IMPERIAL COLLEGE LONDON

TIMED REMOTE ASSESSMENTS 2020-2021

MEng Honours Degree in Electronic and Information Engineering Part IV
MEng Honours Degree in Mathematics and Computer Science Part IV
MEng Honours Degrees in Computing Part IV
MSc Advanced Computing
MSc Artificial Intelligence
MSc in Computing (Specialism)
for Internal Students of the Imperial College of Science, Technology and Medicine

This paper is also taken for the relevant assessments for the Associateship of the City and Guilds of London Institute

PAPER COMP70015=COMP97065=COMP97066

MATHEMATICS FOR MACHINE LEARNING

Tuesday 8 December 2020, 10:00
Duration: 105 minutes
Includes 15 minutes for access and submission

Answer ALL TWO questions
Open book assessment

By completing and submitting work for this assessment, candidates confirm that the submitted work is entirely their own and they have not (i) used the services of any agency or person(s) providing specimen, model or ghostwritten work in the preparation of the work they have submitted for this assessment, (ii) given assistance in accessing this paper or in providing specimen, model or ghostwritten answers to other candidates submitting work for this assessment.

Paper contains 2 questions

Always provide justifications and show any intermediate work for your answers. A correct but unsupported answer may not receive any marks.

You may find the following useful:

- Bernoulli distribution

$$p(x|\mu) = \mu^x(1-\mu)^{1-x}\,, \quad x \in \{0,1\}$$

- Binomial distribution

$$p(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}$$

- Beta distribution

$$\text{Beta}(\mu|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1}$$

- Gamma distribution

$$\text{Gamma}(\tau|a,b) = \frac{1}{\Gamma(a)}b^a\tau^{a-1}\exp(-b\tau)$$

- Gaussian distribution

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\big(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\big)$$

- Wishart distribution

$$\mathcal{W}(\boldsymbol{\Sigma}|\boldsymbol{W},\nu) = B|\boldsymbol{\Sigma}|^{\frac{\nu-D-1}{2}}\exp\big(-\tfrac{1}{2}\text{tr}(\boldsymbol{W}^{-1}\boldsymbol{\Sigma})\big)$$

- Gaussian conditioning. For a joint Gaussian density

$$p\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix};\begin{bmatrix}\mathbf{m_x}\\ \mathbf{m_y}\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{xy}}\\ \boldsymbol{\Sigma}_{\mathbf{yx}} & \boldsymbol{\Sigma}_{\mathbf{yy}}\end{bmatrix}\right), \tag{1}$$

we have the conditional density

$$p(\mathbf{x}\,|\,\mathbf{y}) = \mathcal{N}\Big(\mathbf{x};\quad \mathbf{m_x}+\boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}(\mathbf{y}-\mathbf{m_y}),\quad \boldsymbol{\Sigma}_{\mathbf{xx}}-\boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\boldsymbol{\Sigma}_{\mathbf{yx}}\Big)\,. \tag{2}$$

1 a Consider a test for disease. If the disease is present ($D = 1$), then we have a 99% chance of the test being positive ($T = 1$), i.e. $P(T = 1|D = 1) = 0.99$. Similarly, if the disease is not present ($D = 0$), then we have a 99% chance of the test being negative ($T = 0$), i.e. $P(T = 0|D = 0) = 0.99$. A random person has a 1% chance of having the disease, i.e. $P(D = 1) = 0.01$. What is the probability for a random person to have the disease if they test positive? I.e. what is $P(D = 1|T = 1)$?

b Consider a sequence of random variables defined as

$$x_{t+1} = -kx_t + \epsilon_t\,, \qquad \epsilon_t \sim \mathcal{N}\left(0, \sigma^2\right)\,, \tag{3}$$

with ϵ_t being independent from all other variables.

