UNIVERSITY OF BIRMINGHAM

School of Computer Science

First Year Undergraduate

06-34238

34238 LC Artificial Intelligence 1

Resit Examinations 2022

[Answer all questions]

Question 1 Clustering

(a) Use hierarchical agglomerative clustering with complete linkage to cluster a 1-dimensional dataset with the following points: 3, 7, 8, 11, 17, 25, 27.

(i) Show your step-by-step calculation of how clusters are formed. If there is a tie, select the first pair from left to right. **[6 marks]**

(ii) Draw the resulting dendrogram with heights on one side. Assume that we want to cluster this dataset into 3 clusters. Which 3 clusters would the dendrogram give us? **[4 marks]**

i)

	3	7	8	11	17	25	27
3	0	4	5	8	14	22	24
7		0	1	4	10	18	20
8			0	3	9	17	19
11				0	6	14	16
17					0	8	10
25						0	2
27							0

↓

	3	7,8	11	17	25	27
3	0	5	8	14	22	24
7,8		0	4	10	18	20
11			0	6	14	16
17				0	8	10
25					0	2
27						0

↓

	3	7,8	11	17	25,27
3	0	5	8	14	24
7,8		0	4	10	20
11			0	6	16
17				0	10
25,27					0

	3	7,8,11	17	25,27
3	0	8	14	24
7,8,11		0	10	20
17			0	10
25,27				0

↓

	3,7,8,11	17	25,27
3,7,8,11	0	14	24
17		0	10
25,27			0

↓

	3,7,8,11	17,25,27
3,7,8,11	0	24
17,25,27		0

7,8 @ 1

25, 27 @ 2

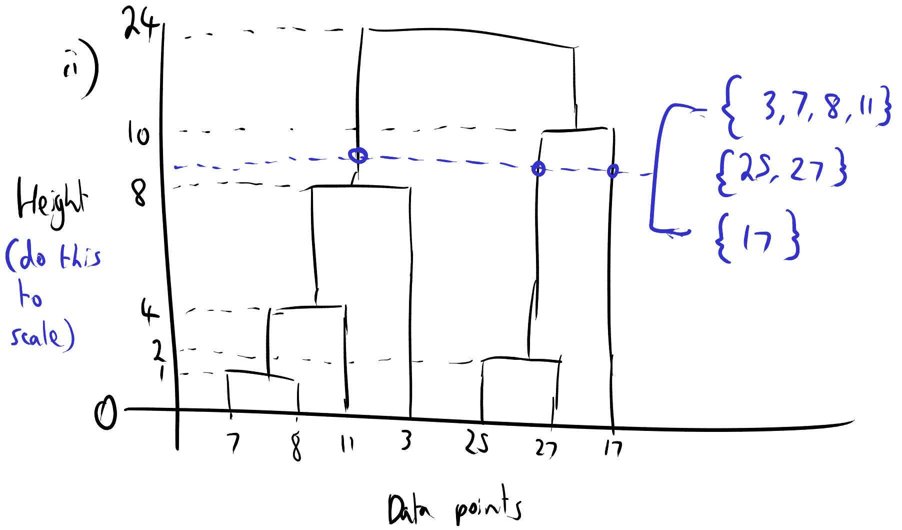
7,8,11 @ 4

3,7,8,11 @ 8

17,25,27 @ 10

3,7,8,11,17,25,27 @ 24

ii)



7,8 @ 1

25, 27 @ 2

7,8,11 @ 4

3,7,8,11 @ 8

17,25,27 @ 10

3,7,8,11,17,25,27 @ 24

(b) Now use K-means to cluster this dataset with the following settings: 1) the number of clusters is 3; 2) the three initial cluster centroids are 7, 8, 25. Will K-means give the same three clusters as the hierarchical agglomerative clustering result from the previous question? **Show the calculations.** **[5 marks]**

Shay

1)b)

points

centroids	3	7	8	11	17	25	27
7	4	0	1	4	10	18	20
8	5 ~~X~~	1	0	3	9	17	19
25	22	18	17	14	8	0	2
7	4	0	1	4	10	18	20
7.33	4.33	0.33	0.67	3.67	9.67	17.67	19.67
23	20	16	15	12	6	2	4
5	2	2	3	6	12	20	22
9.5	6.5	2.5	1.5	1.5	7.5	15.5	17.5
23	20	16	15	12	6	2	4

{3, 7}, {8, 11}, {17, 25, 27} (k-means)

Danny

b) Centroids: 7, 8, 25

Distances: (re-fit)

	3	7	8	11	17	25	27
7	4	0	1	4	10	18	20
8	5	1	0	3	9	17	19
25	22	18	17	14	8	0	2

Squared Euclidean Distance?

New clusters: {3,7}, {8,11}, {17,25,27}

New centroids: 5, 9.5, 23

Re-fit:

	3	7	8	11	17	25	27
5	2	2	3	6	12	20	22
9.5	6.5	2.5	1.5	1.5	7.5	15.5	17.5
23	20	16	15	12	6	2	4

New clusters: {3,7}, {8,11}, {17,25,27}

These clusters are the same as the previous iteration, so the algorithm terminates.

