CST1
COMPUTER SCIENCE TRIPOS Part IB

Wednesday 16 June 2021 11:30 to 14:30 BST

COMPUTER SCIENCE Paper 6

*Answer **five** questions.*

*Submit each question answer in a **separate** PDF. As the file name, use your candidate number, paper and question number (e.g., `1234A-p6-q6.pdf`). Also write your candidate number, paper and question number at the start of each PDF.*

You must follow the official form and conduct instructions for this online examination

1 Artificial Intelligence

The standard *linear regression* model uses a hypothesis

$$h(\mathbf{x}, \mathbf{w}, b) = \mathbf{w}^T \mathbf{x} + b$$

to fit m examples $((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$ by minimizing the error

$$E(\mathbf{w}, b) = \sum_{i=1}^{m} \left(y_i - h(\mathbf{x}_i, \mathbf{w}, b)\right)^2 .$$

(*a*) Derive a *gradient descent* algorithm for training the linear regression model described. [5 marks]

(*b*) In the application of interest, you believe that it is desirable to train such that the learned parameters have $||\mathbf{w}|| \simeq 1$. Suggest a modification to $E(\mathbf{w}, b)$ that facilitates this, and derive the corresponding gradient descent training algorithm. [5 marks]

(*c*) Describe the components of a general *heuristic search* problem. [4 marks]

(*d*) You are faced with a heuristic search problem, but the heuristics you have so far developed are less effective than desired. Suggest two ways in which supervised machine learning might be used to develop a better heuristic, mentioning if necessary any corresponding disadvantages of using the approach. You may assume that a collection of problems to be solved by the heuristic search is available. [6 marks]

2 Artificial Intelligence

A *Boolean satisfiability problem* has four variables, x_1, x_2, x_3 and x_4. A literal l can be a variable or its negation, denoted $\bar{l}$. The formula of interest, in conjunctive normal form (CNF), is

$$f = (x_2 \vee \overline{x_3}) \wedge (\overline{x_2} \vee x_3) \wedge (x_1 \vee x_2 \vee \overline{x_4}). \tag{1}$$

The aim is to find assignments to the variables such that f is true under the usual rules for Boolean operations. This question addresses the use of more general *constraint satisfaction* to solve this problem.

(*a*) Give a general description of a *constraint satisfaction problem (CSP)*. [3 marks]

(*b*) Explain how a Boolean satisfiability problem in CNF form and with n variables can be converted to a CSP, also having n variables and having a suitable constraint for each clause. Illustrate your answer using the 4-variable formula f in (1). [3 marks]

(*c*) Explain, again using a constraint corresponding to a clause from (1), how general constraints can be converted to binary constraints. Provide a graph illustrating the problem from (1) after it has been converted to a CSP with only binary constraints. [4 marks]

(*d*) Explain, how *forward checking* works in the context of a general CSP. How does this benefit a CSP solver? [3 marks]

(*e*) Using the CSP equivalent you developed for (1), with only binary constraints, demonstrate how *forward checking* works using the sequence of assignments $x_1 = F$, $x_2 = F$, $x_4 = T$. [5 marks]

(*f*) How would you expect the solution obtained when applying forward checking to be affected if constraints were allowed to propagate more widely? [2 marks]

3 Complexity Theory

(a) Define the set of Boolean expressions 2CNF and the language 2SAT over them. [2 marks]

(b) For a Boolean expression ϕ in 2CNF, let $G(\phi)$ be the directed graph with *vertices* the variables of ϕ and their negation, and with *edges* (a, b) if, and only if, there is a clause $(\neg a \vee b)$ or $(b \vee \neg a)$ in ϕ. Note that an edge (a, b) is in $G(\phi)$ if, and only if, so is the edge $(\neg b, \neg a)$.

Prove that a Boolean expression ϕ in 2CNF is unsatisfiable if, and only if, there is a variable x in ϕ such that there are paths from x to $\neg x$ and from $\neg x$ to x in $G(\phi)$. [*Hint:* Recall that the proposition $(\neg P \vee Q)$ is equivalently the implication $(P \rightarrow Q)$.] [12 marks]

(c) Argue as to whether or not 2SAT is in NL, in P, and in NP. Your answer may use the fact that NL is closed under complementation. [6 marks]

4 Complexity Theory

(*a*) For a complexity class $\mathcal{C}$, let co-$\mathcal{C} = \{ \overline{L} \mid L \in \mathcal{C} \}$ and say that $\mathcal{C}$ is closed under complementation whenever $\mathcal{C} =$ co-$\mathcal{C}$.

