Name :

Study program :

ID. NR. :

---

1. For each of the following sub-questions, you are asked to provide a *short but essential* answer. You should not need more than five sentences per answer.

a. Explain shortly how Bayes rule relates to machine learning. In your answer, you may assume a model $\mathcal{M}$ with prior distribution $p(\mathcal{M})$ and an observed data set D.

b. Explain the relation between Bayesian estimation, Maximum a Posteriori (MAP) estimation and Maximum Likelihood (ML) estimation. You may assume a context of a given model structure with unknown parameters θ and an observed data set D.

The following two sub-questions relate to a (Factor Analysis) model $x_n = \Lambda z_n + v_n$ for an observed data set $D = \{x_1, \ldots, x_N\}$. The modeling assumptions include $z_n \sim \mathcal{N}(0, I)$, $v_n \sim \mathcal{N}(0, \Psi)$ and $\varepsilon[z_n v_n^T] = 0$.

c. Show that the covariance matrix of the observed data x_n is equal to $\Lambda\Lambda^T + \Psi$.

d. Why is this model not interesting for unconstrained Ψ? How does probabilistic PCA handle this problem?

e. Which of the following statements are justified? You can pick more than one and read the sign '$\sim$' as: 'is similar to'. (Just pick the correct statements; no explanation needed).
1: discriminative classification $\sim$ density estimation
2: generative classification $\sim$ density estimation
3: hidden Markov model $\sim$ factor analysis through time
4: Kalman filtering $\sim$ unsupervised regression through time
5: clustering $\sim$ supervised classification

2. (EM for 2-component Gaussian mixture). Consider an observed IID data set $D = \{x_1, \ldots, x_N\}$ and a proposed model,

$$\begin{aligned} p(x_n) &= \sum_{k=0}^{1} p(x_n, z_n = k|\pi) \\ &= p(z_n = 1|\pi)p(x_n|z_n = 1) + p(z_n = 0|\pi)p(x_n|z_n = 0) \\ &= \pi\mathcal{N}_1(x_n) + (1-\pi)\mathcal{N}_0(x_n) \end{aligned}$$

where we used shorthand notation $\mathcal{N}_k(x_n) \equiv (2\pi\sigma_k^2)^{-1/2} \exp\left(-(x_n - \mu_k)^2/(2\sigma_k^2)\right)$ for the Gaussian distribution. We assume that the parameters $\theta = (\mu_0, \sigma_0^2, \mu_1, \sigma_1^2)$ are known, but the *mixing proportion* parameter π is unknown. The random variable $z_n \in \{0, 1\}$ is an *unobserved* 'cluster label'. In this question we will derive an EM-algorithm for maximum likelihood estimation of π. Let's assume that a estimate $\hat{\pi} = \pi^{(j)}$ is available from the previous iteration. We will now focus on the $(j+1)$-th iteration in the EM algorithm.

a. Describe shortly the E- and M-steps in the $(j+1)$-th iteration of the EM-algorithm. In particular, complete the following equation set (fill in the stars) for the $(j+1)$-th iteration and shortly describe the meaning of the equations: (Note: the expression $\langle f(x)\rangle_{p(x)}$ stands for the expectation of $f(x)$ w.r.t. probability distribution $p(x)$.)

$$\begin{aligned} q_n^{(j+1)} &= p(\star|\star) \quad \text{(E-step)} \\ \pi^{(j+1)} &= \arg\max_{\pi} \langle\star\rangle_\star \quad \text{(M-step)} \end{aligned}$$

b. Work out $p(x_n, z_n = 1|\pi)$ (hint: use product rule). Work out $p(x_n, z_n = 0|\pi)$. And now work out the joint distribution $p(x_n, z_n|\pi)$ to a Bernoulli distribution (as in eq.A1, see formula cheat sheet). In this question, you need to work out the probabilities in terms of z_n, $\mathcal{N}_0(x_n)$, $\mathcal{N}_1(x_n)$ and π.

