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 three sentences per answer.

a. Consider a binary classification problem with two classes $\{y_1, y_2\}$ and input vector x. We are given a data set to train the parameters θ for a likelihood model of the form

$$p(y_k = 1|x, \theta) = \frac{1}{1 + e^{-\theta_k^T x}}$$

There a two fundamentally different ways to train θ, namely through a generative model or by discriminative training.
(1) Explain shortly how we train θ through a generative model. No need to work out all equations for Gaussian models, but explain the strategy in probabilistic modeling terms.
(2) Explain shortly how we train θ through a discriminative approach.

> (1) In a generative model, the class posterior is obtained through Bayes rule,
>
> $$p(y_k = 1|x, \theta) \propto p(x|y_k = 1, \theta)p(y_k = 1|\theta)$$
>
> In terms of ML training, this means we maximize the *joint* log-likelihood $\sum_n \log p(x_n, y_n|\theta)$ wrt θ. This leads to a structured breakdown of the model (and parameters) into a class-conditional likelihood $p(x|y_k = 1, \theta)$ and class priors $p(y_k = 1|\theta)$.
> (2) In a discriminative model, the posterior class density $p(y_k = 1|x, \theta)$ is directly trained, i.o.w. we maximize the *conditional* log-likelihood $\sum_n \log p(y_{nk}|x_n\theta)$. There's no structured model breakdown.

b. 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.

> $$\underbrace{p(\mathcal{M}|D)}_{\text{posterior}} = \frac{p(D|\mathcal{M})}{p(D)} \underbrace{p(\mathcal{M})}_{\text{prior}}$$
>
> Bayes rule relates what we know about a model before (prior) and after (posterior) having seen the data. The difference between the prior and posterior distributions for the model can be interpreted as a 'machine learning' effect. (Alternative answers are also possible).

c. What is the difference between supervised and unsupervised learning? Express the goals of these two learning methods in terms of a probability distribution. (I'm looking here for a statement such as: " Given ..., the goals of supervised/unsupervised learning is to estimate $p(\cdot|\cdot)$".)

> Given data $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ and a model $p(y|x, \theta)$, the goal of supervised learning is to estimate $p(\theta|D)$. Given data $D = \{x_1, \ldots, x_N\}$ and a model $p(x|\theta)$, the goal of unsupervised learning is to estimate $p(\theta|D)$.

d. In a particular model with hidden variables, the log-likelihood can be worked out to the following expression:

$$L(\theta; D) = \sum_n \log \left(\sum_k \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right)$$

Do you prefer a gradient descent or EM algorithm to estimate maximum likelihood values for the parameters? Explain your answer. (No need to work out the equations.)

Since this expression does not degenerate into simple MVGs, the EM approach is in practice preferred.

e. The maximum likelihood estimate (MLE) of the class-conditional mean in a classification problem can be expressed as

$$\hat{\mu}_k = \frac{\sum_n y_n^k x_n}{\sum_n y_n^k}$$

and the M-step update for the cluster mean in a clustering problem is given by

$$\hat{\mu}_k = \frac{\sum_n \gamma_n^k x_n}{\sum_n \gamma_n^k}$$

Explain the relation between y_n^k and γ_n^k. Is y_n^k a binary variable? And what about γ_n^k?

y_n^k are *binary* indicator variables, given by

$$y_n^k = \begin{cases} 1 & \text{if} \quad Y_n = k \\ 0 & \text{else} \end{cases}$$

γ_n^k are *soft* indicators, given by $\gamma_n^k = p(Z_n = k|x_n, \theta)$, where Z_n refers to the unobserved nth class label.

2. The lifetime $x > 0$ of a light bulb is postulated to be exponentially distributed with unknown mean $\mu > 0$, i.e.

$$p(x|\mu) = \frac{1}{\mu}\, e^{-x/\mu}$$

In order to estimate μ, the lifetimes $\mathbf{X} = \{x_1, \ldots, x_N\}$ of N independent bulbs are observed.

a. Work out the log-likelihood $\log p(\mathbf{X}|\mu)$.

$$\begin{aligned} \log \prod_{n=1}^{N} p(x_n|\mu) &= \sum_n \log\left(\frac{1}{\mu}\exp(-\frac{x_n}{\mu})\right) \\ &= -N\log\mu - \frac{1}{\mu}\sum_n x_n \\ &= -N\left(\log\mu + \frac{\bar{x}}{\mu}\right) \end{aligned}$$

b. What is the maximum likelihood estimate for μ based on observations $\mathbf{X}$?