Argue as to whether the following statements are true, false, or unknown.

(*i*) All deterministic time complexity classes are closed under complementation. [3 marks]

(*ii*) All non-deterministic time complexity classes are closed under complementation. [3 marks]

(*b*) For a mapping $f : \Sigma \to \Sigma$ on an alphabet Σ and a language $L \subseteq \Sigma^*$, define $f[L] = \{ f^\natural(w) \in \Sigma^* \mid w \in L \}$ where $f^\natural(a_1 \cdots a_n) = f(a_1) \cdots f(a_n)$.

Prove that $L \in$ NP implies $f[L] \in$ NP. [4 marks]

(*c*) Consider the following decision problem.

> Q: Given natural numbers m and n in $\mathbb{N}$, and bits $a_{i,j}^{(k)}$ and b_k in $\{0,1\}$ for $1 \leq k \leq m$ and $1 \leq i,j \leq n$, determine whether the system of equations $\sum_{1 \leq i,j \leq n} a_{i,j}^{(k)}\, x_i\, x_j = b_k$ $(1 \leq k \leq m)$ with unknowns $x_1, \ldots, x_n$ has a solution in arithmetic modulo 2.

(*i*) Prove that Q is in NP. [3 marks]

(*ii*) By means of a polynomial-time reduction from the problem 3CNF, or otherwise, prove that Q is NP-hard. [*Hint:* Note, for instance, that $x = \neg y$ in the Boolean algebra $\{0,1\}$ if, and only if, $xx + yy = 1$ in arithmetic modulo 2.] [7 marks]

You may use standard results provided that you state them clearly.

5 Computation Theory

(*a*) For each $n, e \in \mathbb{N}$, let $\varphi_e^{(n)}$ denote the partial function $\mathbb{N}^n \rightharpoonup \mathbb{N}$ computed by the register machine with index e using registers $\mathtt{R}_1, \ldots, \mathtt{R}_n$ to store the n arguments and register $\mathtt{R}_0$ to store the result, if any.

Explain why for each $m, n \in \mathbb{N}$ there is a totally defined register machine computable function $S_{m,n} : \mathbb{N}^{1+m} \to \mathbb{N}$ with the property that for all $(e, \vec{x}) \in \mathbb{N}^{1+m}$ and $\vec{y} \in \mathbb{N}^n$

$$\varphi^{(n)}_{S_{m,n}(e,\vec{x})}(\vec{y}) \equiv \varphi_e^{(m+n)}(\vec{x}, \vec{y}) \tag{1}$$

where $\equiv$ denotes Kleene equivalence: for all $z \in \mathbb{N}$, the left-hand side is defined and equal to z if and only if the right-hand side is defined and equal to z. Your explanation should make clear what assumptions you are making about the relationship between numbers and register machine programs. [10 marks]

(*b*) Let $f : \mathbb{N}^{1+m+n} \rightharpoonup \mathbb{N}$ be a register machine computable partial function of $1+m+n$ arguments for some $m, n \in \mathbb{N}$.

(*i*) Why is the partial function $\hat{f} : \mathbb{N}^{1+m+n} \rightharpoonup \mathbb{N}$ satisfying for all $(z, \vec{x}, \vec{y}) \in \mathbb{N}^{1+m+n}$

$$\hat{f}(z, \vec{x}, \vec{y}) \equiv f(S_{1+m,n}(z, z, \vec{x}), \vec{x}, \vec{y}) \tag{2}$$

register machine computable? [3 marks]

(*ii*) By considering $S_{1+m,n}(e, e, \vec{x})$ where e is an index for the partial function $\hat{f}$ in part (*b*)(*i*), prove that there is a totally-defined register machine computable function $\mathsf{fix}\, f : \mathbb{N}^m \to \mathbb{N}$ with the property that for all $\vec{x} \in \mathbb{N}^m$ and $\vec{y} \in \mathbb{N}^n$

$$\varphi^{(n)}_{\mathsf{fix}\, f(\vec{x})}(\vec{y}) \equiv f(\mathsf{fix}\, f(\vec{x}), \vec{x}, \vec{y}) \tag{3}$$

