UNIVERSITY OF AMSTERDAM

Faculty of Science

Exam

Machine Learning 2

Master AI year 1

Final Exam
Date: May 30, 2018
Time: 09:00-12:00

Number of pages: 6 (including front page)
Number of questions: 4
Maximum number of points to earn: 60
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**: 2 handwritten double-sided A4-size cheat sheets, 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 Selection bias and Simpson's paradox

/16

Let A, B, S be three binary random variables.

a) Reichenbach's Principle says: an observed dependence between two events A and B can be explained by (i) A causing B, (ii) B causing A, (iii) A and B being effects of a common cause (or any combination of the three). Draw the three causal Bayesian networks corresponding with the three elementary explanations in Reichenbach's Principle. /1

b) Does $p(A, B \mid S = 0) \neq p(A \mid S = 0)p(B \mid S = 0)$ imply that $A \not\perp\!\!\!\perp B \mid S$? /1

c) Does $p(A, B \mid S = 0) = p(A \mid S = 0)p(B \mid S = 0)$ imply that $A \perp\!\!\!\perp B \mid S$? /1

Interpret the three binary variables now as follows: A means "start engine broken", B means "empty battery", S means "car starts". A car mechanic notices that for the cars that don't start ($S = 0$), there is a strong dependence between the start engine being broken and the battery being empty. This is an example of *selection bias*.

d) Draw a causal Bayesian network with variables A, B, S that most accurately models the situation for the car mechanic, assuming that $A \perp\!\!\!\perp B$. /1

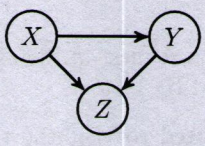
e) What is a fourth possible explanation of an observed dependence between two events that is *not* suggested by Reichenbach's Principle? /1

Suppose you are a teacher who wants to investigate whether learning for the midterm exam has a positive effect on the probability of passing the midterm. You take a survey amongst all students that showed up for the *final* exam. Interestingly, you find that in this group of students, the percentage of students that passed the midterm *without* learning is *higher* than the percentage of students that passed the midterm and learnt for it. You are wondering whether this counterintuitive finding could be due to selection bias, and you decide to get more data. You take a survey by email amongst the other students in the course, i.e., the ones who did *not* show up for the final exam. Surprisingly, you obtain a similar counterintuitive finding for the group of students that did not show up for the final exam: the conditional probability of passing is higher for those who didn't learn than for those who learnt.

In order to model this situation, define the three binary variables:

X: "learnt for midterm exam"
Y: "passed midterm exam"
Z: "shows up for final exam"

and consider the following causal Bayesian network:



f) What is the difference between a Bayesian network and a *causal* Bayesian network? /1

g) Write down how the joint distribution $p(X, Y, Z)$ factorizes according to this causal Bayesian network. /1

h) Write down an explicit expression for $p(Y = 1 \mid \mathrm{do}(X = x))$ in terms of the factors in the factorization of $p(X, Y, Z)$. /1

i) Write down an explicit expression for $p(Y = 1 \mid X = x, Z = z)$ in terms of the factors in the factorization of $p(X, Y, Z)$. /2

We parameterize the distribution as follows:

$p(X = 1) = p$

X	$p(Y = 1 \mid X)$
0	v
1	w

X	Y	$p(Z = 1 \mid X, Y)$
0	0	a
0	1	b
1	0	c
1	1	d

j) Give values for the model parameters such that we get an instance of Simpson's paradox when we condition on Z. More precisely, give a choice for the parameters (p, v, w, a, b, c, d) such that

$$p(Y = 1 \mid X = 1, Z = 0) < p(Y = 1 \mid X = 0, Z = 0),$$
$$p(Y = 1 \mid X = 1, Z = 1) < p(Y = 1 \mid X = 0, Z = 1)$$

yet

$$p(Y = 1 \mid X = 1) > p(Y = 1 \mid X = 0)?$$

Justify your answer with a calculation.
Hint: consider e.g. the case $c = 1 - b$, $d = 1 - a$, $w = 1 - v$. /4

k) The *causal effect* of learning for the midterm exam on passing for the midterm exam is given by

$$p(Y = 1 \mid \mathrm{do}(X = 1)) - p(Y = 1 \mid \mathrm{do}(X = 0)).$$

Express it in terms of the model parameters. For model parameters that satisfy Simpson's paradox as in the previous question, would learning for the midterm exam have a positive effect, negative effect, or no effect at all on passing for the midterm? /2

