Name :

Study program :

ID. NR. :

1.

a. Consider a data set $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ where we assume that each sample $\mathbf{x}_n$ is IID distributed by a multivariate Gaussian (MVG), $\mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}, \Sigma)$. Proof that the maximum likelihood estimate (MLE) of the mean value of this distribution is given by

$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_n \mathbf{x}_n \tag{1}$$

b. Consider now a data set $\mathcal{D} = \{(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_N, t_N)\}$ with 1-of-K notation for the discrete classes, i.e.,

$$t_{nk} = \begin{cases} 1 & \text{if } t_n \text{ in class } \mathcal{C}_k \\ 0 & \text{else} \end{cases}$$

together with class-conditional distribution $p(\mathbf{x}|\mathcal{C}_k, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ and multinomial prior $p(\mathcal{C}_k|\boldsymbol{\pi}) = \pi_k$.
Proof that the joint log-likelihood is given by

$$\log p(\mathcal{D}|\boldsymbol{\theta}) = \sum_{n,k} t_{nk} \log \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}) + \sum_{n,k} t_{nk} \log \pi_k$$

c. Show now that the MLE of the *class-conditional* mean is given by

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_n t_{nk} \mathbf{x}_n}{\sum_n t_{nk}} \tag{2}$$

d. Explain this formula (eqn 2) in relation to eqn 1, the MLE for the mean of a MVG.

e. In the lecture notes, we also discussed the MLE for a *clustering* problem and derived (for the i-th iteration of the EM algorithm):

$$\hat{\boldsymbol{\mu}}_k^{(i)} = \frac{\sum_n \gamma_{nk}^{(i)} \mathbf{x}_n}{\sum_n \gamma_{nk}^{(i)}} \tag{3}$$

(i) What does $\gamma_{nk}^{(i)}$ represent?
(ii) Express $\gamma_{nk}^{(i)}$ in terms of z_{nk} and $\mathbf{x}_n$
(iii) Why the iterative EM algorithm?

2. Consider an IID data set $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$. We will model this data set by a model $y_n = \theta^T f(x_n) + e_n$, where $f(x_n)$ is an M-dimensional feature vector of input x_n; y_n is a scalar output and $e_n \sim \mathcal{N}(0, \sigma^2)$. (Note the list of formula's at the final page of this exam).

a. Rewrite the model in matrix form by lumping input features in a matrix $F = [f(x_1), \ldots, f(x_N)]^T$, outputs and noise in the vectors $y = [y_1, \ldots, y_N]^T$ and $e = [e_1, \ldots, e_N]^T$, respectively.

b. Now derive an expression for the log-likelihood $\log p(y| F, \theta, \sigma^2)$.

c. Proof that the maximum likelihood estimate for the parameters is given by

$$\hat{\theta}_{ml} = (F^T F)^{-1} F^T y$$

d. What is the predicted output value y_{new}, given an observation x_{new} and the maximum likelihood parameters $\hat{\theta}_{ml}$. Work this expression out in terms of F, y and $f(x_{\text{new}})$.

e. Suppose that, before the data set D was observed, we had reason to assume a prior distribution $p(\theta) = \mathcal{N}(0, \sigma_0^2)$. Derive the Maximum a posteriori (MAP) estimate $\hat{\theta}_{map}$.(hint: work this out in the log domain.)

3.

a. (a) Why is Principal Components Analysis more popular than Factor Analysis in signal and image processing applications?
(b) What is the difference between supervised and unsupervised learning?
Mark the following two statements with a TRUE or FALSE flag.

(c) If X and Y are independent Gaussian distributed variables, then $Z = 3X + Y$ is also a Gaussian distributed variable.
(d) The sum of two Gaussian functions is always also a Gaussian function.

