A1) (0,0): $z = w_0x_0 + w_1x_1 + w_2x_2 = -1.5(1) + 1(0) + 1(0) = 1.5$.
Output is -1.

(1,0): $z = w_0x_0 + w_1x_1 + w_2x_2 = -1.5(1) + 1(1) + 1(0) = -0.5$
Output is -1.

(0,1): $z = w_0x_0 + w_1x_1 + w_2x_2 = -1.5(1) + 1(0) + 1(1) = -0.5$
Output is -1.

(1,1): $z = w_0x_0 + w_1x_1 + w_2x_2 = -1.5(1) + 1(1) + 1(1) = 0.5$
Output is 1.

A2) AND: $x_1 + x_2 - 1$

NOT: $2x_1 + 2x_2 - 1$

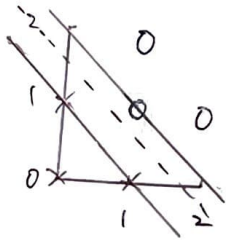
NAND: $-x_1 - x_2 + 2$

NOR: $-x_1 - x_2 + 1$

A3) No, it is not possible to classify it using a single neuron because it acts as a XNOR gate. In a single layer perceptron, implementing the XNOR operation is impossible.

A4)

O: class 1 X: class 2



(Also, we can maximize the margin of the classifier using the support vectors.)

The support vectors are (1,0), (0,1), (1,1).
Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane.

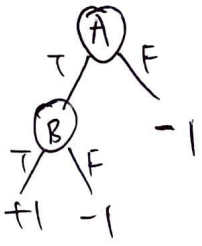
A5) The entropy

$$= -\left[0.5 \log_2(0.5) + 4 \times 0.125 \log_2(0.125)\right]$$

$$= 2$$

Entropy represents the machine learning metric which measures the unpredictability or impurity. It is related to randomness in the information being processed.

A6) e.g. A and B



For perceptron solution, it is necessary to design a decision function and assign weights to the features. Also, all features are used to do the calculation. However, decision tree is built by using one feature in each splitting. The decision tree solution is simpler since we only need to decide true/false actions and finally get the result.

A7)

H	H	E	A
S	B	Br	N
T	D	Br	N
T	B	Bl	Y
S	D	Bl	N
T	R	Bl	Y
T	B	Br	N
S	B	Bl	Y

Gini impurity $= 1 - \left[\left(\frac{3}{8}\right)^2 + \left(\frac{5}{8}\right)^2\right]$

eyes = Blue? $= 0.47$

T F

H	H	E	A
T	B	Bl	Y
T	D	Bl	N
S	D	Bl	N
T	R	Bl	Y
S	B	Bl	Y

Gini impurity $= 1 - \left[\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right]$

$= 0.48$

H	H	E	A
S	B	Br	N
T	D	Br	N
T	B	Br	N

Gini impurity $= 1 - \left[\left(\frac{3}{3}\right)^2\right]$

$= 0$

predict not attractive look

Height = Tall?

T F

H	H	E	A
T	B	Bl	Y
T	D	Bl	N
T	R	Bl	Y

Gini impurity $= 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right]$

$= 0.44$

H	H	E	A
S	D	Bl	N
S	B	Bl	Y

Gini impurity $= 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right]$

$= 0.5$

Hair = Dark? = R

Hair = Potic?

H	H	E	A
T	D	Bl	N

Gini impurity = 0

not attractive look

Gini impurity = 0

H	H	E	A
T	B	Bl	Y
T	R	Bl	Y

attractive look

T

H	H	E	A
S	D	Bl	N

Gini impurity = 0

predict not attractive look

H	H	E	A
S	B	Bl	Y

Gini impurity = 0

predict attractive look

A8) 1. $z = w_0x_0 + w_1x_1 + w_2x_2 = -6(1) + 0.05(40) + 1(3.5)$

$= -0.5$

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

$$\phi(-0.5) = \frac{1}{1+e^{-(-0.5)}}$$

$$= 0.378$$

2. $\phi(z) = 0.5$

$$\frac{1}{1+e^{-z}} = 0.5$$

$$z = 0$$

$$w_0x_0 + w_1x_1 + w_2x_2 = 0$$

$$-6(1) + 0.05(x_1) + 1(3.5) = 0$$

$$x_1 = 50$$

∴ 50 hours are required

A9) In this case (kNN with $k=1$), the training error is 0%. We have $P(Y=j \mid X=x_i) = I(y_i=j)$ which is equal to 1 if $y_i=j$ and 0 if not. There is no error made on the training data so the traing error rate is 0%, which means that the test error rate is 36%. On the other hand, the test error rate for logistic regression is 30%. Therefore, it is better to choose logistic regression due to lower test error rate.

A10) The Gradient Descent algorithms might suffer from this, because the cost function will have the shape of an elongated bowl. In other words, the Gradient Descent algorithms will take a long time to converge. To solve this problem, standardization is required, which aims to scale the data before the model is trained.

A11) We can increase the number of estimators or reduce the regularization hyperparameters of the base estimators. Also, increasing the learning rate is necessary.

A12) With out-of-bag evaluation, each predictor in a bagging ensemble is evaluated using instances which was not trained on. Therefore, we can have a fairly unbiased evaluation of the ensemble without an additional validation set. Having more instances for training, we can make the ensemble better.

A13) Hard voting classifier classifies data based on class labels and the weights associated with each classifier, while soft voting classifier relies on the probabilities and the weights. In hard voting, every individual classifier votes for a class and the majority wins. In soft voting, the target label with the greatest sum of weighted probabilities wins the vote