SEMESTER 2 EXAMINATION 2019/20

ADVANCED MACHINE LEARNING

Duration 120 mins (2 hours)

This paper is a WRITE-ON examination paper.

You ***must*** write your Student ID on this Page and must not write your name anywhere on the paper.

All answers should be written within the designated boxes in this examination paper and sufficient space is provided for each question.

If, for some reason, space is required to complete or correct an answer to a question, use the “Additional Space” provided on the facing or adjacent page to the question. Clearly indicate which question the answer corresponds to.

No credit will be given for answers presented elsewhere and without clear indication of to what question they correspond. Blue answer books may be used for scratch; they will be discarded without being looked at.

Answer all questions. Section A (worth 40 marks) are a series of questions with short answers. Section B (worth 60 marks) involve longer questions

Student ID:

Question	Mark	*Arithmetic checked*	*Double Marked*
Total:			

University approved calculators MAY be used.

A foreign language translation dictionary (paper version) is permitted provided it contains no notes, additions or annotations.

16 page examination paper

Section A

Question A 1

(a) Explain why over expressive machines are likely to generalise poorly when the number of training examples is small. *(5 marks)*

$\overline{}$
5

(b) Explain why CNNs capture the structure of typical image datasets. *(5 marks)*

$\overline{}$
5

(c) Explain the major ways (i) gradient descent (ii) Newton's method and (iii) quasi-Netwon methods differ in terms of the information they use. *(5 marks)*

i

ii

iii

$\overline{}$
5

Additional space. Do not use unless necessary. Clearly mark corresponding question.

(d) In stochastic gradient descent (SGD) explain what are mini-batches and their possible advantages and disadvantages. *(5 marks)*

$\overline{5}$

(e) Describe the Karush-Kuhn-Tucker (KKT) conditions for constrained optimisation. *(5 marks)*

$\overline{5}$

(f) Show that the Dirichlet distribution given by $\mathrm{Dir}(\boldsymbol{p}|\alpha) = \Gamma(\alpha_0) \prod_{i=1}^{d} \frac{p_i^{\alpha_i - 1}}{\Gamma(\alpha_i)}$, where $\boldsymbol{p} = (p_1, p_2, \ldots, p_d)$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ and $\alpha_0 = \sum_{i=1}^{d} \alpha_i$ is a conjugate prior to the multinomial likelihood $\mathrm{Binom}(\boldsymbol{k}|n, p) = n! \prod_{i=1}^{d} \frac{p_i^{k_i}}{k_i!}$, where $\boldsymbol{k}$ is a vector of counts $(k_1, k_2, \ldots, k_d)$ with $\sum_{i=1}^{d} k_i = n$. Derive update equations for the parameters $\boldsymbol{\alpha}'$ of the posterior distribution after observing counts $\boldsymbol{k}$. *(5 marks)*

$\overline{5}$

Additional space. Do not use unless necessary. Clearly mark corresponding question.

(g) Prove that the set of positive semi-definite matrices is a convex set. *(5 marks)*

$\overline{}$ / 5

(h) Explain what are the hyper-parameters of a Gaussian process and why they are relatively easy to learn. *(5 marks)*

$\overline{}$ / 5

End of question A1

(a)	$\frac{}{5}$	(b)	$\frac{}{5}$	(c)	$\frac{}{5}$	(d)	$\frac{}{5}$	(e)	$\frac{}{5}$	(f)	$\frac{}{5}$	(g)	$\frac{}{5}$	(h)	$\frac{}{5}$	Total	$\frac{}{40}$

Additional space. Do not use unless necessary. Clearly mark corresponding question.

Section B

Question B 2

(a) If $\{X_i | i = 1, 2, m\}$ is a set of correlated random variables such that

$$\langle X_i \rangle = \mu \qquad\qquad \langle (X_i - \mu)(X_i - j) \rangle = \begin{cases} \sigma^2 & \text{if } i = j \\ \rho\,\sigma^2 & \text{if } i \neq j \end{cases}$$

show

$$\left\langle \left(\frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right)^2 \right\rangle = \rho\,\sigma^2 + \frac{(1-\rho)\,\sigma^2}{n}$$

(10 marks)

$\overline{10}$

(b) Using the result derived in part (a) explain why ensembling many machines can reduce the variance in the bias-variance dilemma if the machine predictions are not heavily correlated. Use this to explain the success of random forest. *(10 marks)*

$\overline{}$ / 10

End of question B2

(a) $\frac{}{10}$ (b) $\frac{}{10}$ Total $\frac{}{20}$

Additional space. Do not use unless necessary. Clearly mark corresponding question.

Question B 3

(a) We can write the loss for ridge regression as

$$L(\boldsymbol{w}) = \|\mathbf{X}\,\boldsymbol{w} - \boldsymbol{y}\|^2 + \eta\,\|\boldsymbol{w}\|^2$$

where $\mathbf{X}$ is the design matrix and $\boldsymbol{y}$ is a vector of target values. Calculate the weights, $\boldsymbol{w}^*$, that minimise the loss. By writing $\mathbf{X} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^{\mathsf{T}}$ show that we can write $\boldsymbol{w}^* = \mathbf{V}\,\hat{\mathbf{S}}^{+}\mathbf{U}^{\mathsf{T}}\boldsymbol{y}$, where the elements of $\hat{\mathbf{S}}^{+}$ are zero everywhere except for the diagonal where $\hat{S}^{+}_{ii} = s_i/(s_i^2 + \eta)$ with s_i being the singular values of $\mathbf{X}$ (i.e. $s_i = S_{ii}$). *(15 marks)*

$\overline{15}$

(b) Use the result that you derive to explain how adding the L_2 regulariser $\eta \|\boldsymbol{w}\|^2$ improves the conditioning of the solution and is likely to improve generalisation.
(5 marks)

$\overline{}$
5

End of question B3

(a) $\overline{15}$ (b) $\overline{5}$ Total $\overline{20}$

Additional space. Do not use unless necessary. Clearly mark corresponding question.

Question B 4

(a) Write a Lagrangian for the linear programming problem of choosing $\boldsymbol{x}$ to minimise $\boldsymbol{c}^{\mathsf{T}}\boldsymbol{x}$, subject to the constraints $\mathbf{M}\,\boldsymbol{x} = \boldsymbol{b}$. Show that the Lagrangian can be rewritten to obtain the dual problem where the roles of the variables $\boldsymbol{x}$ and the Lagrange multipliers are exchanged. Write down the dual problem as a maximisation problem plus a new set of constraints. *(5 marks)*

$\overline{5}$

(b) Describe the Wasserstein distance as a linear programming problem and describe the dual problem. Describe how this is used in the Wasserstein GAN. *(15 marks)*

$\overline{15}$

End of question B4

(a) $\frac{}{5}$ (b) $\frac{}{15}$ Total $\frac{}{20}$

Additional space. Do not use unless necessary. Clearly mark corresponding question.