2 Variational Bayes for the Rasch Model /11

Consider an exam in which student s answers question q either correctly ($x_{sq} = 1$) or incorrectly ($x_{sq} = 0$). For N students and Q questions, the performance of all students is given by the $N \times Q$ binary matrix $\boldsymbol{X}$. A simple and common way to evaluate the ability of each student is to define the ability a_s of student s as the fraction of questions that the student answered correctly. However, that would not take into account that some questions may be more difficult than others; a student that answered difficult questions should have a higher ability than a student who answered the same number of easy questions. However, *a priori* we do not know the difficulty of the questions. The Rasch Model is a simple model that allows one to estimate the difficulty of the questions and the student's abilities simultaneously from the data. To account for inherent differences in question difficulty, the Rasch model posits that the probability that a student s gets a question q correct is based on the student's ability α_s and the difficulty of the question β_q:

$$p(x_{sq} = 1 \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \sigma(\alpha_s - \beta_q)$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the logistic function. Under this model, the higher the ability of the student is above the difficulty of the question, the more likely it is that the student will answer the question correctly. The model assumes that the data are i.i.d. given the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$.

a) Draw the Rasch model as a generative directed graphical model using plate notation, treating $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as parameters. Clearly distinguish variables from parameters, observed from latent variables, and indices that are "looped over" in plates. /3

The likelihood of the data $\boldsymbol{X}$ given the model parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$ is given by:

$$p(\boldsymbol{X} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{s=1}^{N} \prod_{q=1}^{Q} \sigma(\alpha_s - \beta_q)^{x_{sq}} \sigma(\beta_q - \alpha_s)^{1-x_{sq}}$$

b) Derive an explicit expression for the log-likelihood of the data $\boldsymbol{X}$ given the parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$. /1

c) One way to calculate the maximum likelihood estimate for the parameters is by using gradient descent. Give the explicit pseudocode that one could implement directly without having to do any remaining calculations.
Hint: you may use that $\frac{d}{dx} \ln \sigma(x) = \sigma(-x)$. /2

If there is only a small amount of data, the Rasch model may overfit. We will now turn to a Bayesian extension of the Rasch model, which assumes Gaussian priors for the parameters:

$$p(\boldsymbol{\alpha}) = \prod_{s} \mathcal{N}(\alpha_s \mid 0, \sigma^2), \qquad p(\boldsymbol{\beta}) = \prod_{q} \mathcal{N}(\beta_q \mid 0, \tau^2)$$

with hyperparameters σ^2 and τ^2.

d) Draw the Bayesian extension of the Rasch model as a generative directed graphical model using plate notation. Clearly distinguish variables from parameters, observed from latent variables, and indices that are "looped over" in plates. /3

The posterior cannot be calculated analytically, and approximations are required. One way to do so is to use Variational Bayes. Remember that for a fully factorizing approximation $p(\boldsymbol{\theta} \mid \boldsymbol{X}) \approx \prod_i q_i(\theta_i)$ the VB update equations are given by

$$q_i(\theta_i) \propto \exp\left(\mathbb{E}_{q_{\setminus i}}[\ln p(\boldsymbol{X}, \boldsymbol{\theta})]\right).$$

e) Taking

$$q(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \left(\prod_{s} q_s(\alpha_s)\right)\left(\prod_{q} q_q(\beta_q)\right)$$

with $q_s(\alpha_s) = \mathcal{N}(\alpha_s \mid \mu_s, \sigma_s^2)$ and $q_q(\beta_q) = \mathcal{N}(\beta_q \mid \nu_q, \tau_q^2)$, simplify (as far as possible) the r.h.s. of the VB update equation for $q_s(\alpha_s)$. (Don't attempt to solve the remaining integral—it is not analytically tractable.) /2

Further approximations are necessary, for example, local bounds can be used to bound the remaining integrals of the form $\int \mathcal{N}(x \mid \mu, \sigma^2) \ln \sigma(\gamma x)\, dx$. Another possibility would be to treat the Bayesian extension of the Rasch model using sampling methods.

