UNIVERSITY OF SOUTHAMPTON COMP6208W1

SEMESTER 2 EXAMINATION 2020 - 2021

ADVANCED MACHINE LEARNING

DURATION 120 MINS (2 Hours)

This paper contains 4 questions

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

An outline marking scheme is shown in brackets to the right of each question.

This examination is worth 80%. The coursework was worth 20%.

This is a take home examination. Detailed instructions are given on line.

University approved calculators MAY be used.

A foreign language dictionary is permitted ONLY IF it is a paper version of a direct 'Word to Word' translation dictionary AND it contains no notes, additions or annotations.

5 page examination paper.

Section A

Question A1.

(a) Explain what it means to normalise the input features and why this is often a good thing to do. [5 marks]

(b) What is the purpose of cross-validation and what is its cost? [5 marks]

(c) To find the maximum or minimum of a function 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]

(d) Explain why CNNs capture the structure of typical image datasets. [5 marks]

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

(f) Show that an empirical covariance matrix, $\mathbf{C}$, can be written as $\mathbf{X}\mathbf{X}^{\mathsf{T}}$ and hence prove that its eigenvalues are non-negative. [5 marks]

(g) Prove that $f(x) = \exp(c\,x)$ is a convex-up function. [5 marks]

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

Section B

Question B1.

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

$$\mathbb{E}[X_i] = \mu \qquad \mathbb{E}[(X_i - \mu)(X_j - \mu)] = \begin{cases} \sigma^2 & \text{if } i = j \\ \rho\,\sigma^2 & \text{if } i \neq j \end{cases}$$

show

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)^2\right] = \rho\,\sigma^2 + \frac{(1-\rho)\,\sigma^2}{n}$$

[10 marks]

(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]

TURN OVER

Question B2.

(a) Consider gradient descent

$$x^{(t+1)} = x^{(t)} - r\, f'(x^{(t)})$$

acting on a function $f(x) = x^2$. Explain what happens starting from some point $x^{(0)}$ if (i) $0 < r < \frac{1}{2}$, (ii) $r = \frac{1}{2}$, (iii) $\frac{1}{2} < r < 1$, (iv) $r = 1$ and (v) $r > 1$. [10 marks]

(b) Describe the problems that can arise in finding the optimum of a high-dimensional loss function and solution to them.

[10 marks]

Question B3.

(a) Explain how Markov chain Monte-Carlo (MCMC) techniques are used to solve Bayesian inference problems and what problem do they solve.
[10 marks]

(b) Describe the Metropolis algorithm and show that it satisfies detailed balance. Explain why this is important.
[10 marks]

END OF PAPER