$\mu = \bar{x}$

In a separate experiment M independent bulbs were tested, but the individual lifetimes were not recorded. We will use the symbols $\mathbf{Z} = \{z_1, \ldots, z_M\}$ for the *unobserved* lifetimes of bulbs $N+1, \ldots, N+M$ and $\{\mathbf{X}, \mathbf{Z}\} = \{x_1, \ldots, x_N, z_1, \ldots, z_M\}$ for the complete data set. Instead of lifetimes, we only recorded if a bulb had failed at time t. We record $y_m = 1$ if bulb $N+m$ still burned at time t and $y_m = 0$ if the bulb had already failed at time t. We will now derive an EM algorithm to estimate μ, based all $N+M$ observations.

c. Complete the following two formula's for the EM algorithm:

E-step : evaluate $p(\cdot|\cdot, \mu^{\text{old}})$

M-step : $\mu^{\text{new}} = \arg\max_{\cdot} \sum_{\cdot} p(\cdot|\cdot,\cdot)\log p(\cdot,\cdot|\cdot)$

E-step: evaluate $p(\mathbf{Z}|\mathbf{X},\mu^{\text{old}})$
M-step: $\mu^{\text{new}} = \arg\max_{\mu} \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\mu^{\text{old}}) \log p(\mathbf{Z},\mathbf{X}|\mu)$

d. Proof that the expected complete-data log-likelihood $\mathbb{E}[\log p(\mathbf{X},\mathbf{Z}|\mu)]$ equals

$$-(N+M)\log\mu - \frac{1}{\mu}\left(N\bar{x} + \sum_{m=1}^{M}\mathbb{E}[z_m]\right)$$

where $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$.

$$\begin{aligned}
\log p(\mathbf{X},\mathbf{Z}|\mu) &= \log\left(\prod_{n=1}^{N} p(x_n|\mu)\prod_{m=1}^{M} p(z_m|\mu)\right) \\
&= \sum_{n=1}^{N}\log\left(\frac{1}{\mu}\exp(-\frac{x_n}{\mu})\right) + \sum_{m=1}^{M}\log\left(\frac{1}{\mu}\exp(-\frac{z_n}{\mu})\right) \\
&= -(N+M)\log\mu - \frac{1}{\mu}\left(N\bar{x} + \sum_{m=1}^{M} z_m\right) \\
\mathbb{E}[p(\mathbf{X},\mathbf{Z}|\mu)] &= -(N+M)\log\mu - \frac{1}{\mu}\left(N\bar{x} + \sum_{m=1}^{M}\mathbb{E}[z_m]\right)
\end{aligned}$$

e. You can find the (local) optimum of $\mathbb{E}[\log p(\mathbf{X},\mathbf{Z}|\mu)]$ by setting its derivative w.r.t. μ to zero. Now differentiate $\mathbb{E}[\log p(\mathbf{X},\mathbf{Z}|\mu)]$ (see answer 2d) w.r.t. μ and set to zero to obtain the re-estimation formula (**M-step**) for μ.

$$\frac{\partial}{\partial\mu}\mathbb{E}[p(\mathbf{X},\mathbf{Z}|\mu)] = -\frac{N+M}{\mu} + \frac{1}{\mu^2}\left(N\bar{x} + \sum_{m=1}^{M}\mathbb{E}[z_m]\right)$$

Set to zero to obtain

$$\mu = \frac{1}{N+M}\left(N\bar{x} + \sum_{m=1}^{M}\mathbb{E}[z_m]\right)$$

We do not derive the $\mathbb{E}[z_m]$ for the **E-step**. Use the following equation instead

$$\mathbb{E}[z_m] = \begin{cases} t+\mu & \text{if } y_m = 1 \\ \mu - \frac{t\exp\left(-\frac{t}{\mu}\right)}{1-\exp\left(-\frac{t}{\mu}\right)} & \text{if } y_m = 0 \end{cases}$$

f. In total we found that r out of M bulbs had failed at time t. Derive an expression for $\sum_{m=1}^{M}\mathbb{E}[z_m]$ in terms of r and M.