K-Means gives a different clustering result.

Question 2 Supervised Learning

(a) Often in practice, the attributes of a data set are normalised or standardised, in a pre-processing step, before running a classification or a regression method. Give one example scenario of a classifier or regressor where you think this pre-processing step is important, and one example scenario where you think it is not. **Justify your answers.** **[7 marks]**

kNN

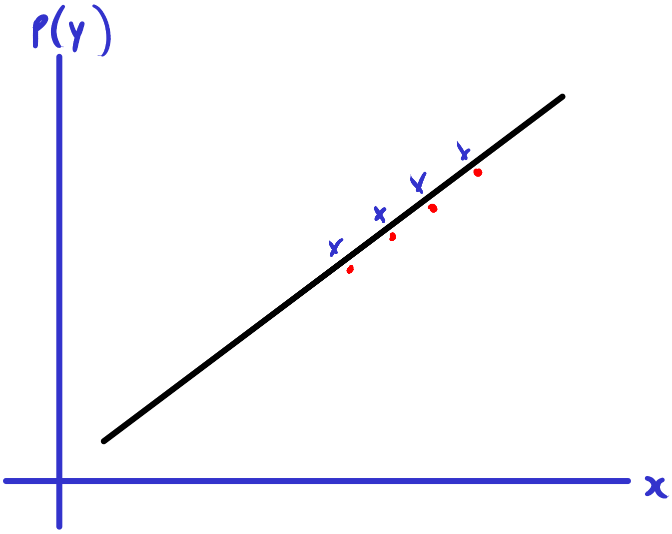
Prevent any variable from dominating the distance metrics

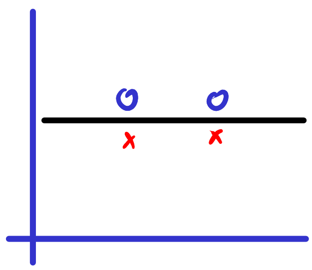
Shay

2) a) This preprocessing step would be important if we are trying to model how ~~a person's~~ an adult's age and weight affects their income via a linear regression model. Both the ages and weights will have different ranges, e.g. ages will be 18+ and weights may be anywhere up to a hundred kg. By ~~normalised~~ normalising the data, we can weight all of these attributes equally.

The preprocessing step wouldn't be ~~required~~ important if we only had 1 independent variable, for example just using age to predict or model income, as the weights of the model would adjust accordingly and no different weightages of attributes need to be considered.

(b) Consider the following two classification methods: Logistic Regression (LR), and k-Nearest Neighbour with $k = 1$ employing the Euclidean distance (1NN). 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 1NN is maximal. **[8 marks]**





Question 3 Search Strategies

Tic-Tac-Toe is a common game played between two individuals. In this 2-player game, players play in turn by marking the spaces either with a O or a X in one of the squares of a 3×3 grid. The condition for victory is to place three of the same marks (O or X) in a horizontal, vertical or diagonal row. A draw can happen when a victory condition is not achieved and all the squares are full.

This problem can be formulated as a search problem as follows:

- The initial state and one possible goal state for this game are depicted in the figure below. A goal state is defined by either a victory condition or a draw.
- Actions: you can place a O in an empty square if the previous action was placing a X; you can place a X in an empty square if the previous action was placing a O. When you expand the nodes, choose the next node corresponding to the action in the following order: place a O in an empty square, place a X in an empty square, then keep alternating between the two. Always start placing the marks (O or X) from the top-left square, then top-middle square, then row by row following the same pattern.
- Nodes are identified by the arrangement of Os and Xs in the grid.
- The cost of each action is equal to 1. Always avoid loopy paths.

	X	
	O	
X	O	X

Initial State

X	X	O
X	O	O
X	O	X

A Goal State

(a) Generate the depth first tree until the goal node is found. Write down the steps to solve the problem, from the initial state to the goal state. When expanding the nodes, use the grid to identify the nodes. **[10 marks]**

(b) Write down the solution for this problem. What is the cost of this solution? **[5 marks]**

