UNIVERSITY OF AMSTERDAM

Faculty of Science

Exam

Machine Learning 2

Midterm Exam
Date: April 26, 2018
Time: 9:00-11:00

Number of pages: 5 (including front page)
Number of questions: 4
Maximum number of points to earn: 90
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:

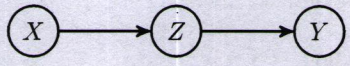
$$I(X;Y;Z) := I(X;Y) - I(X;Y|Z).$$

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 Exponential Families

/26

Consider the family of binomial distributions with a fixed number of trials K ($K \geq 1$), and parameter $\pi \in [0,1]$:

$$\mathrm{Bin}(x|\pi) = \binom{K}{x} \cdot \pi^x \cdot (1-\pi)^{K-x},$$

with $x \in \{0,1,2,\ldots,K\}$ and the binomial coefficient

$$\binom{K}{x} := \begin{cases} \frac{K!}{x! \cdot (K-x)!} & \text{for} \quad x \in \{0,1,2,\ldots,K\}, \\ 0 & \text{otherwise.} \end{cases}$$

a) Show that the above family of binomial distributions can be written as an exponential family with one natural parameter η. (Note that K is fixed and not considered a varying parameter of the above family.) Write η as a function of π and vice versa. What is the corresponding sufficient statistic, the base function and log-partition function (as a function of η)? /10

b) Use the log-partition function to derive mean and variance of the sufficient statistic. /4

c) Write down the explicit form of the family of conjugate priors (up to normalizing constant). Clearly indicate the hyper-parameters. /4

d) Derive the update rule for the hyper-parameters when passing from the (conjugate) prior distribution to the corresponding posterior distribution after observing N i.i.d. samples from the binomial distribution. /4

e) Decide if the following family of distributions with parameters $a < b$ can be written as an exponential family:

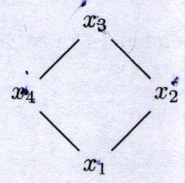
$$p(x|a,b) := \begin{cases} \frac{1}{b-a} & \text{for} \quad a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

Reason why. /4

3 Graphical Models

/21

Consider the following Markov Random Field:



a) Write down the corresponding factorization of the joint density $p(x_1, x_2, x_3, x_4)$ in terms of the potential functions that correspond to the MRF. /3

b) (i) Write down all independences of the form $A \perp\!\!\!\perp B$ (where A and B are two different variables from $\{x_1, x_2, x_3, x_4\}$) that hold according to the MRF. /2

(ii) Write down all conditional independences of the form $A \perp\!\!\!\perp B \,|\, C$ (where A and B are two different variables, and C is a non-empty set of different variables, such that $\{A\}, \{B\}$ and C are mutually disjoint) that hold according to the MRF. /4

c) Draw the factor graph representation corresponding to the MRF. /2

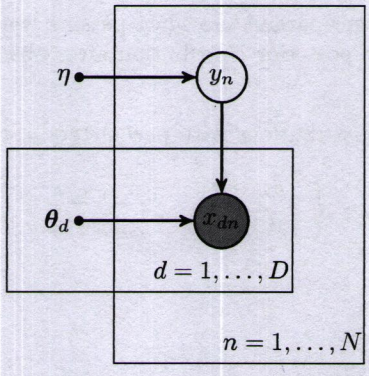
d) Draw a Bayesian Network with a minimal number of edges that can represent the exact joint density $p(x_1, x_2, x_3, x_4)$ modeled by the MRF, no matter how the potential functions (parameters) of the MRF are chosen.
Hint: we are ***not*** *asking you for a Bayesian network that entails exactly the same (conditional) independences as the MRF. Note that in general, a Bayesian network with a complete DAG can represent* ***any*** *joint density, although it might not imply all conditional independences in that joint density. If it wouldn't be for the requirement of a minimal number of edges, that would already be a possible answer to the question.* /4

e) Same question as b), but now for the Bayesian network you provided in the last question (rather than for the MRF). /6

4 Naïve Bayes Classifier

/27

Consider the Naïve Bayes Classifier for D observed features $x_1, \ldots, x_D$ and one latent class label y, for a dataset with N data points. It can be represented as the following graphical model:



where we introduced separate variables for each data point, and parameters $\boldsymbol{\eta}$ and $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D)$.

a) Which of the following conditional independences must hold according to this graphical model? **/4**

(i) $x_{1n} \perp\!\!\!\perp x_{2n}$

(ii) $x_{1n} \perp\!\!\!\perp x_{2n} \,|\, y_n$

(iii) $x_{1n} \perp\!\!\!\perp x_{1m}$ with $m \neq n$

(iv) $x_{1n} \perp\!\!\!\perp x_{1m} \,|\, y_n, y_m$ with $m \neq n$

b) Write down an expression for the likelihood of the data $(\boldsymbol{X}, \boldsymbol{y}) = (\boldsymbol{x}_n, y_n)_{n=1}^N$, formulating it as explicitly as possible in terms of the basic "building blocks" of the model. **/4**

c) Once the model parameters have been learnt from the data, we can predict the most likely class label y^* of a new data point $\boldsymbol{x}^*$. Provide an equation that describes which calculation yields the optimal prediction, formulating it as explicitly as possible in terms of the basic "building blocks" of the model.
Hint: use the famous rule by an English statistician, philosopher and Presbyterian minister that the name of this classifier refers to. **/4**

Let us now pick a particular parameterization: assume that the y_n are binary with $\text{Bern}(y_n \,|\, \eta)$ distribution, and that the x_{dn} are conditional Gaussians with conditional distribution $p(x_{dn} \,|\, y_n) = \mathcal{N}(x_{dn} \,|\, \beta_d y_n + \gamma_d, \sigma_d^2)$, with $\boldsymbol{\theta}_d := (\beta_d, \gamma_d, \sigma_d)$. Remember that

$$\text{Bern}(y \,|\, \eta) = \eta^y (1 - \eta)^{1-y}$$

and

$$\mathcal{N}(x \,|\, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

d) Derive the maximum likelihood equation for η (i.e., the equation that this parameter must satisfy when the likelihood is maximized), assuming that the other parameters (σ_d, γ_d, and β_d, for $d = 1, \ldots, D$) have fixed values. Solve the equation with respect to η. **/6**

e) Calculate the ML estimate for η from the following data: **/1**

x_1	x_2	x_3	y
1.5	2.0	1.2	0
2.3	0.0	1.7	0
7.4	2.0	1.4	1

In case of missing data, one can use the model to perform "imputation", i.e., replacing missing values with their most likely values given the available data. In the rest of this exercise, assume for simplicity that there are only $D = 3$ features. Suppose that x_1^*, x_2^* are observed, but x_3^* and y^* are missing (latent).

f) Draw the relevant directed graphical model for a single data point (x_1, x_2, x_3, y), omitting the parameters, clearly indicating which variables are observed and which are latent, and *without* using plate notation. **/2**

g) Draw the corresponding factor graph representation and express the factors in terms of the model parameters. Introduce separate factors for representing the observed values $x_1 = x_1^*$ and $x_2 = x_2^*$. **/6**