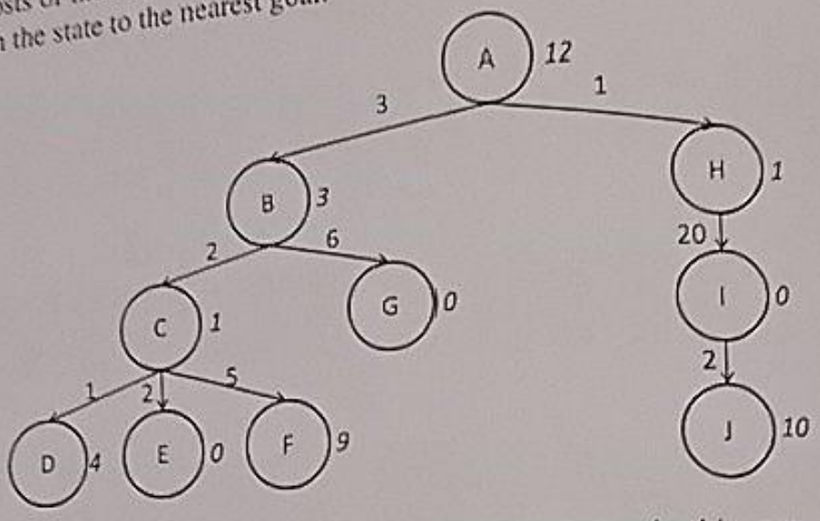
1. (20 marks)

i) (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 G 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.

ii) (*4 marks*) To be *complete* in general, least-cost search requires costs to be positive.

Carefully explain why zero or negative costs might result in it not being complete.

iii) (*4 marks*) To be *optimal* in general, least-cost search requires costs to be non-negative.

Carefully explain why negative costs might result in it not being optimal.

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

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

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

	$\mathbf{x}^{(1)}$	$\mathbf{x}^{(2)}$	$\mathbf{x}^{(3)}$	$\mathbf{x}^{(4)}$	$\mathbf{x}^{(5)}$
$\mathbf{x}^{(1)}$	$\sqrt{0}$	$\sqrt{11}$	$\sqrt{19}$	$\sqrt{16}$	$\sqrt{17}$
$\mathbf{x}^{(2)}$	$\sqrt{11}$	$\sqrt{0}$	$\sqrt{8}$	$\sqrt{3}$	$\sqrt{2}$
$\mathbf{x}^{(3)}$	$\sqrt{19}$	$\sqrt{8}$	$\sqrt{0}$	$\sqrt{11}$	$\sqrt{6}$
$\mathbf{x}^{(4)}$	$\sqrt{16}$	$\sqrt{3}$	$\sqrt{11}$	$\sqrt{0}$	$\sqrt{1}$
$\mathbf{x}^{(5)}$	$\sqrt{17}$	$\sqrt{2}$	$\sqrt{6}$	$\sqrt{1}$	$\sqrt{0}$

i) (*10 marks*) The k-means clustering algorithm is to be applied to this dataset with $k = 2$. The initial centroids are $\mathbf{x}^{(2)}$ and $\mathbf{x}^{(5)}$

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*) Explain why k-means clustering is sensitive to *outliers*.

ii) (*10 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.

c. (*2 marks*) Explain which strategy (complete-linkage or average-linkage) is most sensitive to outliers.

3. (*20 marks*)

i) (*2 marks*) Describe how a labeled training set differs from an unlabeled one.

ii) (*2 marks*) Describe how regression differs from classification.

iii) (*8 marks*) Consider the following labeled training set:

$$\mathbf{X} = \begin{bmatrix} 1 & 3 & 1 \\ 1 & 1 & 1 \\ 1 & 3 & 0 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 9 \\ 8 \\ 8 \end{bmatrix}$$

A Gradient Descent algorithm initializes β randomly to $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$. Calculate the *loss* for ordinary least squares regression. For credit, show all working.

iv) (*3 marks*) We learned four ways to fit an ordinary least squares regression model: the Normal Equation, Batch Gradient Descent, Stochastic Gradient Descent and Mini-Batch Gradient Descent. Which of the four would you **not** use if you have a training set with a million features? Explain why not.

v) (*5 marks*) We learned three variants of Gradient Descent: Batch, Stochastic and Mini-Batch. In the case of ordinary least squares regression, do all three reach the same final model provided the learning rate is not too high and you run them long enough? Explain your answer.

4. (*20 marks*)

i) (*6 marks*) The following dataset has one nominal-valued feature, *sex*, which has values female and male, and a class label, *degree*, which has values BSc, MSc and PhD:

sex	*degree*
male	BSc
female	MSc
female	MSc
female	PhD
male	PhD

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 class label.

ii) (*6 marks*) This question concerns experiment methodologies for classification.

a. (*2 marks*) Describe stratified k-fold cross-validation.

b. (*4 marks*) Give the advantages and disadvantages of k-fold cross-validation relative to stratified k-fold cross-validation.

iii) (*8 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 receive an A. We fit a model by logistic regression and produce estimated coefficients $\beta_0 = -30$, $\beta_1 = 1$, $\beta_2 = 0.2$.

a. (*3 marks*) What, according to the model, is the probability that a student who studies for 10 hours and has an average grade so far of 80 gets an A?

b. (*5 marks*) How many hours would a student who has an average grade so far of 80 need to study to have a 75% chance of getting an A?

In both cases, for credit, show your working.