[7 marks]

6 Computation Theory

A set A equipped with a binary operation $@ : A \times A \to A$ is a *combinatory algebra* if there are elements $K, S \in A$ satisfying for all $a, b, c \in A$

$$@(@(K, a), b) = a \tag{1}$$

$$@(@(@(S, a), b), c) = @(@(a, c), @(b, c)) \tag{2}$$

(*a*) Show that there is a binary operation on the set of equivalence classes of closed λ-terms for the equivalence relation of β-conversion that makes it a combinatory algebra. [5 marks]

(*b*) Show that every combinatory algebra A contains an element I satisfying

$$@(I, a) = a \tag{3}$$

for all $a \in A$. [Hint: what does (2) tell us when $a = b = K$?] [2 marks]

(*c*) For an arbitrary combinatory algebra A, let $A[x]$ denote the set of expressions given by the grammar

$$e ::= x \mid \ulcorner a \urcorner \mid (ee)$$

where x is some fixed symbol not in A and a ranges over the elements of A. Given $e \in A[x]$ and $a \in A$, let $e[x := a]$ denote the element of A resulting from interpreting occurrences of x in e by a, interpreting the expressions of the form $\ulcorner a' \urcorner$ by a' and interpreting expressions of the form (ee') using @.

(*i*) Give the clauses in a definition of $e[x := a]$ by recursion on the structure of e. [2 marks]

(*ii*) For each $e \in A[x]$ show how to define an element $\Lambda_x e \in A$ with the property that

$$@(\Lambda_x e, a) = e[x := a] \tag{7}$$

for all $a \in A$. [6 marks]

(*d*) Recall the usual encoding of Booleans in λ-calculus. Using Part (*c*)(*ii*), show that in any combinatory algebra A there are elements $\mathit{True}, \mathit{False} \in A$ and a function $\mathit{If} : A \times A \to A$ satisfying

$$@(\mathit{If}(a, b), \mathit{True}) = a \tag{11}$$

$$@(\mathit{If}(a, b), \mathit{False}) = b \tag{12}$$

for all $a, b \in A$ [5 marks]

7 Data Science

(*a*) Let x_t be the number of new COVID infections on date t. We anticipate approximately exponential growth or decay, $x_{t+1} \approx (1+\lambda)x_t$, and we would like to estimate λ from a dataset $(x_1, \ldots, x_T)$.

(*i*) Find the maximum likelihood estimator for λ for the model

$$X_{t+1} \sim \text{Poisson}\big((1+\lambda)x_t\big)$$

[2 marks]

(*ii*) Find the maximum likelihood estimator for λ for the model

$$X_{t+1} \sim \text{Normal}\big((1+\lambda)x_t,\ (\sigma x_t)^2\big)$$

[3 marks]

(*iii*) For the latter model, explain how to compute a 95% confidence interval for λ. Explain the resampling step carefully. [4 marks]

(*b*) We don't actually know the number of new infections x_t on date t: we only know the number of new positive test results, y_t. We anticipate $y_t \approx \beta_{\text{dow}(t)} x_t$, where dow($t$) gives the day of the week for date t. We would like to estimate not only λ but also $\beta_{\text{Mon}}, \ldots, \beta_{\text{Sun}}$ from the dataset $(y_1, \ldots, y_T)$.

(*i*) Propose a probability model for Y_{t+1} in terms of y_t. [5 marks]

(*ii*) Explain briefly how to estimate the parameters of your model. In your answer, you should consider whether or not the parameters are identifiable. [6 marks]

8 Data Science

(*a*) In a COVID vaccine trial, n_0 subjects were given a placebo and n_1 were given the vaccine; x_0 of the placebo subjects developed the disease and x_1 of the vaccinated subjects. Considering the probability model $X_k \sim \text{Binomial}(n_k, p_k)$, the vaccine efficacy is defined to be $e = 1 - p_1/p_0$.

