1. Consider the following multi-layer perceptron network. Assume that the neurons have a sigmoid activation function, perform a forward pass and a backword pass on the network. Assume that the actual output of y is 1 and learning rate is 0.9. Perform another forward pass.

→ Activation function : sigmoid
Actual output $y = 1$
Learning Rate $\eta = 0.9$

→ Net input to H_4,

$$H_4 = 1(0.2) + 0(0.4) + 1(-0.5) + (-0.4)$$
$$= -0.7$$
$$H_4 = \frac{1}{1 + e^{-(-0.7)}} = 0.3318122$$

→ Net input to H_5

$$H_5 = 1(-0.3) + 0(0.1) + 1(0.2) + 0.2$$
$$= 0.1$$
$$H_5 = \frac{1}{1 + e^{-(0.1)}} = 0.5249792$$

→ Net input to O_6

$$O_6 = 0.3318122(-0.3) + 0.5249792(-0.2) + 0.1$$
$$= -0.1045395$$
$$O_6 = \frac{1}{1 + e^{-(-0.1045395)}} = 0.473889$$

$$\delta_6 = O_6(1 - O_6)(1 - O_6) = 0.473889 = 0.1311690$$

$$\Delta w_{46} = \eta \cdot \delta_6 \cdot H_4 = 0.9 \times 0.474 \times 0.332 = 0.1416$$
$$\Delta w_{56} = \eta \cdot \delta_6 \cdot H_5 = 0.9 \times 0.474 \times 0.525 = 0.223965$$

$\Delta\theta_6 = \eta \cdot \delta_8 = 0.9 \times 0.1312 = 0.11181$

$\Delta w_{46} = \eta \cdot \delta_8 \cdot H_4 = 0.9 \times 0.1312 \times 0.332 = 0.0392$

$\Delta w_{56} = \eta \cdot \delta_8 \cdot H_5 = 0.9 \times 0.1312 \times 0.522 = 0.061992$

$\delta_4 = H_4(1 - H_4)(\delta_6 \cdot w_{46})$

$= 0.3318(1 - 0.3318)(0.1312 \cdot 0.0392)$

$= 0.00114$

$\delta_5 = H_5(1 - H_5)(\delta_6 \cdot w_{56})$

$= 0.525(1 - 0.525)(0.131 \times 0.062)$

$= 0.002025$

$w_{46} = -0.3 + 0.0392 = -0.2608$

$w_{56} = -0.2 + 0.061992 = -0.138$

$\theta_6 = 0.1 + 0.11181 = 0.2181$

$\Delta w_{14} = \eta \cdot \delta_4 \cdot x_1 = 0.9(0.00114)(1) = 0.001026$

$\Delta w_{24} = \eta \cdot \delta_4 \cdot x_2 = 0$

$\Delta w_{34} = \eta \cdot \delta_4 \cdot x_3 = 0.9(0.001414)(1) = 0.001026$

$\Delta w_{15} = \eta \cdot \delta_5 \cdot x_1 = 0.9 \times 0.002025 \times 1 = 0.0018225$

$\Delta w_{25} = \eta \cdot \delta_5 \cdot x_2 = 0$

$\Delta w_{35} = \eta \cdot \delta_5 \cdot x_3 = 0.9 \times 0.002025 = 0.0018225$

$w_{14} = 0.2 + 0.001026 = 0.201026$

$w_{24} = 0.4 + 0 = 0.4$

$w_{34} = -0.5 + 0.001026 = -0.498974$

$w_{24}^{15} = -0.3 + 0.0018225 = -0.2981775$

$w_{25} = 0.1$

$w_{35} = 0.2 + 0.0018225 = 0.2018225$

$H_4 = x_1 \times w_{14} + x_2 \times w_{24} + x_3 \times w_{34}$ =

$= 1(0.201026) + 0(0.4) + 1(-0.498974)$

$= -0.297948$

$H_4 = \frac{1}{1+e^{-(-0.297948)}} = 0.4260592$

$H_5 = x_1 \times w_{15} + x_2 \times w_{25} + x_3 \times w_{35}$

$= 1(-0.2981775) + 0 + 1(0.2018225)$

$= -0.096355$

$H_5 = \frac{1}{1+e^{-(-0.096355)}} = 0.4759299$

$O_6 = H_4 \times w_{46} + H_5 \times w_{56} + \theta_6$

$= 0.4260592 \times (-0.2608) + 0.4759299(-0.138) + 0.22$

$= 0.04320$

$O_6 = \frac{1}{1+e^{-(0.0432)}} = 0.5107997$

2. Apply K-nearest neighbour classifier on below mentioned fitness dataset and classify the test record based on the values of input feature.