$$\sum_{m=1}^{M}\mathbb{E}[z_m] = (M-r)(t+\mu^{\text{old}}) + r\left(\mu^{\text{old}} - \frac{t\exp\left(-\frac{t}{\mu^{\text{old}}}\right)}{1-\exp\left(-\frac{t}{\mu^{\text{old}}}\right)}\right)$$

g. Put the results of the last two exercises together and derive the re-estimation formula (**M-step**) for μ (in terms of a previous estimate of μ^{old}).

$$\mu^{\text{new}} = \frac{1}{N+M}\left(N\bar{x} + (M-r)(t+\mu^{\text{old}}) + r\left(\mu^{\text{old}} - \frac{t\exp\left(-\frac{t}{\mu^{\text{old}}}\right)}{1-\exp\left(-\frac{t}{\mu^{\text{old}}}\right)}\right)\right)$$

3. Let B be a positive real valued random variable with probability density

$$p_B(b) = e^{-b}, \quad \text{for all } b > 0.$$

Also A is a real valued random variable with conditional density

$$p_{A|B}(a|b) = \sqrt{\frac{b}{\pi}}e^{-a^2 b}, \quad \text{for all } a \in (-\infty, \infty) \text{ and } b \in (0, \infty).$$

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

$$p_A(a) = \int_0^\infty p_B(b) p_{A|B}(a|b)\, db = \int_0^\infty \sqrt{\frac{b}{\pi}} e^{-b(a^2+1)}\, db$$

b. Approximate $p_A(a)$ using the Laplace approximation.
Give the detailed derivation, not just the answer.

First we define for notational efficiency

$$f_a(b) = \sqrt{\frac{b}{\pi}} e^{-b(a^2+1)}.$$

In order to find the maximum we take the first derivative w.r.t. b.

$$\begin{aligned}\frac{\partial}{\partial b} f_a &= \frac{1}{2\pi}\sqrt{\frac{\pi}{b}} e^{-b(a^2+1)} - \sqrt{\frac{b}{\pi}}(a^2+1)e^{-b(a^2+1)} \\ &= e^{-b(a^2+1)}\left(\frac{1}{2}\sqrt{\frac{1}{\pi b}} - (a^2+1)\sqrt{\frac{b}{\pi}}\right)\end{aligned}$$

Solving for zero we get

$$\begin{aligned}\frac{1}{2}\sqrt{\frac{1}{\pi b}} &= (a^2+1)\sqrt{\frac{b}{\pi}} \\ \sqrt{\frac{1}{b^2}} &= \frac{1}{b} = 2(a^2+1) \\ b_{\text{opt}} &= \frac{1}{2(a^2+1)}\end{aligned}$$

For the Laplace approximation we need the (negative of the) second derivative w.r.t. b of $\ln f_a(b)$, evaluated in b_{opt}.

$$\begin{aligned}g_a(b) = \ln f_a(b) &= -b(a^2+1) + \frac{1}{2}\ln\frac{b}{\pi}. \\ \frac{\partial}{\partial b} g_a(b) &= -(a^2+1) + \frac{1}{2}\frac{\pi}{b}\frac{1}{\pi} = -a^2 - 1 + \frac{1}{2b} \\ \frac{\partial^2}{\partial b^2} g_a(b) &= -\frac{1}{2b^2} \\ A_{\text{Laplace}} &= \frac{1}{2b_{\text{opt}}^2} = 2(a^2+1)^2.\end{aligned}$$

So we find

$$\begin{aligned} p_A(a) &= \int_0^\infty f_a(b)\,db \\ &\approx f_a(b_{\text{opt}})\sqrt{\frac{2\pi}{A_{\text{Laplace}}}} \\ &= \sqrt{\frac{1}{2e}}\frac{1}{(a^2+1)^{\frac{3}{2}}} \\ &= 1.16582\frac{1}{(a^2+1)^{\frac{3}{2}}} \end{aligned}$$

Compare this to the actual value !

$$\begin{aligned} p_A(a) &= \int_0^\infty f_a(b)\,db \\ &= \frac{1}{2}\frac{1}{(a^2+1)^{\frac{3}{2}}} \end{aligned}$$

4. The *Bayesian Information Criterion* is results in

$$\underbrace{\log\frac{p(\mathcal{M}_1|x^N)}{p(\mathcal{M}_2|x^N)}}_{(*1)} \approx \underbrace{\log\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)}}_{(*2)} + \underbrace{\log\frac{p(x^N|\mathcal{M}_1,\underline{\hat{\theta}}_1)}{p(x^N|\mathcal{M}_2,\underline{\hat{\theta}}_2)}}_{(*3)} + \underbrace{\frac{1}{2}(k_1-k_2)\log N}_{(*4)}.$$

Here x^N is a binary data sequence of length N, k_1 and k_2 are the number of free parameters in respectively model $\mathcal{M}_1$ and $\mathcal{M}_2$, and $\underline{\hat{\theta}}_1$ and $\underline{\hat{\theta}}_2$ are the estimated (ML) parameter vectors.

a. Explain the four terms marked by (*1), (*2), (*3), and (*4).

