Coursework for Introduction to AI

This coursework assesses your fundamental understanding of the algorithms and methods that you will have encountered in the unit. Q1 to Q3 concern machine learning, while Q4 to Q6 concern search, knowledge representation and reasoning, and Markov decision processes. You should submit the coursework in the form of a written report addressing each of the questions. Overall the report should not be more than 10 pages and should be submitted via blackboard. The deadline is 29th May at 13:00.

Q1: K-Nearest Neighbours and Naive Bayes

Q1(a)

Explain why k-nearest neighbours tends to overfit the training data when $k = 1$ and underfit it when k is equal to the number of training data points.
4 marks

Q1(b)

Consider applying a k-nearest neighbour regression algorithm to estimate a 1-D function f on the basis data of the form (x, y) where $y = f(x)$ so if $x_1, \ldots, x_k$ are the k nearest values to x then;

$$\hat{f}(x) = \frac{\sum_{i=1}^{k} w_i y_i}{\sum_{i=1}^{k} w_i} \text{ where } w_i = \frac{1}{||x - x_i||}$$

Suppose you have the following training data $\{(2, 4), (5, 25), (7, 49)\}$. Write down expressions for $\hat{f}$ as a piecewise function of x for $k = 1, 2$ and 3. If the true function is $f(x) = x^2$ then what is the optimal value for k in this case?
10 marks

	x_1	x_2	x_3	x_4
x_1	0	0.1	0.5	0.8
x_2	0.1	0	0.15	0.7
x_3	0.5	0.15	0	0.2
x_4	0.8	0.7	0.2	0

Table 1: Table of distances for Q2

Q1(c)

For a binary classification problem with classes c_1 and c_2 and suppose that n_1 out of N training data points are of class c_1. By assuming a uniform prior distribution on the probability of C_1, derive the following Laplace's rule estimate of the probability of c_1 given the training data.

$$P(c_1) = \frac{n_1 + 1}{N + 2}$$

Note: You can assume that for x and y natural numbers,

$$\int_0^1 p^x (1-p)^y dp = \frac{x!y!}{(x+y+1)!}$$

6 marks

Q2: Clustering and Distance

Q2(a)

Consider data element $\{x_1, x_2, x_3, x_4\}$ with the distances between them as shown in table 1. Assume that distance is extended from elements to sets of elements using max so that $d(S,T) = \max\{d(x,y) : x \in S, y \in T\}$ then perform hierarchical clustering on these four elements. Repeat, but this time using min so that $d(S,T) = \min\{d(x,y) : x \in S, y \in T\}$. What do you think are the natural clusters in both cases?

5 marks

Q2(b)

For the feature space Ω, let $d : \Omega \times \Omega \to \mathbb{R}^+$ be a distance metric. Assume that d is then extended to sets of elements such that for $S, T \subseteq \Omega$;

$$d(S,T) = \max\{d(x,y) : x \in S, y \in T\}$$

Suppose we now apply d in a bottom up (agglomerative) hierarchical clustering algorithm and at a certain level there are three clusters C_1, C_2 and C_3. The closest pair of clusters are C_1 and C_2 which are then merged so that at the next level we have two cluster C_3 and $C_1 \cup C_2$. Show that in this case

$$d(C_3, C_1 \cup C_2) \geq d(C_1, C_2)$$

Comment on why this property is important when plotting a dendrogram.
5 marks

Q3: Linear Regression and Decision Trees

Q3(a)

Given a data set of N elements of the form (x, y) where $x, y \in \mathbb{R}$ show that the least squared model of the form $y = ax^2 + b$ has the parameters;

$$a = \frac{\overline{yx^2} - \overline{y}\,\overline{x^2}}{\overline{x^4} - (\overline{x^2})^2} \text{ and } b = \overline{y} - a\overline{x^2}$$

where,

$$\overline{y} = \frac{1}{N}\sum_{(x,y)} y,\ \overline{x^2} = \frac{1}{N}\sum_{(x,y)} x^2,\ \overline{yx^2} = \frac{1}{N}\sum_{(x,y)} yx^2 \text{ and } \overline{x^4} = \frac{1}{N}\sum_{(x,y)} x^4$$

8 marks

Q3(b)

Use the least squared parameters given in part (a) to fit a model of the form $y = ax^2 + b$ to the following data $\{(0, 1), (5, 23), (7, 40), (10, 90)\}$.
4 marks

Q3(c)

Consider a classification problem with k classes $c_1, \ldots, c_k$. For a data set D, let p_i denote the proportion of elements in D labelled as class c_i then the Gini impurity of D is given by;

$$G(D) = 1 - \sum_{i=1}^{k} p_i^2$$

Let D_1 and D_2 form a partition of D so that $D_1 \cup D_2 = D$ and $D_1 \cap D_2 = \emptyset$. Show that;

$$G(D) \geq \frac{|D_1|}{|D|} G(D_1) + \frac{|D_2|}{|D|} G(D_2)$$

Explain the significance of this property when using Gini impurity in decision trees.
8 marks

