Calculators may be used in this examination provided they are not capable of being used to store alphabetical information other than hexadecimal numbers

UNIVERSITY OF BIRMINGHAM

School of Computer Science

Artificial Intelligence 1

Main Summer Examinations 2024

Time allowed: 2 hours

[Answer all questions]

Note

Answer ALL questions. Each question will be marked out of 20. The paper will be marked out of 60, which will be rescaled to a mark out of 100.

Question 1 Supervised Learning

(a) Consider a data set with 4 instances ($\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \mathbf{x}^{(4)}$), and each is described by 3 attributes (x_1: age, x_2: weight, x_3: gender). The task is to classify them into a medical condition (either yes or no). The detailed data vectors can be found in Table 1. The labels of the first three instances are known.

(i) Can you use k-Nearest Neighbour (k-NN) to find out the label of $\mathbf{x}^{(4)}$? Please use normalisation and Manhattan distance for numeric attributes, Hamming distance for discrete attributes, and set $k = 1$. The Manhattan distance between two d-dimensional vectors is defined as:

$$D\left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}\right) = \sum_{i=1}^{d} \left| x_i^{(1)} - x_i^{(2)} \right|$$

	age	weight	gender	label y
$\mathbf{x}^{(1)}$	10	150	male	yes
$\mathbf{x}^{(2)}$	15	180	female	yes
$\mathbf{x}^{(3)}$	7	100	female	no
$\mathbf{x}^{(4)}$	10	100	male	?

Table 1: Data Set

(ii) If k is set to 2 in question (i), a draw would happen (i.e. one neighbour has label "yes" and the other neighbour has label "no"). Can you improve the labeling method to solve this situation (but without changing k), and label $\mathbf{x}^{(4)}$ using the suggested method? **[10 marks]**

(b) Consider the following two classification methods: Logistic Regression (LR), and k-Nearest Neighbour with k = 1 (1-NN) employing the Euclidean distance. Create and draw a 2D labelled classification data set on which the leave-one-out validation error of LR is zero, but the leave-one-out validation error of 1-NN is maximal. Describe what "leave-one-out" validation strategy is, and justify your design. **[10 marks]**

Question 2 Unsupervised Learning

(a) Consider a data set consisting of the following feature vectors: $\mathbf{x}^{(1)} = (1, 6)$, $\mathbf{x}^{(2)} = (0, 4)$, $\mathbf{x}^{(3)} = (2, 1)$ and $\mathbf{x}^{(4)} = (6, 3)$, which are grouped into two clusters $C_1 = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$ and $C_2 = (\mathbf{x}^{(3)}, \mathbf{x}^{(4)})$. Compute the within cluster sum of squares (WCSS) of the above cluster assignment. Starting from this cluster assignment, if you run one iteration of K-means algorithm, what will be the resulting new clusters?

Hint: The WCSS of a cluster assignment $\mathcal{C}$ consisting of K clusters is the

$$\text{WCSS}(\mathcal{C}) = \sum_{C \in \mathcal{C}} \sum_{e \in C} d^2_{\text{Euc}}(e, \text{Centroid}(C)),$$

where $d^2_{\text{Euc}}(e, \text{Centroid}(C))$ denotes the squared Euclidean distance between example e and the centroid $\text{Centroid}(C)$ of cluster C.

[10 marks]

(b) Table 2 summarizes the output of clustering the data points $\{4, 10, 1, 5, 20, 15\}$ into three clusters with labels C_1, C_2 and C_3 by some clustering algorithm. The last column indicates the true labels R, G, B of the points supplied by an external agent.

Data	Cluster Label	True Label
4	C_1	R
10	C_1	R
1	C_2	G
5	C_2	B
20	C_3	G
15	C_3	R

Table 2: Data points, Cluster Labels and True Labels

(i) From the information available in Table 2, which cluster has the highest purity? Explain your answer.

(ii) Starting with the cluster assignment as given in Table 2, which two clusters will merge together next in a (a) single-linkage dendrogram, (b) complete linkage dendrogram? Work out your answer for (a) and (b) respectively.

[10 marks]

Question 3 Search & Optimisation

(a) Least squares estimation is an optimisation method that aims to minimise the sum of the squared distances to fit a curve to n data points:

$$\min g(\mathbf{w}) = \sum_{i=1}^{n} (y^{(i)} - f_{\mathrm{w}}(x^{(i)}))^2,$$

where the curve is given by $f_{\mathrm{w}}(x^{(i)})$. In 2D, this can be seen as a linear regression problem, where $f_{\mathrm{w}}(x^{(i)}) = w_1 x^{(i)} + w_0$.

(i) Solve the optimisation problem corresponding to the least squares minimisation in a 3D space where the goal is to fit a plane to n 3D data points.

Hint: recall that the equation of a plane in 3D space is $f_{\mathrm{w}}(x^{(i)}, y^{(i)}) = w_2 x^{(i)} + w_1 y^{(i)} + w_0$; start with the first formula and solve the corresponding minimisation problem in the form $Aw = b$.

(ii) Provide the equation of the plane, i.e., the solution of the optimisation problem, given the following data points:

$$(x^{(i)}, y^{(i)}, z^{(i)}) = (2, 1, 0), (3, 2, 0), (1, 3, 0),$$

where $x^{(i)}, y^{(i)}, z^{(i)}$ are the x-, y- and z-coordinates of the plane.

[10 marks]

(b) Consider again the least square estimation method discussed in part (a) when we want to fit a line to 2D data points. Provide a formulation of this problem as a search problem. Imagine we want to use A* to solve this problem: discuss which heuristic is consistent in this case and why.

[10 marks]

End of Paper

This page intentionally left blank.

Do not complete the attendance slip, fill in the front of the answer book or turn over the question paper until you are told to do so

Important Reminders

- Coats/outwear should be placed in the designated area.
- Unauthorised materials (e.g. notes or Tippex) must be placed in the designated area.
- Check that you do not have any unauthorised materials with you (e.g. in your pockets, pencil case).
- Mobile phones and smart watches **must** be switched off and placed in the designated area or under your desk. They must not be left on your person or in your pockets.
- You are not permitted to use a mobile phone as a clock. If you have difficulty seeing a clock, please alert an Invigilator.
- You are not permitted to have writing on your hand, arm or other body part.
- Check that you do not have writing on your hand, arm or other body part – if you do, you must inform an Invigilator immediately
- Alert an Invigilator immediately if you find any unauthorised item upon you during the examination.

Any students found with non-permitted items upon their person during the examination, or who fail to comply with Examination rules may be subject to Student Conduct procedures.