c. Show that the complete-data log-likelihood $\ell_c(\pi) = \sum_n \log p(x_n, z_n|\pi)$ can be worked out to

$$\ell_c(\pi) = \sum_n z_n \log \frac{\pi \mathcal{N}_1(x_n)}{(1-\pi)\mathcal{N}_0(x_n)} + \sum_n \log(1-\pi)\mathcal{N}_0(x_n) \tag{1}$$

To finalize the E-step, we now take the expectation of the complete-data log-likelihood with respect to the posterior distribution $p(z_n|x_n, \pi^{(j)})$. It follows from Eq.1 that we need to compute the expected value of z_n. We'll compute the expected value of z_n in two stages:

d. First show that the expectation $\sum_{\{z_n\}} z_n \cdot p(z_n|x_n, \pi^{(j)})$ can be worked out as follows:

$$\sum_{\{z_n\}} z_n p(z_n|x_n, \pi^{(j)}) = p(z_n = 1|x_n, \pi^{(j)})$$

e. And now use Bayes rule to work out an expression for $p(z_n = 1|x_n, \pi^{(j)})$ in terms of $\pi^{(j)}$, $\mathcal{N}_0(x_n)$ and $\mathcal{N}_1(x_n)$.

If we use shorthand notation $\zeta_n = p(z_n = 1|x_n, \pi^{(j)})$, then the expected complete-data log-likelihood can be written as

$$\langle \ell_c(\pi) \rangle = \sum_n \zeta_n \log \frac{\pi \mathcal{N}_1(x_n)}{(1-\pi)\mathcal{N}_0(x_n)} + \sum_n \log(1-\pi)\mathcal{N}_0(x_n)$$

f. Set $\partial \langle \ell_c(\pi) \rangle / \partial \pi = 0$ and obtain an expression for $\pi^{(j+1)}$ in terms of $\pi^{(j)}$, $\mathcal{N}_0(x_n)$ and $\mathcal{N}_1(x_n)$ (i.e. write down the $(j+1)$-th iteration of the M-step).

3. You observe some data x^n. You ask two experts to explain the data.

Expert A uses a data compression system that needs 1537 bits to describe the parameters of the model and 438 bits to describe the data given the model.

Expert B gives you a system that needs 1325 bits for the parameters and 650 bits for the data, given the model.

a. Which expert's result do you prefer?
Explain (briefly) why you select that experts result.

b. You ask two additional experts.
Expert C gives you a model with a parameter description length of 1471 bits and a data description that needs 450 bits.
Expert D gives you a model with a parameter description length of 1464 bits and a data description that needs 543 bits.
Of the four experts A to D, which result do you prefer, and why?

4. Let X be a real valued random variable with probability density

$$p_X(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}, \quad \text{for all } x.$$

Also Y is a real valued random variable with conditional density

$$p_{Y|X}(y|x) = \frac{e^{-(y-x)^2/2}}{\sqrt{2\pi}}, \quad \text{for all } x \text{ and } y.$$

a. Give an (integral) expression for $p_Y(y)$.
Do not try to evaluate the integral.

b. Approximate $p_Y(y)$ using the Laplace approximation.
Give the detailed derivation, not just the answer.
Hint: You may use the following results.

Let

$$g(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}, \quad \text{and}$$
$$h(x) = \frac{e^{-(y-x)^2/2}}{\sqrt{2\pi}}, \quad \text{for some real value } y.$$

Then

$$\frac{\partial}{\partial x} g(x) = -x g(x)$$
$$\frac{\partial^2}{\partial x^2} g(x) = (x^2 - 1) g(x)$$
$$\frac{\partial}{\partial x} h(x) = (y - x) h(x)$$
$$\frac{\partial^2}{\partial x^2} h(x) = ((y - x)^2 - 1) h(x)$$

5. We implement an e-mail spam filter using two features that we can extract from an e-mail. A feature can be the occurrence of a particular word or phrase in the e-mail.