4. Consider a sequence x^n generated by an exponential model $\mathcal{M}$ with an unknown parameter μ. So

$$p(x|\mathcal{M}, \mu) = \frac{1}{\mu} e^{-x/\mu}, \quad \text{for a single symbol } x,$$

and thus

$$p(x^n|\mathcal{M}, \mu) = \frac{1}{\mu^n} \prod_{i=1}^{n} e^{-x_i/\mu}, \quad \text{for a sequence } x^n.$$

a. Derive an expression for the log-likelihood $\ell(\mu) \equiv \log p(x^n|\mathcal{M}, \mu)$ as a function of the average value $\bar{x} = (1/n) \sum_{i=1}^{n} x_i$.

b. What is the maximum likelihood estimate, $\hat{\mu}_{ML}$, for μ based on observations x^n?

c. Let your observations be

$$\begin{aligned} x^{15} = (&0.7578, 0.2808, 3.4246, 0.1069, 0.6905, 0.9240, 0.2466, 0.7749, \\ &3.1880, 0.5657, 0.6044, 2.2380, 1.8625, 0.2467, 2.6036). \end{aligned} \tag{4}$$

So $\sum_{i=1}^{15} x_i = 18.5161$. Determine the ML estimate $\hat{\mu}_{ML}$ in this case.

d. Also, derive an expression for the ML sequence probability $p(x^{15}|\mathcal{M}, \hat{\mu}_{ML})$ for x^{15} as given in equation 4.

e. You are now given a *prior* over μ, namely

$$p(\mu|\mathcal{M}) = \begin{cases} 1; \text{if } \mu \in [1, 2], \\ 0; \text{if } \mu \notin [1, 2]. \end{cases}$$

Derive the Laplace approximation of

$$p(x^{15}|\mathcal{M}) = \int_0^\infty p(\mu|\mathcal{M}) p(x^{15}|\mathcal{M}, \mu)\, d\mu,$$

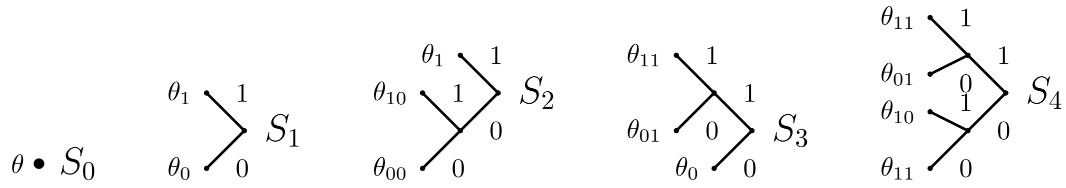
where x^{15} is again as given in equation 4.

5. We observe a binary sequence $x^{15} = 110\,111\,101\,100\,001$. (The spacing is just for ease of reading and has no other meaning.) Assume that this sequence is preceded by two zeros as the initial context.

Consider a context model $\mathcal{S}$ of depth 2 and the "CTW prior" $P(S_i)$ given as:

$$\begin{aligned} \Delta_2(S) &= 2|S| - 1 - |\{s \in S : |s| = 2\}|, \\ P(S_i) &= 2^{-\Delta_2(S_i)} \end{aligned}$$

The following five tree structures are possible models, S_0, S_1, S_2, S_3, and S_4:



a. Compute the probability of this sequence in the recursive "CTW" manner. So, compute recursively

$$P_w^\lambda(x^{15}) = \sum_{i=0}^{4} P(S_i)P(x^{15}|S_i).$$

b. Determine the a-posteriori model probabilities for these five models.

Appendix: formula's

$$\begin{aligned}
|A^{-1}| &= |A|^{-1} \\
\nabla_A \log|A| &= (A^T)^{-1} = (A^{-1})^T \\
\mathrm{Tr}[ABC] &= \mathrm{Tr}[CAB] = \mathrm{Tr}[BCA] \\
\nabla_A \mathrm{Tr}[AB] &= \nabla_A \mathrm{Tr}[BA] = B^T \\
\nabla_A \mathrm{Tr}[ABA^T] &= A(B + B^T) \\
\nabla_x x^T A x &= (A + A^T)x \\
\nabla_X a^T X b &= \nabla_X \mathrm{Tr}[ba^T X] = ab^T
\end{aligned}$$

Multivariate gaussian

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

Points that can be scored per question:

Question 1: a) 2 points; b) 2 points; c) 1 point; d) 1 point; e) 3 points (total: 9).

Question 2: a) 1 point; b) 2 point; c) 1 point; d) 1 point; e) 2 points (total: 7).

Question 3: each sub-question a through d: 1 point (total: 4).

Question 4: a) 3 points; b) 2 points; c) 1 point; d) 1 point; e) 3 points. Total 10 points.

Question 5: a) 5 points; b) 5 points. Total 10 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.