

Hint: You may use the symmetry of the higher interaction information and properties of the graphical model. In addition, you may use that $I(X;Y \mid Z) \geq 0$ *and* $I(X;Y \mid Z) = 0 \iff X \perp\!\!\!\perp Y \mid Z$.

2 A simple Bayesian network **/20**

Consider the following Bayesian network:



a) Write down the factorization of the joint probability density $p(x_1, x_2, x_3, x_4)$ implied by the Bayesian network. **/1**

b) Which of the following (conditional) independences necessarily hold in the Bayesian network?

(i) $x_1 \perp\!\!\!\perp x_2 | \varnothing$

(ii) $x_1 \perp\!\!\!\perp x_2 | x_3$

(iii) $x_1 \perp\!\!\!\perp x_2 | x_4$

(iv) $x_1 \perp\!\!\!\perp x_2 | \{x_3, x_4\}$

(v) $x_1 \perp\!\!\!\perp x_4 | \varnothing$

(vi) $x_1 \perp\!\!\!\perp x_4 | x_3$

(vii) $x_1 \perp\!\!\!\perp x_4 | x_2$

(viii) $x_1 \perp\!\!\!\perp x_4 | \{x_2, x_3\}$

(ix) $x_2 \perp\!\!\!\perp x_3 | \varnothing$

(x) $x_2 \perp\!\!\!\perp x_3 | x_1$

(xi) $x_2 \perp\!\!\!\perp x_3 | x_4$

(xii) $x_2 \perp\!\!\!\perp x_3 | \{x_1, x_4\}$

/4

c) Draw the Markov Random Field representation of the Bayesian network. List all maximal cliques and express the corresponding potential functions in terms of conditional probability densities. **/3**

d) Which of the following (conditional) independences necessarily hold according to the graph of the Markov Random Field?

(i) $x_1 \perp\!\!\!\perp x_2 | \varnothing$

(ii) $x_1 \perp\!\!\!\perp x_2 | x_3$

(iii) $x_1 \perp\!\!\!\perp x_2 | x_4$

(iv) $x_1 \perp\!\!\!\perp x_2 | \{x_3, x_4\}$

(v) $x_1 \perp\!\!\!\perp x_4 | \varnothing$

(vi) $x_1 \perp\!\!\!\perp x_4 | x_3$

(vii) $x_1 \perp\!\!\!\perp x_4 | x_2$

(viii) $x_1 \perp\!\!\!\perp x_4 | \{x_2, x_3\}$

(ix) $x_2 \perp\!\!\!\perp x_3 | \varnothing$

(x) $x_2 \perp\!\!\!\perp x_3 | x_1$

(xi) $x_2 \perp\!\!\!\perp x_3 | x_4$

(xii) $x_2 \perp\!\!\!\perp x_3 | \{x_1, x_4\}$

/4

e) Draw the factor graph representation of the Bayesian network. Express each factor in terms of conditional probability densities. **/2**

f) Write down explicitly the steps that the variable elimination algorithm makes to calculate the marginal distribution $p(x_1)$ when eliminating first x_2, then x_3 and finally x_4. Clearly indicate the new factors introduced by the variable elimination algorithm and on which variables they depend. **/3**

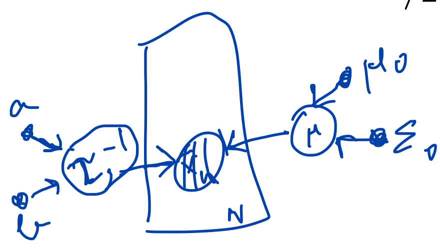
g) The Loopy Belief Propagation algorithm (also known as Sum-Product Algorithm) works by passing messages along the edges of the factor graph representation. Write down the sum-product message update equations for all messages that depend on variable x_1. You may assume the variables to be discrete. **/3**

3 Sampling

/23

Consider a model with a Gaussian likelihood:

$$p(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \,|\, \boldsymbol{\mu}, \tau) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{x}_n \,|\, \boldsymbol{\mu}, \tau^{-1}\boldsymbol{C})$$



where all $\boldsymbol{x}_n \in \mathbb{R}^D$, and the following prior distribution for $\boldsymbol{\mu}$ and τ:

$$p(\boldsymbol{\mu} \,|\, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) = \mathcal{N}(\boldsymbol{\mu} \,|\, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$$
$$p(\tau \,|\, a, b) = \mathrm{Gam}(\tau \,|\, a, b)$$

The model has parameters $\boldsymbol{C}$, $\boldsymbol{\mu}_0$, $\boldsymbol{\Sigma}_0$, a, b. Suppose we have a dataset $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)$ of i.i.d. samples. We will construct a Gibbs sampler to sample the posterior $p(\boldsymbol{\mu}, \tau \,|\, \boldsymbol{X})$ from this model.

Hint: some properties of the probability distributions used in this exercise:

$$\mathrm{Gam}(\tau \,|\, a, b) = \frac{1}{\Gamma(a)} b^a \tau^{a-1} \exp(-b\tau),$$
$$\mathrm{E}(\tau \,|\, a, b) = \frac{a}{b}, \quad \mathrm{Var}(\tau \,|\, a, b) = \frac{a}{b^2}.$$
$$\mathcal{N}(\boldsymbol{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right),$$
$$\mathrm{E}(\boldsymbol{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \boldsymbol{\mu}, \quad \mathrm{Cov}(\boldsymbol{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \boldsymbol{\Sigma}.$$

a) Draw the model in plate notation. Clearly distinguish observed variables, latent variables, parameters, and make clear which variable subscripts are "looped over" if you use plates. **/3**

b) Write down an explicit expression for the joint probability $p(\boldsymbol{\mu}, \tau, \boldsymbol{X})$. **/1**