If you assume any identities on expectations or variances, clearly state them. Only use one identity per step.

i) In terms of the variance of x_t (denoted v_t), find the variance of x_{t+1}.

ii) Given that $x_0 = 0$, find the joint density $p(x_1, x_2, x_3)$. State **1)** the form of the density and why it has this density, and **2)** any parameters that are needed to specify it.

iii) Find the conditional density $p(x_3|x_1)$.

c A linear autoencoder tries to compress images $\mathbf{x} \in \mathbb{R}^D$ into a lower-dimensional code $\mathbf{z} \in \mathbb{R}^L$ (i.e. $L < D$) with a linear map $\boldsymbol{B} \in \mathbb{R}^{L\times D}$. A decoder $\boldsymbol{C} \in \mathbb{R}^{D\times L}$ maps back to the image space to get a reconstruction $\hat{\mathbf{x}}$. Both $\boldsymbol{B}$ and $\boldsymbol{C}$ are chosen by minimising the *reconstruction loss*:

$$L(\boldsymbol{B}, \boldsymbol{C}) = \sum_{n=1}^{N} ||\mathbf{x} - \hat{\mathbf{x}}||^2\,, \qquad \hat{\mathbf{x}} = \boldsymbol{C}\mathbf{z}\,, \qquad \mathbf{z} = \boldsymbol{B}\mathbf{x}\,. \tag{4}$$

i) Using a consistent defintion of the multivariate chain rule:

A) represent the derivatives $\frac{\partial L}{\partial \boldsymbol{C}}$ and $\frac{\partial L}{\partial \boldsymbol{B}}$ in terms of the intermediate derivatives on the variables $\hat{\mathbf{x}}$ and $\mathbf{z}$,

B) give the shape of each intermediate gradient (4 total),

C) Give the scalar summation form of each possible application of the multivariate chain rule in this derivative (3 total).

ii) Find the explicit form of all gradients needed in the chain rules (4 total). You are free to use indexed scalar notation, or vector notation. Choose whatever you can most clearly describe the answer with.

The three parts carry, respectively, 10%, 45%, and 45% of the marks.

2 Component Analysis & SVMs

a Assume you are given a set of n data samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$. The minimum enclosing hyper-sphere in a space defined by a positive definite kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$ on the above data can be found by the following constrained optimisation problem

$$\begin{array}{ll} \min_{R,\mathbf{a},\xi_i} & R^2 + C\sum_{i=1}^n \xi_i \\ \text{subject to} & (\phi(\mathbf{x}_i) - \mathbf{a})^T(\phi(\mathbf{x}_i) - \mathbf{a}) \leq R^2 + \xi_i, \forall \xi_i \geq 0, \ \ \forall i \in \{1, \ldots, n\} \end{array}$$

where R is the radius of the hyper-sphere, $\mathbf{a}$ is the center of the hyper-sphere and the variable C gives the trade-of between simplicity (or volume of the sphere) and the number of errors. Formulate the Lagrangian of the above optimisation problem and derive its dual. Compute $\mathbf{a}$ and R.

Assume the minimum enclosing hyper-ellipse optimisation problem below

$$\begin{array}{ll} \min_{R,\mathbf{a},\xi_i} & R^2 + C\sum_{i=1}^n \xi_i \\ \text{subject to} & (\mathbf{x}_i - \mathbf{a})^T \mathbf{S}_t^{-1}(\mathbf{x}_i - \mathbf{a}) \leq R^2 + \xi_i, \forall \xi_i \geq 0. \end{array}$$

where $\mathbf{S}_t$ is the covariance matrix of the data which is assumed to be invertible.

Find an appropriate kernel in order to transform the above problem into a special case of kernel minimum enclosing hyper-sphere problem. How can we solve the problem in case that $\mathbf{S}_t$ is singular?

b Assume data samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ split in C different classes. Assume you are given a positive definite kernel k which defines an implicit Hilbert space on the vectors $\phi(\mathbf{x}_i) \in \mathcal{H}$. Assume a different Kernel Discriminant Analysis (KDA) optimisation problem as

$$\begin{array}{c} \max_{\mathbf{W}_\Phi} \ \mathrm{tr}(\mathbf{W}_\Phi^T \mathbf{S}_b^\Phi \mathbf{W}_\Phi) - \mathrm{tr}(\mathbf{W}_\Phi^T \mathbf{S}_w^\Phi \mathbf{W}_\Phi) \\ \text{subject to } \ \mathbf{W}_\Phi^T \mathbf{W}_\Phi = \mathbf{I}. \end{array} \tag{2}$$

where $\mathbf{S}_b^\Phi$ and $\mathbf{S}_w^\Phi$ are the between and within class scatter matrices defined in the Hilbert space of the vectors $\phi(\mathbf{x}_i) \in \mathcal{H}$. Find the solution to the optimisation problem. Then, write how given a test sample $\mathbf{x}_t$ using the above solution you would extract features.

The two parts carry equal marks.