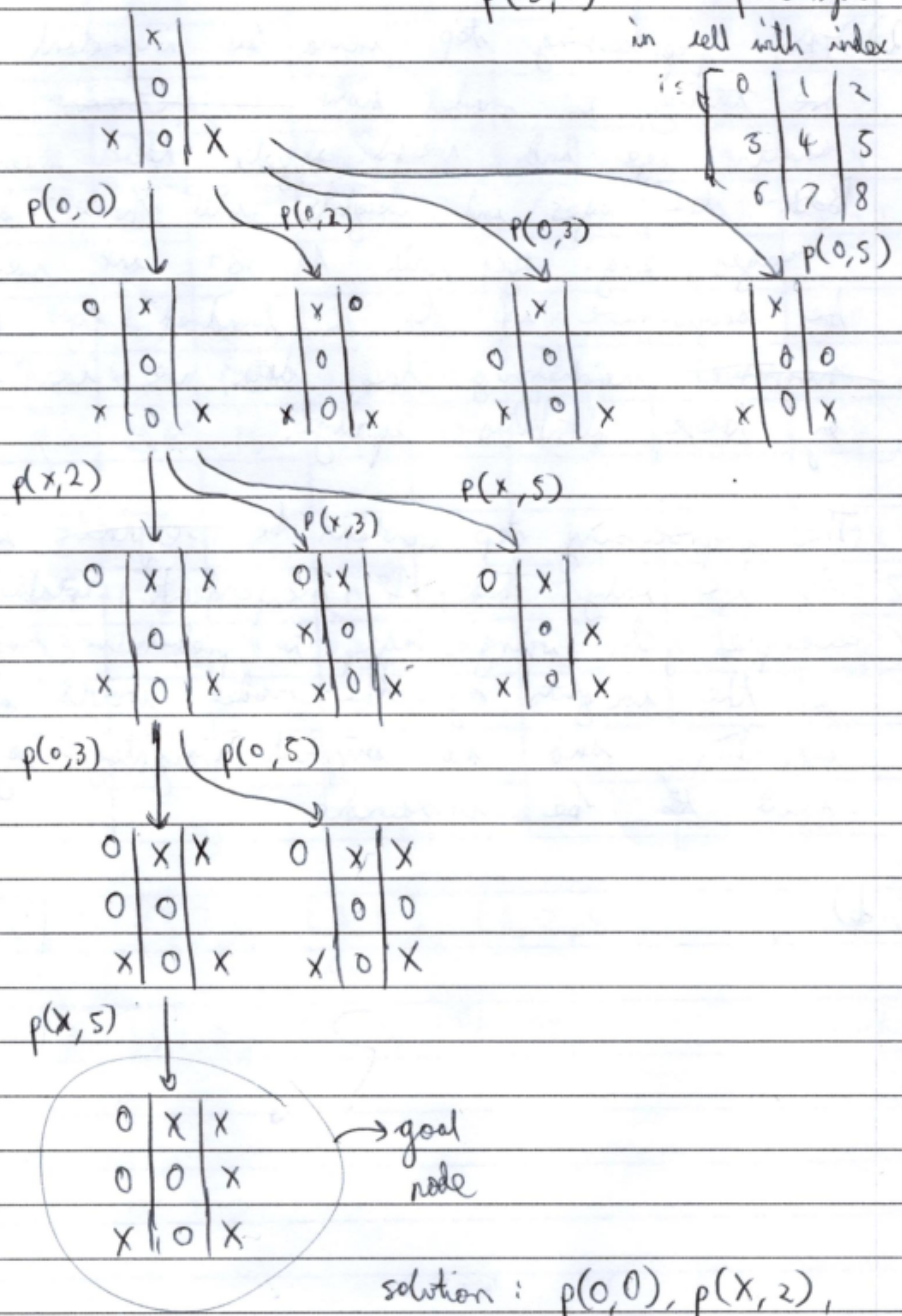
Shay

3) a)

$p(s, i)$ means place symbol s in cell with index i

i:

0	1	2
3	4	5
6	7	8



b)

solution : $p(0,0)$, $p(X,2)$, $p(0,3)$, $p(X,5)$

cost $= 1 + 1 + 1 + 1$

$= 4$

Question 4 Optimisation Problem Formulation

Consider an illustrative problem where you wish to allocate n contractors to m tasks in a project, where $n > 0$, $m > 0$ and $n > m$. Each contractor i, $0 < i \leq n$, has a cost c_i corresponding to the amount of money they charge per hour. Each task j, $0 < j \leq m$, requires h_j hours of work. Each contractor must be allocated to at most one task, whereas each task must be allocated to one and only one contractor. We would like to find an allocation of contractors to tasks that minimises the total cost of the project, in terms of monies paid to contractors.

Consider that someone tried to formulate this optimisation problem as follows:

$$\text{minimize} \quad f(\mathbf{x}) = \sum_{j=1}^{m} h_j \times c_{x_j},$$

where $\mathbf{x}$ is a vector of size m. Each position x_j, $0 < j \leq m$, of this vector contains an integer number i, $0 < i \leq n$, corresponding to a contractor i allocated to task j.

However, this problem formulation does not list any explicit constraints.

(a) Why is this an issue for this problem? **[7 marks]**

(b) Propose an adjustment of this problem formulation to overcome this issue. **Explain your proposed adjustment.** **[8 marks]**

a) We aren't checking that each contractor is assigned to at most one task.

Victor

b) $\boxed{\cancel{y_j(x)=\sum_{i=0}^{n} I(x_i=j)=1, \forall j. 0<j\leq m}}$

Where I(a=b) is 1 if a==b and 0 otherwise.

Contractors
1 Adwit
2 Victor
3 Danny
4 Jizheng

Tasks
1 Revise
2 Leetcode
3 Mop

$h=(2, 4, 10)$

$x=(1, 3, 4)$

$c=(5, 5, 5, 30)$

$$\underbrace{y_i(x)=\left(\sum_{j=0}^{n} I(x_j=i)\right)}_{\text{how many tasks is contractor } i \text{ assigned to?}} \leq 1. \ \forall i \ 0<i\leq n$$

how many tasks is contractor i assigned to?