EGT3
ENGINEERING TRIPOS PART IIB

Wednesday 26 April 2023   09.30 to 11.10

Module 4M24

COMPUTATIONAL STATISTICS AND MACHINE LEARNING

Answer not more than ***three*** *questions.*

All questions carry the same number of marks.

The ***approximate*** *percentage of marks allocated to each part of a question is indicated in the right margin.*

Write your candidate number ***not*** *your name on the cover sheet.*

STATIONERY REQUIREMENTS
Write on single-sided paper.

SPECIAL REQUIREMENTS TO BE SUPPLIED FOR THIS EXAM
CUED approved calculator allowed.
Engineering Data Books.

10 minutes reading time is allowed for this paper at the start of the exam.

You may not start to read the questions printed on the subsequent pages of this question paper until instructed to do so.

You may not remove any stationery from the Examination Room.

1 The expectation of a function with respect to a probability measure takes the integral form

$$\mathcal{I} = \mathbb{E}\{f(X)\} = \int_a^b f(x)p(x)dx$$

where $X \in \mathbb{R}$ is a univariate random variable with probability density $p(\cdot)$, and dx denotes Lebesgue measure on $\mathbb{R}$. An estimate $\hat{\mathcal{I}}$ of the integral $\mathcal{I}$ can be obtained with a generalised Monte Carlo scheme such that

$$\hat{\mathcal{I}} = \sum_{n=1}^{N} w_n f(x_n)$$

where $x_1, \cdots, x_N$ are i.i.d from $p(\cdot)$, and $w_n \in \mathbb{R}$ for all $n = 1, \cdots, N$.

(a) We denote the $N \times 1$ vector $\mathbf{w} = [w_1, \cdots, w_N]^\mathsf{T}$ and the $N \times 1$ vector of ones as $\mathbf{1} = [1, \cdots, 1]^\mathsf{T}$.

(i) By defining the elements of the $N \times N$ dimensional matrix $\mathbf{C}$ show that the expected squared deviation of $\hat{\mathcal{I}}$ around the true value of $\mathcal{I}$, $\mathbb{E}\{(\hat{\mathcal{I}} - \mathcal{I})^2\}$, takes the form $\mathcal{I}^2 \left[\mathbf{w}^\mathsf{T}\mathbf{C}\mathbf{w} - 2\mathbf{w}^\mathsf{T}\mathbf{1} + 1\right]$. [30%]

(ii) Derive an expression for the value of $\mathbf{w}$ yielding the expected minimum deviation estimate of $\mathcal{I}$. [10%]

(b) By noting that the matrix $\mathbf{C}$ can be written in the form $\mathbf{C} = \epsilon\mathbf{I} + \mathbf{1}\mathbf{1}^\mathsf{T}$ where $\mathbf{I}$ is the $N \times N$ identity matrix and $\epsilon \in \mathbb{R}$ is a scalar, use the Sherman-Morrison formula

$$\left(\mathbf{A} + \mathbf{b}\mathbf{c}^\mathsf{T}\right)^{-1} = \mathbf{A}^{-1} - \frac{1}{\left(1 + \mathbf{c}^\mathsf{T}\mathbf{A}^{-1}\mathbf{b}\right)}\mathbf{A}^{-1}\mathbf{b}\mathbf{c}^\mathsf{T}\mathbf{A}^{-1}$$

to show that the minimum deviance estimate $\hat{\mathcal{I}}_{MD}$ for $\mathcal{I}$ takes the form

$$\hat{\mathcal{I}}_{MD} = \frac{1}{\epsilon + N}\sum_{n=1}^{N} f(x_n)$$

clearly defining the scalar value ϵ. Assess the general suitability of $\hat{\mathcal{I}}_{MD}$ as a practical estimator that can be implemented. [40%]

(c) (i) Assess whether $\hat{\mathcal{I}}_{MD}$ is an unbiased estimator by taking the expectation of $\hat{\mathcal{I}}_{MD}$. [10%]

(ii) What can be said about the estimator $\hat{\mathcal{I}}_{MD}$ as $N \to \infty$? [10%]

2 The Metropolis-Hastings algorithm provides a means to draw samples from $\mathbb{R}$ with a probability distribution $\pi^*(dy)$ having density $\pi(y)$ with respect to Lebesgue Measure dy.