3 Sampling

/19

Consider the following probability density function:

$$p(x) = \frac{2}{x^2} \qquad x \in (1, 2) \tag{1}$$

Suppose your programming language only offers you a way to sample random numbers from the uniform distribution on $(0, 1)$.

a) Show how the transformation method can be used to obtain samples from the density in equation (1). Make sure that you explicitly state the required transformation and that your answer is as explicit as possible. **/3**

b) Provide a rejection sampler that samples from the density in equation (1). Give pseudo-code and make it as explicit as possible. In particular, clearly state what happens with "rejections". **/2**

Now we consider sampling a multivariate distribution with density $p(x_1, \ldots, x_K)$. A common approach to sampling from such distributions is by using a Markov chain with transition probabilities $T(\boldsymbol{x}' \mid \boldsymbol{x})$.

c) Show that a transition probability that satisfies detailed balance with respect to the distribution $p(\boldsymbol{x})$, i.e.,

$$p(\boldsymbol{x})T(\boldsymbol{x}' \mid \boldsymbol{x}) = p(\boldsymbol{x}')T(\boldsymbol{x} \mid \boldsymbol{x}') \qquad \forall \boldsymbol{x}, \boldsymbol{x}'$$

will leave that distribution invariant. **/1**

One sampler that uses a Markov chain is the Metropolis-Hastings algorithm. We consider here the special case with a *single* proposal distribution.

d) Describe how the Metropolis-Hastings algorithm with proposal distribution $q(\boldsymbol{x}' \mid \boldsymbol{x})$ would sample from $p(x_1, \ldots, x_K)$. Give explicit pseudocode. In particular, clearly describe what happens with "rejections". **/3**

e) Does one need to know the normalizing constant of $p(\boldsymbol{x})$ in order to apply the Metropolis-Hastings algorithm? Explain your answer. **/1**

f) Express the transition probabilities $T(\boldsymbol{x}' \mid \boldsymbol{x})$ in terms of the proposal distribution $q(\boldsymbol{x}' \mid \boldsymbol{x})$. *Hint: what happens in case of a 'reject'?* **/2**

g) Show that the Metropolis-Hastings algorithm satisfies the property of detailed balance with respect to $p(\boldsymbol{x})$. **/3**

Another popular sampler is the Gibbs sampler.

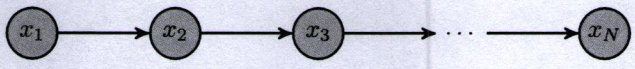
h) Describe the Gibbs sampler in this context using explicit pseudocode. **/2**

i) Show that a single step in the Gibbs sampler can be seen as a special case of a single step of a Metropolis-Hastings sampler. **/2**

4 Markov chain

/14

Consider a Markov chain with observed variables $\boldsymbol{X} = (x_1, \ldots, x_N)$:



a) Write down the expression for the joint probability distribution associated with this graphical model. **/1**

b) Show that the Markov property holds, i.e., that for $n = 1, \ldots, N-1$: **/1**

$$p(x_{n+1}|x_1, \ldots, x_n) = p(x_{n+1}|x_n)$$

Let all x_n take values in $\{1, 2, \ldots, K\}$. Assuming stationarity, we can parameterize

$$p(x_n|x_{n-1}, \boldsymbol{A}) = A_{x_{n-1}, x_n}$$

where $\boldsymbol{A} \in \mathbb{R}^{K \times K}$ is a matrix of transition probabilities. For the initial state, we write

$$p(x_1|\boldsymbol{\pi}) = \pi_{x_1}$$

where $\boldsymbol{\pi} \in \mathbb{R}^K$ parameterizes the probability distribution of the initial state.

c) Write down an explicit expression for the logarithm of the likelihood $p(\boldsymbol{X}|\boldsymbol{A}, \boldsymbol{\pi})$. **/1**

d) What constraints should the entries of the matrix $\boldsymbol{A}$ satisfy? **/2**

e) Maximize the log-likelihood with respect to the transition probabilities parameters $\boldsymbol{A}$ and provide an explicit expression for the maximum likelihood estimate $\hat{\boldsymbol{A}}$. Provide a complete derivation of your answer. **/4**

f) Write down an explicit expression for the (frequentist) predictive distribution for the upcoming observation x_{N+1} given the estimated $\hat{\boldsymbol{A}}$, and the observed values $x_1, \ldots, x_N$. **/1**

g) Let $M \in \{2, 3, \ldots\}$ be an integer. Write down an explicit expression for the predictive distribution of x_{N+M} given only observations $x_1, \ldots, x_N$ and the estimated $\hat{\boldsymbol{A}}$. **/3**

h) The answer to the previous question can be calculated by a well-known algorithm. Give a name of that algorithm. **/1**