c) Write down the pseudocode for a Gibbs sampler that samples from $p(\boldsymbol{\mu}, \tau \,|\, \boldsymbol{X})$. **/2**

d) Are the samples that the Gibbs sampler generates independent of each other? **/1**

e) Show that the conditional distribution $p(\boldsymbol{\mu} \,|\, \tau, \boldsymbol{X})$ is a Gaussian distribution and give an explicit expression of its parameters. **/4**

f) Show that the conditional distribution $p(\tau \,|\, \boldsymbol{\mu}, \boldsymbol{X})$ is a Gamma distribution and give an explicit expression of its parameters. **/4**

g) In order to implement the Gibbs sampler, you need to be able to sample from a Gamma distribution. Which sampling method to sample from $\mathrm{Gam}(\tau \,|\, a, b)$ would be suitable in the context of the Gibbs sampler? What will be the main challenge in implementing it? **/2**

Suppose you do not have a subroutine that samples from a Gamma distribution, but you do have a subroutine `randnorm(`μ`,`σ`)` that samples from a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ with given mean μ and standard deviation σ. One way to sample from $\mathrm{Gam}(a, b)$ is to use importance sampling.

h) Give an expression for the importance weights in terms of the occurring parameters. **/1**

i) Explain how the samples output by the importance sampler can be used in order to approximate the expectation value

$$\mathrm{E}(f(\tau)) = \int f(\tau)\mathrm{Gam}(\tau|a,b)d\tau$$

where $f(\tau)$ is some function of τ. **/1**

j) Give a detailed explanation of how you would implement that importance sampler using `randnorm`. How would you choose μ and σ in order to make the sampler efficient? **/3**

k) Why is the importance sampler not suitable to use within the context of the Gibbs sampler? **/1**

4 Variational EM for mixture of Bernoulli distributions **/30**

Consider a multivariate Bernoulli distribution

$$p(\boldsymbol{x}|\boldsymbol{\mu}) = \prod_{i=1}^{D} \mu_i^{x_i}(1-\mu_i)^{1-x_i}$$

where $\boldsymbol{x} = (x_1, \ldots, x_D)$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_D)$, with $\mu_i \in [0,1], x_i \in \{0,1\}$ for $i = 1, \ldots, D$.

a) What is the mean of $\boldsymbol{x}$ under this distribution? **/1**

b) What is the covariance matrix of $\boldsymbol{x}$ under this distribution? **/2**

Now consider a mixture of K of these multivariate Bernoulli distributions

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}|\boldsymbol{\mu}_k)$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$, and

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i}(1-\mu_{ki})^{1-x_i}$$

c) What is the mean of $\boldsymbol{x}$ under this mixture distribution? **/1**

Suppose we are given a data set $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$.

d) Write down the log-likelihood function for this model. Make the expression as explicit as possible, and use brackets to remove any ambiguity regarding what is summed over in the expression. **/3**

e) Why doesn't standard maximum-likelihood work here? **/1**

We will use the Variational EM algorithm to learn the parameters of the model. For each datapoint $\boldsymbol{x}_n$, introduce a latent variable $\boldsymbol{z}_n = (z_{n1}, \ldots, z_{nK})$ which is a one-of-K coded binary vector that indicates the latent class of that datapoint. In other words: the latent variable $\boldsymbol{z}_n$ has K components, all of which are 0 except for the k'th one that is 1, where k is the latent class for data point $\boldsymbol{x}_n$. Using these conventions, for data point $\boldsymbol{x}_n$ and associated latent class $\boldsymbol{z}_n$, we can write:

$$p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\mu}, \boldsymbol{\pi}) = p(\boldsymbol{z}_n|\boldsymbol{\pi})p(\boldsymbol{x}_n|\boldsymbol{z}_n, \boldsymbol{\mu}) = \prod_{k=1}^{K} \pi_k^{z_{nk}} p(\boldsymbol{x}_n|\boldsymbol{\mu}_k)^{z_{nk}}$$

f) Write down the complete-data log-likelihood function for this model. Make the expression as explicit as possible, and use brackets to remove any ambiguity regarding what is summed over in the expression. **/3**

g) Draw the corresponding graphical model using plate notation. Clearly distinguish observed variables, latent variables, parameters, and make clear which variable subscripts are "looped over" if you use plates. **/3**

h) Write down the general form of the VEM objective function (ELBO) $\mathcal{B}(q, \boldsymbol{\theta})$ and show that in this model we have the equality: **/3**

$$\mathcal{B}(\{q_n\}, \boldsymbol{\pi}, \boldsymbol{\mu}) = \sum_{n=1}^{N} \sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \sum_{k=1}^{K} z_{nk} \left(\log \pi_k + \sum_{i=1}^{D} \big(x_{ni} \log \mu_{ki} + (1 - x_{ni}) \log(1 - \mu_{ki}) \big) \right) - \sum_{n=1}^{N} \sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \log q_n(\boldsymbol{z}_n)$$

i) Include Lagrange multipliers for all constraints in the model and construct the Lagrangian $\tilde{\mathcal{B}}$ from $\mathcal{B}$. Make the Lagrangian as explicit as possible. **/3**

j) Work out the details of the E-step, i.e., optimize $\tilde{\mathcal{B}}$ with respect to q_n for all $n = 1, \ldots, N$. Solve the equation. What is the interpretation of $q_n(\boldsymbol{z}_n)$? **/4**

k) Work out the details of the M-step for $\boldsymbol{\pi}$, i.e., optimize $\tilde{\mathcal{B}}$ with respect to π_k for all k. Solve the equation. **/3**

l) Work out the details of the M-step for $\boldsymbol{\mu}$, i.e., optimize $\tilde{\mathcal{B}}$ with respect to μ_{ki} for all k, i. Solve the equation. **/3**