1. *(20 marks)*

i) *(8 marks)* State whether each of the following statements is *true* or *false*. To obtain credit, you **must** in each case **explain** your answer correctly and in detail.

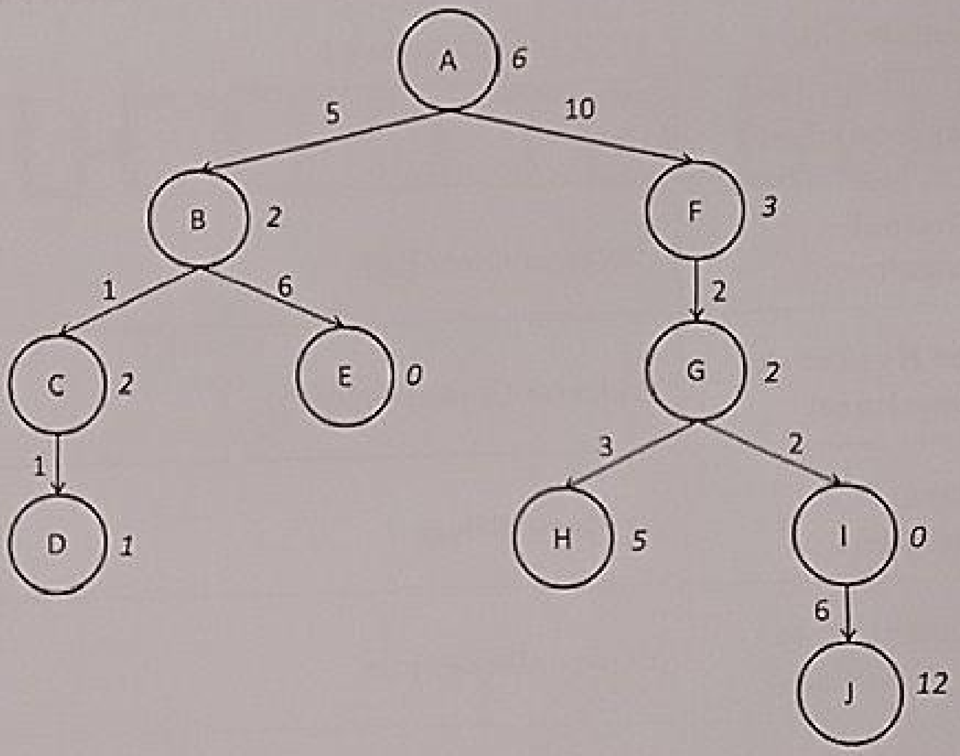
a. The terms *state space* and *search tree* are synonymous (i.e. they mean the same).

b. A goal state in a state space always has no successors.

c. A goal node in a search tree is always a leaf node.

d. At the point when state space search terminates, the agenda will always be empty.

ii) *(12 marks)* Consider the following state space in which the states are shown as nodes labeled A through J. A is the initial state and E and I are goal states. The numbers alongside the edges represent the costs of moving between the states. To the right of every state is the estimated cost of the path from the state to the nearest goal.



Show how each of the following search strategies finds a solution in this state space by writing down, in order, the names of the nodes removed from the agenda. Assume the search halts when the goal state is removed from the agenda. (In some cases, multiple answers are possible. You need give only one such answer in each case.)

a. *(1 mark)* Breadth-first;

b. *(1 mark)* Depth-first;

c. *(3 marks)* Least-cost search;

d. *(3 marks)* Greedy search, i.e. heuristic search using $f(n) = h(n)$ as the evaluation function, where $h(n)$ is the estimated cost of the cheapest path from node n to a goal; and

e. *(4 marks)* Heuristic search using $f(n) = g(n) + h(n)$ as the evaluation function, where $g(n)$ is the cost of the path to node n, and $h(n)$ is the estimated cost of the path from node n to the nearest goal.

2. (*20 marks*) Consider the following unlabeled dataset with $m = 5$ examples and $n = 3$ features:

$$\mathbf{x}^{(1)} = \begin{bmatrix} 3 \\ 3 \\ 4 \end{bmatrix} \quad \mathbf{x}^{(2)} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \quad \mathbf{x}^{(3)} = \begin{bmatrix} 2 \\ 3 \\ 3 \end{bmatrix} \quad \mathbf{x}^{(4)} = \begin{bmatrix} 0 \\ 3 \\ 2 \end{bmatrix} \quad \mathbf{x}^{(5)} = \begin{bmatrix} 0 \\ 2 \\ 2 \end{bmatrix}$$

The all-pairs Euclidean distances are given in the following table:

	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	$\mathbf{x}^{(3)}$	$\mathbf{x}^{(4)}$	$\mathbf{x}^{(5)}$
$\mathbf{x}^{(1)}$	$\sqrt{0}$	$\sqrt{14}$	$\sqrt{2}$	$\sqrt{13}$	$\sqrt{14}$
$\mathbf{x}^{(2)}$	$\sqrt{14}$	$\sqrt{0}$	$\sqrt{6}$	$\sqrt{3}$	$\sqrt{2}$
$\mathbf{x}^{(3)}$	$\sqrt{2}$	$\sqrt{6}$	$\sqrt{0}$	$\sqrt{5}$	$\sqrt{6}$
$\mathbf{x}^{(4)}$	$\sqrt{13}$	$\sqrt{3}$	$\sqrt{5}$	$\sqrt{0}$	$\sqrt{1}$
$\mathbf{x}^{(5)}$	$\sqrt{14}$	$\sqrt{2}$	$\sqrt{6}$	$\sqrt{1}$	$\sqrt{0}$

i) (*12 marks*) The k-means clustering algorithm is to be applied to this dataset with $k = 2$. The initial centroids are $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(3)}$

a. (*4 marks*) *Assignment step:* give the members of the two clusters after the *assignment step* is run for the first time.

b. (*4 marks*) *Update step:* Give the centroids of the two clusters that you just found.

c. (*2 marks*) Give two problems that can arise when the initial centroids for k-means are chosen at random.

d. (*2 marks*) Describe an alternative to choosing the initial centroids at random that might overcome these problems.

ii) (*8 marks*) Agglomerative clustering is to be applied to the same dataset. Suppose that, after a couple of iterations, the algorithm has created the following three clusters:

$$C_1 = \{\mathbf{x}^{(1)}\} \; C_2 = \{\mathbf{x}^{(2)}, \mathbf{x}^{(3)}\} \; C_3 = \{\mathbf{x}^{(4)}, \mathbf{x}^{(5)}\}$$

a. (*4 marks*) If the algorithm is using *complete-linkage*, which two of these three clusters will be merged? For credit, show all your working.

b. (*4 marks*) If instead the algorithm is using *average-linkage*, which two of these three clusters will be merged? For credit, show all your working.

3. (*20 marks*)

i) (*3 marks*) Give a real-life domain in which *regression* might be useful. Describe the target and three of the features, one of which should be numeric, one should be binary-valued, and one should be nominal with more than two values.

ii) (*3 marks*) Give a real-life domain, different from the one you gave in part (i), in which (non-binary) *multiclass classification* might be useful. Describe the class labels and three of the features, one of which should be numeric, one should be binary-valued, and one should be nominal with more than two values.

iii) (*8 marks*) We run Stochastic Gradient Descent twice on the same training set, obtaining two different linear models: in one case $\beta = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}$, and in the other case $\beta = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix}$. Consider the following *test set*:

$$\mathbf{X} = \begin{bmatrix} 1 & 3 & 2 \\ 1 & 3 & 1 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 20 \\ 10 \end{bmatrix}$$

Calculate the *mean absolute error* for both models and say which is better. For credit, show your working.

iv) (*6 marks*) We learned three variants of Gradient Descent: Batch, Stochastic and Mini-Batch. In the case of ordinary least squares regression:

a. Which will reach the vicinity of the global minimum fastest?

b. Which will actually converge?

c. How can you improve the likelihood that the others converge?

For credit, explain your answers.

4. *(20 marks)*

i) *(6 marks)* The following dataset has one nominal-valued feature, *dish*, which has values beef, chicken and lamb, and a class label, *wine*, which has values red and white:

dish	*wine*
beef	red
beef	red
chicken	white
chicken	red
lamb	red

Show what the dataset will look like after you have prepared it for a logistic regression classifier that only allows numeric values for both the feature and the dependent variable.

ii) *(5 marks)* Recall that hypotheses in logistic regression are represented as

$$h_\beta(\mathbf{x}) = s(\mathbf{x}\beta)$$

where

$$s(z) = \frac{1}{1+e^{-z}}$$

Suppose we collect data for a group of students in a programming module with features $\mathbf{x}_1$ being hours studied, $\mathbf{x}_2$ being average grade so far, and y being whether they receive an A (positive class) or not (negative class). We fit a model by logistic regression and produce estimated coefficients $\beta_0 = -30, \beta_1 = 0.5, \beta_2 = 0.1$.

a. *(3 marks)* Consider a student $\mathbf{x}$ who studies for 30 hours and has an average grade so far of 60. Calculate $s(h_\beta(\mathbf{x}))$. For credit, show your working.

b. *(2 marks)* What does the number that you calculated in part (i) mean?

iii) *(9 marks)* Below is the confusion matrix for a binary classifier, A:

		Predicted Class	
		0	1
Actual	0	25	10
Class	1	20	45

a. *(2 marks)* What is its accuracy?

b. *(2 marks)* How many True Positives are there?

c. *(2 marks)* How many False Positives are there?

d. *(3 marks)* Below is the confusion matrix for another binary classifier, B:

		Predicted Class	
		0	1
Actual	0	45	20
Class	1	10	25

Explain the considerations you would take into account if you had to choose between classifiers A and B.