(*1) This ratio of model posteriors (given the data) allows us to select the most appropriate model of the two options.

(*2) This ratio shows our initial preference of the first model relative to the second one.

(*3) This is the log-likelihood ratio of the two models after observing the data.

(*4) This is the correction term needed to compare models of different complexity.

b. The binary data $x^N = x_1, x_2, \ldots, x_N$ is generated by a Bernoulli process, i.e.

$$p(x^N|\mathcal{M},\theta) = (1-\theta)^{n(0|x^N)}\theta^{n(1|x^N)}.$$

The parameter prior $p(\theta|\mathcal{M})$ is given by the Beta distribution:

$$p(\theta|\mathcal{M}) = \frac{1}{\pi}\frac{1}{\sqrt{\theta(1-\theta)}}.$$

Let $N = 10$ and $x^{10} = 1001101101$. Determine $p(x^N|\mathcal{M})$. Give the complete derivation starting with the information given above.

$$
\begin{aligned}
p(x^N|\mathcal{M}) &= \int_0^1 p(\theta|\mathcal{M})p(x^N|\mathcal{M},\theta)\,d\theta, \\
&= \int_0^1 \frac{1}{\pi}\frac{1}{\sqrt{\theta(1-\theta)}}(1-\theta)^4\theta^6\,d\theta, \\
&= \frac{\Gamma(4+\frac{1}{2})\Gamma(6+\frac{1}{2})}{\pi\Gamma(11)}, \\
&= \frac{\frac{1}{2}\frac{3}{2}\frac{5}{2}\frac{7}{2}\cdot\frac{1}{2}\frac{3}{2}\frac{5}{2}\frac{7}{2}\frac{9}{2}\frac{11}{2}}{10!}, \\
&= \frac{77}{262144} = 0.0002937.
\end{aligned}
$$

c. Why do we think that the probability estimate $p(x^N|\mathcal{M})$ is a useful and good estimate for the actual, but unknown, probability $p(x^N|\mathcal{M},\theta)$? And how close will $p(x^N|\mathcal{M})$ be to the probability $p(x^N|\mathcal{M},\theta)$, for any x^N and any θ, if $\mathcal{M}$ is a binary memoryless model?

In the lecture noted it is shown that for the memoryless binary model, with any parameter value θ and any sequence x^N holds

$$\log\frac{p(x^N|\mathcal{M},\theta)}{p(x^N|\mathcal{M})} \le \frac{1}{2}\log N + 1.$$

And with the *Capacity-Redundancy theorem* we know that

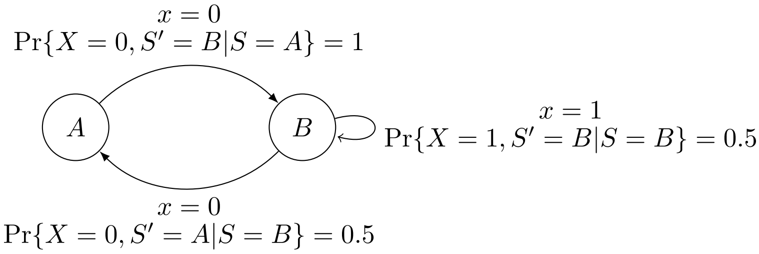
$$\log\frac{p(x^N|\mathcal{M},\theta)}{p(x^N|\mathcal{M})} \ge \frac{1}{2}\log N - \epsilon_N.$$

Here $\epsilon_N \to 0$ as $N \to \infty$.
So, in this sense, the probability estimate is optimal.