(a) For a generic transition kernel of the form $P(x, dy) = p(x, y)dy + r(x)\delta_x(dy)$

(i) State the conditions on $p(x, y)$ that need to be satisfied for $\pi(\cdot)$ to be the invariant density of $P(x, \cdot)$ and define the form of $r(x)$. [15%]

(ii) Prove that $\pi(\cdot)$ is the invariant density of $P(x, \cdot)$ under the conditions satisfied above. [25%]

(b) (i) For a proposal density $q(x, y)$ derive a form for the acceptance probability $\alpha(x, y)$ that satisfies the sufficient conditions to yield $\pi(\cdot)$ as the invariant density of $P(x, \cdot)$. [25%]

(ii) Write out the overall form for the Metropolis-Hastings transition kernel that targets a density $\pi(\cdot)$ with proposal $q(x, y)$ and acceptance probability $\alpha(x, y)$. [15%]

(c) Write out pseudo-code for the Metropolis-Hastings algorithm that uses a symmetric proposal density and targets the density

$$\pi(x) = \frac{1}{\mathcal{Z}}\phi(x) \quad \text{where} \quad \mathcal{Z} = \int_{\mathbb{R}} \phi(x)dx$$

[20%]

(TURN OVER

3 (a) Describe the characteristic of point wise function evaluation in a Reproducing Kernel Hilbert Space (RKHS) and how this differs from the L_2 norm in a Hilbert space of functions. [15%]

(b) Explain how the Moore-Aronszajn theorem relates a reproducing kernel function to an RKHS. Give two defining properties of a reproducing kernel function, and write down the *reproducing* property of such a function. [20%]

(c) Consider discrete counts, y_n with $n = 1, \cdots, N$, of an i.i.d process that follow a conditional Poisson probability function $p(y|\mathbf{x}) = \exp(-\mu(\mathbf{x})) \times \mu(\mathbf{x})^y / y!$ where the function value $\mu(\cdot)$ is modelled in a Reproducing Kernel Hilbert Space with an approximating functional form $\log(\hat{\mu}(\cdot)) = \sum_{n=1}^{N} \alpha_n k(\cdot, \mathbf{x}_n)$. Each $\mathbf{x}_n$ is a vector of features corresponding to each count y_n, and $k(\cdot, \cdot)$ is a reproducing kernel function. The regularised log-likelihood function is given as,

$$\mathcal{L}(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \log(p(y_i|\mathbf{x}_i)) - \frac{\lambda}{2} \| \log(\hat{\mu}) \|^2$$

where the $N \times 1$ vector $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_N]^\mathsf{T}$.

(i) Let the $N \times 1$ vector $\mathbf{y} = [y_1, \cdots, y_N]^\mathsf{T}$ and by defining the $N \times N$ matrix $\mathbf{K}$ and the $N \times 1$ vector $\hat{\boldsymbol{\mu}}$ show that the gradient of the regularised log-likelihood with respect to the weights $\boldsymbol{\alpha}$ takes the vector form of $\mathbf{K}(\mathbf{y} - \hat{\boldsymbol{\mu}} - \lambda\boldsymbol{\alpha})$. [30%]

(ii) Show that the expression for the matrix of second derivatives of the regularised log-likelihood with respect to the weights $\boldsymbol{\alpha}$ takes the form $-\mathbf{K}\mathbf{V}\mathbf{K}^\mathsf{T} - \lambda\mathbf{K}$ by defining the diagonal matrix $\mathbf{V}$. [20%]

(iii) Using the results from (i) and (ii) above write down a Newton iteration scheme to find the maximum regularised likelihood solution for the kernel weights $\boldsymbol{\alpha}$. [15%]

4 The Student t-distribution defined on $\mathbb{R}$ has a probability density function with respect to Lebesgue Measure which is given as

$$p(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

where $x \in \mathbb{R}$, ν is a parameter known as the *degrees of freedom*, and Γ is the gamma function.

(a) Derive the equation describing a Langevin Diffusion defined on $\mathbb{R}$ whose invariant density is that of the Student t-distribution. [40%]

(b) Write out an Unadjusted Langevin Algorithm (ULA) which will converge to a biased version of the Student t-distribution. [20%]

(c) The Metropolis Adjusted Langevin Algorithm (MALA) takes ULA as a proposal mechanism and removes the bias by applying an Accept-Reject step. Define the corresponding MALA acceptance probability for the Student t-distribution target. [40%]

END OF PAPER

THIS PAGE IS BLANK