(*i*) State the maximum likelihood estimators for p_0 and p_1. Give a formula for the maximum likelihood estimator for e. [2 marks]

(*ii*) Explain how to compute a 95% confidence interval for e. Also explain how to test whether $e > 0.5$. [7 marks]

(*b*) Further data about the trial has been made available, and we learn that subjects weren't all enrolled for the same length of time. We are given a full dataset consisting of three features, the predictor variable d_i and the response variables (t_i, c_i) for subject i. Here $d_i = 1$ if the subject received the vaccine and $d_i = 0$ otherwise; $c_i = 1$ if the subject developed the disease and $c_i = 0$ otherwise; and t_i is the day on which the subject developed the disease if $c_i = 1$, and the number of days enrolled in the trial otherwise.

Consider the following probability model. Among vaccinated subjects, the vaccine is effective with probability f and ineffective otherwise. Effectively vaccinated subjects never get the disease. For ineffectively vaccinated subjects, and for subjects on placebo, each day there is a probability q of developing the disease. The parameters f and q are unknown.

(*i*) For a subject i who received placebo, give an expression for the likelihood of the pair (t_i, c_i). [4 marks]

(*ii*) For a subject i who received the vaccine, give an expression for the likelihood of the pair (t_i, c_i). [4 marks]

(*iii*) Give an expression for the log likelihood of the entire dataset. [3 marks]

[*Note:* Your answers to this question should be symbolic. But you may like to know that in the low-dose part of the Oxford–AstraZeneca trial, $n_0 = 1374$, $n_1 = 1367$, $x_0 = 30$, and $x_1 = 3$.]

9 Logic and Proof

(*a*) Suppose we have the propositional symbols P_1, P_2, …, P_n, where $n > 2$, and consider the set of clauses

$$\{\neg P_1, P_2\} \quad \{\neg P_2, P_3\} \quad \cdots \quad \{\neg P_n, P_1\}.$$

(*i*) List the satisfying interpretations (if any) of this set of clauses, with brief justification. [3 marks]

(*ii*) Regarding the set of clauses above as a single propositional formula, and using the variable ordering P_1, P_2, …, P_n, sketch the corresponding BDD. Does the choice of variable ordering here significantly affect the size of the resulting BDD? [5 marks]

(*iii*) Briefly describe the set of clauses that would be generated by a resolution theorem prover, starting with the set of clauses above. [3 marks]

(*b*) For the following set of clauses, either exhibit a model, or show that none exists using resolution. Below, a and b are constants, while y and z are variables.

$$\{P(a, f(a)),\ P(a, b)\}$$
$$\{P(a, f(a)),\ \neg P(z, b),\ P(z, f(a))\}$$
$$\{\neg P(a, f(a)),\ P(g(y), y),\ \neg P(a, y)\}$$
$$\{\neg P(a, f(a)),\ \neg P(g(y), f(a)),\ \neg P(a, y)\}$$

[9 marks]

10 Logic and Proof

(*a*) A mysterious propositional connective, $\odot$, has the following sequent calculus rule, $(\odot l)$:

$$\frac{A, B, \Gamma \Rightarrow \Delta \qquad \Gamma \Rightarrow \Delta, A, B}{A \odot B, \Gamma \Rightarrow \Delta}$$

Present the corresponding right-side sequent calculus rule, $(\odot r)$, along with the truth table for $\odot$. [5 marks]

(*b*) For the following formula, either exhibit a formal proof (using the sequent calculus, augmented with the $(\odot l)$ rule above) or exhibit a falsifying interpretation:

$$\forall x(P(x) \odot Q(x)), \exists x\, P(x) \Rightarrow \exists x\, Q(x)$$

[5 marks]

(*c*) Use the DPLL method to find **all** models (if any) satisfying the following set of formulas.

$$(P \land R) \to Q$$
$$(\neg Q \land R) \to P$$
$$Q \to (P \lor R)$$
$$\neg(P \land Q \land R)$$
$$P \to R$$

[5 marks]

(*d*) The modal logic S5 differs from S4 in requiring the accessibility relation to be symmetric, as well as reflexive and transitive. Present a formula that is valid for S5 but not for S4. Explain why it is valid for S5 and demonstrate that it is not valid for S4 by drawing an S4 modal frame for which it fails. [5 marks]

END OF PAPER