5. Consider the following binary finite state model (Markov source). This model produces outputs X_t where the probability of the next output symbol depends on the current state of the source. We list all non-zero probabilities.

$$
\begin{aligned}
\Pr\{X_t = 0, S_{t+1} = B|S_t = A\} &= 1, \\
\Pr\{X_t = 0, S_{t+1} = A|S_t = B\} &= 0.5, \\
\Pr\{X_t = 1, S_{t+1} = B|S_t = B\} &= 0.5,
\end{aligned}
$$

The following figure depicts this model.



a. Compute the stationary probabilities $q(A)$ and $q(B)$ where

$$q(s) = \lim_{t\to\infty}\Pr\{S_t = s\} \qquad \text{for } s \in \{A, B\}.$$

$$q(A) = \frac{1}{2}q(B),$$
$$q(B) = 1 - q(A)$$

This results in $q(A) = \frac{1}{3}$ and $q(B) = \frac{2}{3}$.

b. Compute the following probabilities assuming that the model is stationary (i.e. $\Pr\{S_1 = A\} = q(A)$ and $\Pr\{S_1 = B\} = q(B)$).

$$\Pr\{X_1 = 1\}$$
$$\Pr\{X_2 = 1|X_1 = 0\}$$

$$\Pr\{X_1 = 1\} = q(A) \cdot 0 + q(B) \cdot \frac{1}{2} = \frac{1}{3}.$$

For the next one we need

$$\Pr\{X^2 = 01\} = q(A) \cdot 1 \cdot \frac{1}{2} + q(B) \cdot \frac{1}{2} \cdot 0 = \frac{1}{6},$$

and we find

$$\Pr\{X_2 = 1|X_1 = 0\} = \frac{\Pr\{X^2 = 01\}}{\Pr\{X_1 = 0\}} = \frac{1/6}{2/3} = \frac{1}{4}.$$

c. Let $\mathcal{M}_0$ be the i.i.d. model with

$$\underline{\theta}_0 = (\Pr\{X_1 = 1\}).$$

Also $\mathcal{M}_1$ is the first order model with

$$\underline{\theta}_1 = (\theta_{10}, \theta_{11}) = (\Pr\{X_2 = 1|X_1 = 0\}, \Pr\{X_2 = 1|X_1 = 1\}).$$

And $\mathcal{M}_2$ is the second order model with

$$\begin{aligned}\underline{\theta}_2 &= (\theta_{200}, \theta_{201}, \theta_{210}, \theta_{211}) \\ &= (\Pr\{X_3 = 1|X_1 = 0, X_2 = 0\}, \Pr\{X_3 = 1|X_1 = 0, X_2 = 1\}, \\ &\qquad \Pr\{X_3 = 1|X_1 = 1, X_2 = 0\}, \Pr\{X_3 = 1|X_1 = 1, X_2 = 1\}).\end{aligned}$$

The Markov model produces a 'typical' sequence so

$$-\log_2 \Pr\{X^n = x^n|\mathcal{M}_i, \underline{\theta}_i\} \approx H_i(X^n),$$

where $H_i(X^n)$ is the entropy rate of the i^{th} model.
Given is that

$$H_0(X^n) = 0.9183 \cdot n$$
$$H_1(X^n) = 0.8742 \cdot n$$
$$H_2(X^n) = 0.7925 \cdot n$$

Determine for what range of n you should use $\mathcal{M}_0$. And when $\mathcal{M}_1$ and when $\mathcal{M}_2$? Use the idea of stochastic complexity and **motivate your answer**.

The remaining conditional probabilities should be computed in order to find the entropies.

Then we find the stochastic complexities

$$\begin{aligned} S.C._0 &= \frac{\log_2 n}{2} + 0.9183 \cdot n \\ S.C._1 &= \frac{2\log_2 n}{2} + 0.8742 \cdot n \\ S.C._2 &= \frac{4\log_2 n}{2} + 0.7925 \cdot n \end{aligned}$$

Equality of $S.C._0$ and $S.C._1$ happens at $n = 69 \sim 70$.
Equality of $S.C._1$ and $S.C._2$ happens at $n = 76 \sim 77$.

So up to $n = 70$ we use $\mathcal{M}_0$, then until $n = 77$ we use $\mathcal{M}_1$ and afterwards we use $\mathcal{M}_2$.

Points that can be scored per question:

Question 1:	each sub-question a through e: 2 points. Total 10 points.
Question 2:	a) 2 points; b) 1 point; c) 2 points; d) 2 points; e) 1 point; f) 1 point; g) 1 point. Total 10 points.
Question 3:	a) 1 point; b) 5 points. Total 6 points.
Question 4:	a) 4 points; b) 2 points; c) 2 points. Total 8 points.
Question 5:	a) 1 point; b) 2 points; c) 3 points. Total 6 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.