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

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

$$\begin{aligned}
\nabla_\mu \log p(\mathcal{D}|\boldsymbol{\theta}) &= -\frac{1}{2}\sum_n \nabla_\mu \left[(\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})\right] \\
&= -\frac{1}{2}\sum_n \nabla_\mu \mathrm{Tr}\left[-2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}_n + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right] \\
&= -\frac{1}{2}\sum_n \left(-2\boldsymbol{\Sigma}^{-1}\mathbf{x}_n + 2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \\
&= \boldsymbol{\Sigma}^{-1}\sum_n (\mathbf{x}_n - \boldsymbol{\mu})
\end{aligned}$$

Set to zero yields

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

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

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

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

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

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

see lecture notes.

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

> Eqn 2 computes the sample proportion, just like eqn 1 , but now only for samples from class k.

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?

> (1) The responsibilty $\gamma_{nk}^{(i)} = \mathbb{E}[z_{nk}|\mathbf{x}_n, \boldsymbol{\theta}^{(i-1)}]$ is a *soft* class indicator.
> (2) It is our best estimate of the binary class indicator t_{nk}, given the input $\mathbf{x}_n$.
> (3) We need the iterative EM algorithm because in clustering we don't have a one-step solution to the maximum likelihood estimation problem.

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.

> $y = F\theta + e$

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

> $$\begin{aligned} \log p(D|\theta, \sigma^2) &= \log \mathcal{N}(y|\, F\theta, \sigma^2) \\ &\propto -\frac{1}{2\sigma^2} (y - F\theta)^T (y - F\theta) \end{aligned}$$

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

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

> Taking the derivative to θ
>
> $$\nabla_\theta \log p(D|\theta) = \frac{1}{\sigma^2} F^T (y - F\theta)$$
>
> Set derivative to zero for maximum likelihood estimate
>
> $$\hat{\theta} = (F^T F)^{-1} F^T y$$

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 $f(x_{\text{new}})$.

> Prediction of new data point: $\hat{y}_{\text{new}} = \hat{\theta}^T f(x_{\text{new}}) = [(F^T F)^{-1} F^T y]^T f(x_{\text{new}})$

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.)

$$\begin{aligned}\log p(\theta|D) &\propto \log p(D|\theta)p(\theta) \\ &\propto -\frac{1}{2\sigma^2}\left(y - F\theta\right)^T\left(y - F\theta\right) + \frac{1}{2\sigma_0^2}\theta^T\theta\end{aligned}$$

Derivative $\nabla_\theta \log p(\theta|D) = -\frac{1}{\sigma^2}F^T(y - F\theta) + (1/\sigma_0^2)\theta$
Set derivative to zero for MAP estimate leads to

$$\hat{\theta}_{MAP} = (F^T F + \frac{\sigma^2}{\sigma_0^2}I)^{-1}F^T y$$

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?

(Alternative answers may also be accepted)
(a) In signal and image processing, the components of a vector are often shifted (delayed) samples. In that case the noise variances are not affected, which is modeled correctly by PCA.
(b) In supervised learning concerns learning a map from inputs to targets. Unsupervised learning concerns analysis of data without targets, such as pattern discovery and compression.

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.

(c) T, (d) F

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 $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$.

$$\begin{aligned}\ell(\mu) &= \log \prod_i p(x_i|\mathcal{M},\mu) = \sum_i \log\left(\frac{1}{\mu}e^{-x_i/\mu}\right)\\ &= -n\log\mu - (1/\mu)\sum_i x_i\\ &= -n\left(\log\mu + \bar{x}/\mu\right)\end{aligned}$$

where $\bar{x} = (1/n)\sum_{i=1}^{N} x_i$.

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

Set $\frac{\partial \ell}{\partial \mu} = -n\left(\frac{1}{\mu} - \bar{x}/\mu^2\right)$ to zero to get $\hat{\mu}_{ML} = \bar{x}$ (the sample mean).

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.

In this case $n = 15$ and $\sum_{i=1}^{15} x_i = 18.5161$, so

$$\hat{\mu}_{ML} = 1.2344.$$

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.

We denote $\sum_{i=1}^{15} x_i = 18.5161$ by $\mathcal{X}$.
Straightforward, just plug-in $\hat{\mu}_{ML}$ into the given expression, so

$$\begin{aligned}p(x^n|\mathcal{M},\hat{\mu}_{ML}) &= \frac{1}{\hat{\mu}_{ML}^n}\prod_{i=1}^{n} e^{-x_i/\hat{\mu}_{ML}}\\ &= \left(\frac{n}{\mathcal{X}}\right)^n e^{-n}.\end{aligned}$$

Actually, a numerical answer is also fine.

$$= 1.2993 \cdot 10^{-8}$$

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.