Q4: Markov Decision Processes

Q4(a)

In a Markov decision process (MDP), what is meant by

- a policy
- an optimal policy
- the expected utility of a policy

3 marks

Q4(b)

Consider a robot in a gridworld trying to reach a goal and avoid obstacles. It can move north, south, east, or west in the grid, but occasionally fails to move in the intended direction. If you modelled this using an MDP, would you use value iteration or policy iteration to solve it optimally? Justify your answer in a sentence (or two).
3 marks

Q4(c)

Now suppose the robot can call on a helicopter (or teleport) to move it to any square, S, on the grid, knowing that it might instead land in a square adjacent to S with a specified probability. Would this change your choice? Again, justify your answer in a sentence (or two).
3 marks

wall	-100	-100	-100	-100	-100	wall
1	0	0	0	0	0	10
wall	-100	-100	-100	-100	-100	wall

wall	-100	-100	-100	-100	-100	wall
1	-17.28	-30.44	-36.56	-25.78	-10.8	10
	←	←	→	→	→	
wall	-100	-100	-100	-100	-100	wall

Figure 1: (a) rewards for the bridge-crossing problem in gridworld. (b) utilities after 100 iterations, and the corresponding optimal policy

Q4(d)

Figure 1(a) shows a narrow bridge represented as a gridworld environment. A robot starts at the left hand side, in the middle row (marked with a reward of 1). The goal is the middle row on the right hand side, marked with a reward of 10. Squares marked with a reward of -100 are terminal nodes, and represent the robot falling off the bridge. The robot can move one square up, down, left or right. When told to move in a specified direction, it moves in the intended direction with probability 0.8 or at 90 degrees to the intended direction with probability 0.1, or at -90 degrees to the intended direction with probability 0.1. With a discount value of 0.9, use value iteration to calculate the utility of each non-terminal grid square after one and two iterations.

6 marks

Q4(e)

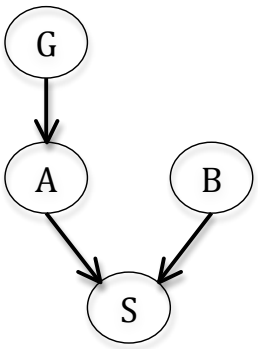
The problem in Q4(d) leads to the optimal policy shown in Figure 1(b), which fails to cross the bridge. What would be the effect of increasing the discount value?

5 marks

Q4(f)

What would be the effect on the policy of increasing the utility of the goal ? Choose a new value for the utility of the goal state so that the optimal policy is to cross the bridge from left to right, and show the utility of each non-terminal grid square after 3 iterations.

8 marks



$P(G)$	
g	0.1
$\neg g$	0.9

$P(B)$	
b	0.4
$\neg b$	0.6

$P(A\|G)$		
g	a	1.0
g	$\neg a$	0.0
$\neg g$	a	0.1
$\neg g$	$\neg a$	0.9

$P(S\|A,B)$			
a	b	s	1.0
a	b	$\neg s$	0.0
a	$\neg b$	s	0.9
a	$\neg b$	$\neg s$	0.1
$\neg a$	b	s	0.8
$\neg a$	b	$\neg s$	0.2
$\neg a$	$\neg b$	s	0.1
$\neg a$	$\neg b$	$\neg s$	0.9

Figure 2: Bayes net and probability tables for question Q5

Q5: Bayes Networks and Knowledge Representation

A patient can have a symptom, S, that is caused by two different diseases, A and B. It is known that the presence of a gene G is important in the manifestation of disease A. The Bayes net and conditional probability tables are shown in figure 2.

Q5(a)

What is the probability that a patient has disease A
3 marks

Q5(b)

What is the probability that a patient has disease A if we know that the patient has disease B
3 marks

Q5(c)

What is the probability that a patient has disease A if we know that the patient has disease B AND symptom S
3 marks

Q6: Search

Consider the 8-puzzle problem, given the initial state shown in figure 3.

Q6(a)

Using breadth first search, show the search tree that would be built down to level 2 (assume level zero is the root of the tree).

Assume that the *blank* moves left if possible, up (if left is not possible), right (if up and left are not possible), else down.
3 marks

Q6(b)

Using depth first search, show the state of the search tree down to level 3 (stop once you have expanded one node that goes to level 3).
3 marks

Q6(c)

Suggest two heuristic functions that could estimate the distance of an 8-puzzle state from the goal state, and evaluate the functions on the initial state shown.
3 marks

Q6(d)

Briefly outline the relative advantages and disadvantages of breadth, depth and heuristic search.
4 marks

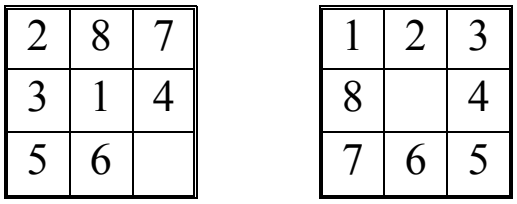


Figure 3: 8-puzzle initial state (left) and goal state (right)