Row	Height	weight	class	Distance
1	167	51	underweight	$\sqrt{(170-167)^2+(57-51)^2} = 6.71$
2	182	62	Normal	$\sqrt{(144+25)} = 13.0$
3	176	69	Normal	$\sqrt{(36+144)} = 13.42$
4	173	64	Normal	$\sqrt{(9+49)} = 7.62$
5	172	65	Normal	$\sqrt{(4+64)} = 8.25$
6	174	56	Underweight	$\sqrt{(16+1)} = 4.12$
7	169	58	Normal	$\sqrt{(1+1)} = 1.41$
8	173	57	Normal	$\sqrt{(9+0)} = 3$
9	170	55	Normal	$\sqrt{(0+4)} = 2$

K = 3

→ Pick three nearest neighbour

Row 7 → Distance = 1.41 → Normal

Row 8 → Distance = 2.00 → Normal

Row 9 → Distance = 3.00 → Normal

∴ (170, 57) → Normal

3. Define support vector machine. Discuss its advantages and application in brief

→ A support vector machine (SVM) is supervised machine learning algorithm used for classification and regression task. The main idea behined SVM is to find the optional hyperplane that best spepraates the data points of different classes in a high-dimensional space.

⇒ Advantages of SVM:

→ Effective in high dimensional space.

→ Memory efficient

→ Versatile.

→ Robust to overfitting

⇒ Applications of SVM:

→ Text and hypertext classification

→ Image classification.

→ Bioinformatics

→ Financial applications.

→ Medical diagnosis

4 Explain SVM classifier algorithm for linearly seperable data. Use diagram to support your discussion.

=> steps of algoritm:

-> Input label training data (x_i, y_i) where $y_i \in \{-1, 1\}$

-> Find the hyperplane $w \cdot x + b = 0$ that seprates the classes.

-> Maximise the margin between the support vectors.

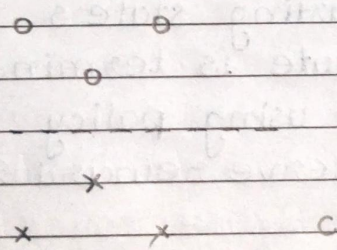
-> Use tohe optimal w and b to classify neco data point

-> Decision Rule:

If $w \cdot a + b \geq 0$ => class +1

Else => class -1

=> Diagram for SVM with linearly spparable data:



5 Discuss the working of Q-learning algorithm in reinforcement learning.

-> Q-learning is a model-free reinforcement learning algorithm that helps an agent learn the optimal action-selection policy by learning the Q-values of state-action pairs. It does not require a model of the environment.

⇒ Q-Learning Intuition :

→ The agent interacts with the environment in discrete time steps.

→ At each time step :

→ The agent observes the current state(s).

→ Choose an action(a) based on some policy

→ Receives a reward(r) and observes the next state(s').

→ Updates the Q-values of the pair using the Q-Learning formula.

⇒ Formula :

$$Q(s,a) \leftarrow Q(s,a) + \alpha [r + \gamma \max_{a'} Q(s',a') - Q(s,a)]$$

⇒ Steps of the Q-Learning Algorithm :

1. Initialize Q-table with zeros for all state-action pairs.
2. For each episode :
   1. Initialize the starting state s
   2. Repeat until the state is terminal :
      → Choose action a using policy
      → Take action a, observe reward r and new state s'
      → Update Q(s,a) using the update rule.
      → set s = s'
6. Discuss the working
6. state and explain the inductive learning hypothesis.

→ Any hypothesis found to approximate the target function well over a sufficiently large train

dataset will also perform well over unseen data from the same distribution.

=> working :

-> You collect a traing dataset

-> You train a model using dataset.

-> You evaluate the model's performance on new data.

-> The inductive learning hypothesis allows you to expect that good training performance indicate good performance on similar, unseen data.

=> Example :

-> Train a spam classifier on 10,000 emails labeled as "spam" or "not spam".

=> Formal ~~signature~~ significance :

-> Evaluating model performance

-> Avoiding overfitting.

-> Developing generalization strategies.

7. what is occam's Razor with respect to hypothesis search ?

-> Among competing hypothesis that explain the observed data quality well, the simplest one should be prefferred.