Given an e-mail E we denote the extracted features by F and G.
$F = 1$ means that feature F is present in the e-mail E.
$F = 0$ means that feature F is absent. And likewise for feature G.
The variable C indicates whether E is spam ($C = 1$) or not ($C = 0$).

We are given 247 e-mails that are already classified. The following table shows how many e-mails contained certain features and the classification.

F	G	C	nr of e-mails
0	0	0	15
0	0	1	28
0	1	0	18
0	1	1	25
1	0	0	8
1	0	1	75
1	1	0	10
1	1	1	68

a. From the table given above you can determine probability estimates using the maximum likelihood estimates. e.g. the probability $P(C = 1)$, i.e. the probability that an email will be spam, is approximated by:

$$P(C = 1) = \frac{\#\text{ of e-mails with } C = 1}{\text{total }\#\text{ of e-mails}} = \frac{196}{247} = 0.7935.$$

Note that the method using a beta prior would be better suited but we'll use the maximum likelihood because it is simpler.
Determine the following estimates.

$$P(F = 1|C = 0), P(F = 1|C = 1),$$
$$P(G = 1|C = 0), P(G = 1|C = 1),$$
$$P(F = 0, G = 0|C = 0), P(F = 0, G = 1|C = 0),$$
$$P(F = 1, G = 0|C = 0), P(F = 1, G = 1|C = 0),$$
$$P(F = 0, G = 0|C = 1), P(F = 0, G = 1|C = 1),$$
$$P(F = 1, G = 0|C = 1), P(F = 1, G = 1|C = 1).$$

Model M_1 for e-mail does not consider any feature. So $P(C)$ can be used to estimate the probability that the next e-mail will be spam or not. We will write that as $P(C|M_1)$.

b. Model M_2 considers only feature F to predict whether the next e-mail will be spam or not.

Use Bayes rule and the probability estimates determined in the previous question to determine an estimate for $P(C|M_2) = P(C|F)$.

Model M_3 considers feature G only and model M_4 considers both F and G. Model M_5 also considers both F and G but assumes that F and G are independent given the classification C.

c. Use Bayes rule again to show how you would calculate $P(C|M_5)$.

d. The models $M_1, M_2, \ldots, M_5$ all have a certain number of free parameters. Determine the number of free parameters for each of the five models.

e. Given the training set of the 247 e-mail as shown in the table above, which of the five models would you prefer? Use an MDL argument in your answer.

HINT: You will need to calculate an estimate for the email entropy for each model. For model M_1 you make an estimate of $H(C)$ using the maximum likelihood estimate $P(C = 1) = 0.7935$. Likewise you calculate for M_2 the entropy $H(C|F)$ and thus you'll need to compute $P(C, F)$. For M_3 you must compute the entropy $H(C|G)$; for M_4 you calculate $H(C|F, G)$ and for M_5 also $H(C|F, G)$ although this will be a different calculation than for M_4.

Appendix: formula sheet

The *Bernoulli distribution* is a discrete distribution having two possible outcomes labeled by $x = 0$ and $x = 1$ in which $x = 1$ ("success") occurs with probability θ and $x = 0$ ("failure") occurs with probability $1 - \theta$. It therefore has probability function

$$p(x|\theta) = \theta^x (1-\theta)^{1-x} \tag{A.1}$$

The *Gaussian distribution* with mean μ and variance σ^2 is defined as

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

Points that can be scored per question:

Question 1: each sub-question 2 points. Total 10 points.

Question 2: a) 2 points; b) 2 points; c) 2 points; d) 1 point; e) 1 point; f) 2 points. Total 10 points.

Question 3: a) 3 points; b) 3 points. Total 6 points.

Question 4: a) 3 points; b) 3 points. Total 6 points.

Question 5: a) 1 point; b) 1 point; c) 2 points; d) 2 points; e) 2 points. Total 8 points.

Max score that can be obtained: 40 points.
The final grade is obtained by dividing the score by 4 and rounding to the nearest integer.