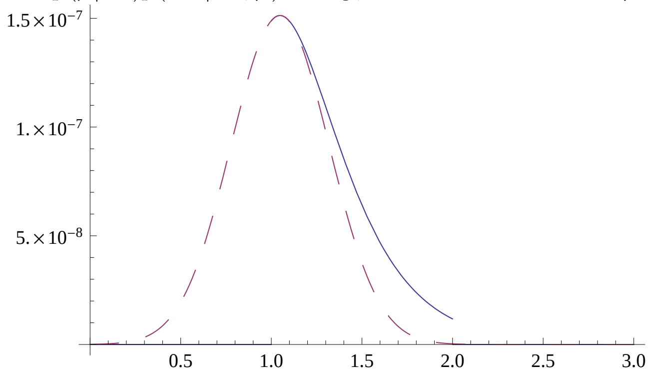
Consider
$$f(\mu) = p(x^n|\mathcal{M}, \mu).$$
We know from (4) that it has a maximum at
$$\hat{\mu}_{ML} = 1.2344$$
which lies in the valid range for $p(\mu|\mathcal{M})$. So we must evaluate the second derivative of $\ln f(\mu)$ in $\hat{\mu}_{ML}$.

$$\begin{aligned}
\ln f(\mu) &= -15 \ln \mu - \frac{\mathcal{X}}{\mu}.\\
\frac{\partial \ln f}{\partial \mu} &= \frac{-15}{\mu} + \frac{\mathcal{X}}{\mu^2}.\\
\frac{\partial^2 \ln f}{\partial \mu^2} &= \frac{15}{\mu^2} - \frac{2\mathcal{X}}{\mu^3}\\
\frac{\partial^2 \ln f(\hat{\mu}_{ML})}{\partial \mu^2} &= -\frac{15^3}{\mathcal{X}^2} < 0.
\end{aligned}$$

Then the Gaussian approximation is
$$\begin{aligned}
g(\mu) &= \left(\frac{n}{\mathcal{X}}\right)^n e^{-n} \exp\left(-\frac{1}{2}(\mu - \frac{\mathcal{X}}{n})^2 \frac{n^3}{\mathcal{X}^2}\right)\\
&= \left(\frac{n}{\mathcal{X}}\right)^n e^{-n(1+(\frac{n}{\mathcal{X}}\mu-1)^2))}
\end{aligned}$$

The integral of $g(\mu)$ gives
$$\begin{aligned}
\int g(\mu)\, d\mu &= \sqrt{\frac{2\pi}{n^3}}\mathcal{X}\left(\frac{n}{\mathcal{X}}\right)^n e^{-n}\\
&= 1.0380 \cdot 10^{-8} \text{ for the given sequence.}
\end{aligned}$$

We show a plot of $p(\mu|\mathcal{M})p(x^{15}|\mathcal{M}, \mu)$ and g, both as a function of μ.

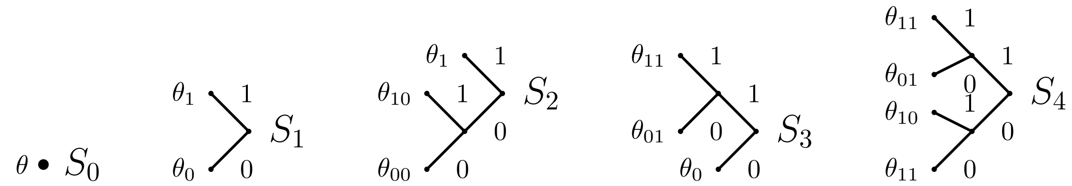


5. We observe a binary sequence $x^{15} = 110\,111\,101\,100\,001$. (The spacing is just for ease of reading and has no other meaning.) Assume that this sequence is preceded by two zeros as the initial context.

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

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

We consider a binary context tree of depth 2. We must collect the number of zeros and ones in eacht of the contexts $\lambda, 0, 1, 00, 10, 01,$ and 11. We find

s	$N(s0\|x^{15}) = a$	$N(s1\|x^{15}) = b$
λ	6	9
0	3	4
1	3	5
00	2	2
10	1	2
01	0	3
11	3	2

From this we calculate the $P_e(a,b)$'s and the P_w^s's.

s	$P_e(a,b)$	P_w^s
00	$2.3438 \cdot 10^{-2}$	$2.3438 \cdot 10^{-2}$
10	$6.25 \cdot 10^{-2}$	$6.25 \cdot 10^{-2}$
01	$3.125 \cdot 10^{-1}$	$3.125 \cdot 10^{-1}$
11	$1.1719 \cdot 10^{-2}$	$1.1719 \cdot 10^{-2}$
0	$2.4414 \cdot 10^{-3}$	$1.9531 \cdot 10^{-3}$
1	$1.3733 \cdot 10^{-3}$	$2.5177 \cdot 10^{-3}$
λ	$8.3596 \cdot 10^{-6}$	$6.6347 \cdot 10^{-6}$

The requested probability is $p_w^\lambda = 6.6347 \cdot 10^{-6}$.

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

For every model we must compute $2^{-\Delta_2(S)} \prod_{s \in S} P_e(a_s, b_s)/P_w^\lambda$.

S	$2^{-\Delta_2(S)}$	$P(x^{15}\|S)$	$P(S\|x^{15})$
S_0	2^{-1}	$8.3596 \cdot 10^{-6}$	0.630
S_1	2^{-3}	$3.3528 \cdot 10^{-6}$	0.063
S_2	2^{-3}	$2.0117 \cdot 10^{-6}$	0.038
S_3	2^{-3}	$8.9407 \cdot 10^{-6}$	0.168
S_4	2^{-3}	$5.3644 \cdot 10^{-6}$	0.101

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.