-> when multiple hypotheses fit the traing data well, occam's Razor suggests that the simplest hypothesis ~~fever~~ is more likely to generalize better to unseen data

=> Benifits :

-> Improve generalization

-> Reduce overfitting.

-> Encouruges interpretability and efficiency.

8. Explain the following terms: (i) restriction bias (ii) Performance bias. what is the preference bias in decision tree?

=> (i) Restriction Bias (also called language bias)

-> Restriction bias refers to the limitations or constraints on the set of hypotheses that a learning algorithm can consider.

=> (ii) Preference Bias (also called search bias).

-> Preference bias is a soft bias where the algorithm has a preference for some hypotheses over others, even if multiple hypotheses fit the data.

=> Preference Bias in Decision Trees:

1. Information gain or gini index:

-> The algorithm prefers attributes that split the data in a way that maximize information gain.

2. Three simplicity:

-> Some implementations may prefer shallower trees or fewer splits to avoid overfitting.

-> Pruning is used to remove parts of the tree that do not provide much predictive power, reflecting a bias toward simpler trees.

9. what is overfitting of an ML model? How can you detect whether overfitting has occurred in a decision tree?

-> Overfitting occurs when a machine learning model learns that training data to well,

including noise and irrelevent patterns, resulting in poor generalization to new, unseen data.

=> How to detect overfitting in decision Trees:

1. Performance crap:

-> Train Accuracy >> Test Accuracy.

-> Large difference between training and validation/test accuracy.

2. Complex Tree structure:

-> The tree has too many nodes or depth.

-> Leaf nodes with very few samples.

3. Cross-validation Results:

-> Significant variance in accuracy accross fold may indicate overfitting

4. visual Inspection:

-> The tree is Too specific to the training data

10 Explain the advantage and disadvantage of using the following attribute selection criteria for decision trees: (i) information gain (ii) Gain Ratio.

=> (i) Information gain:

-> Information gain measures the reduction in entropy achieved by splitting the data based on a particular attribute.

$$\text{Information Gain} = \text{Entropy(parent)} - \sum \left[\frac{|\text{Subset i}|}{|\text{Total}|} \times \text{Entropy(subset i)} \right]$$

=> Advantage:

1. simple and effective.

-> Easy to compute and interpret.

-> Works well with categorical attribute

e) Used in ID3 Algorithm:

→ A core component of the ID3 decision tree algorithm.

→ Quick indentifices attributes that provide the most immediate "purity" improvement.

⇒ Disadvantage:

1. Bias toword many-valued Attributes
2. overfitting on training data.

(iii) Gain Ratio:

→ Gain Ratio is normalized version of information gain that penalize attributes with many values.

Gain Ration = $\frac{\text{Information Gain}}{\text{Split Information}}$

→ Split information measures how broadly and uniformly the data is split.

⇒ Advantages:

→ Reduce Bias.

→ Better generalization

→ Used in C4.5 Algorithm.

⇒ Disadvantages:

→ Can be unstable

→ Computationaly more expensive.

11 consider the following set of training example

Instance	classification	a_1	a_2
1	+	T	T
2	+	T	T
3	-	T	F
4	+	F	F
5	-	F	T
6	-	F	T

a. what is the entropy of this collection of training samples with respect to the target function classification ?

→ Positive (+) : 3

Negative (-) : 3

$$\text{Entropy}(S) = -p_+ \log_2(p_+) - p_- \log(p_-)$$

$$p_+ = \frac{3}{6} = 0.5 \qquad p_- = \frac{3}{6} = 0.5$$

$$\text{Entropy}(S) = -0.5 \log_2(0.5) - 0.5 \log_2(0.5) = 1$$

(b) what is the information gain of a_2 relative to these training example.

→ split data based on a_2

$a_2 = T$: Instance 1,2,5,6 → classes : +, +, -, -

$a_2 = F$: Instance 3,4 → Classes : -, +

→ Entropy for $a_2 = T$

$$\text{Entropy}_T = -0.5 \log_2(0.5) - 0.5 \log_2(0.5) = 1$$

→ Entropy for $a_2 = F$

$$\text{Entropy}_F = -0.5 \log_2(0.5) - 0.5 \log_2(0.5) = 1$$

→ weighted Entropy after split

$$\text{Entropy}_{a_2} = \frac{4}{6} \cdot 1 + \frac{2}{6} \cdot 1 = \frac{4}{6} + \frac{2}{6} = 1$$

→ Information Gain

Gain(U2) = Entropy(S) - Entropy(U2)

= 1 - 1

= 0

→ Information gain of U2 = 0