A1 $\phi(w^T x) = \begin{cases} 1, & \text{if } w^T x \geq 0 \\ 0, & \text{otherwise.} \end{cases}$

$$w = \begin{pmatrix} 1 \\ 1 \\ -1.5 \end{pmatrix}$$

For (0, 0):
$x = (0, 0, 1)^T$
$\phi(w^T x) = \phi(-1.5) = 0$
the output of (0,0) is 0.

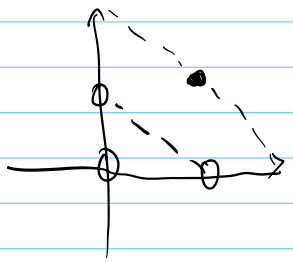
For (1,0):
$x = (1, 0, 1)^T$
$\phi(w^T x) = \phi(-0.5) = 0$
the output of (1,0) is 0.

For (0,1):
$x = (0, 1, 1)^T$
$\phi(w^T x) = \phi(-0.5) = 0$
the output of (0,1) is 0

For (1,1):
$x = (1, 1, 1)^T$
$\phi(w^T x) = \phi(0.5) = 1$
the output of (1,1) is 1.

A2. AND:

A	B	OUT
1	1	1
0	1	0
1	0	0
0	0	0



$$\Rightarrow \begin{cases} w_1 + w_2 + w_3 \geq 0 \\ w_2 + w_3 < 0 \\ w_1 + w_3 < 0 \\ w_3 < 0 \end{cases} \Rightarrow \begin{cases} w_1 > 0 \\ w_2 > 0 \\ |w_3| \leq w_1 + w_2 \\ w_3 < 0 \end{cases} \Rightarrow \text{let} \begin{cases} w_1 = 1 \\ w_2 = 1 \\ w_3 = -1 \end{cases}$$

$w=(1,1,-1)^T$ could work for AND operation.

NOT:

A	OUT
1	0
0	1

$\Rightarrow \begin{cases} w_1 + w_2 < 0 \\ w_2 \geq 0 \end{cases} \Rightarrow \begin{cases} w_1 < 0 \\ |w_1| > w_2 \\ w_2 \geq 0 \end{cases}$

let $\begin{cases} w_1 = -2 \\ w_2 = 1 \end{cases}$

$w=(-2,1)$ satisfies the circumstances of NOT operation.

NAND:

A	B	OUT
1	1	0
0	1	1
1	0	1
0	0	1

$\Rightarrow \begin{cases} w_1 + w_2 + w_3 < 0 \\ w_2 + w_3 \geq 0 \\ w_1 + w_3 \geq 0 \\ w_3 \geq 0 \end{cases}$

$\Rightarrow \begin{cases} w_1 < 0 \\ w_2 < 0 \\ w_1 + w_2 + w_3 < 0 \\ w_3 \geq 0 \\ |w_1| < w_3 \\ |w_2| < |w_3| \end{cases}$ let $\begin{cases} w_1 = -1 \\ w_2 = -1 \\ w_3 = 1.5 \end{cases}$ $w=(-1,-1,1.5)^T$ satisfies the conditions of NAND operation.

NOR:

A	B	OUT
1	1	0
0	1	0
1	0	0
0	0	1

$\Rightarrow \begin{cases} w_1 + w_2 + w_3 < 0 \\ w_2 + w_3 < 0 \\ w_1 + w_3 < 0 \\ w_3 \geq 0 \end{cases}$

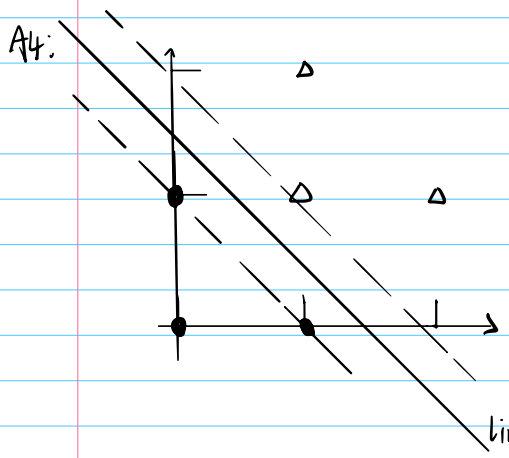
$\Rightarrow \begin{cases} w_1 < 0 \\ w_2 < 0 \\ w_3 \geq 0 \\ |w_1| > w_3 \\ |w_2| > w_3 \\ w_1 + w_2 + w_3 < 0 \end{cases} \Rightarrow$ let $\begin{cases} w_1 = -1 \\ w_2 = -1 \\ w_3 = 0.5 \end{cases}$ $w=(-1,-1,0.5)^T$ satisfies the condition of NOR operation

A3:

A	B	C	OUT
1	1	1	0
0	0	0	0
1	1	0	1
1	0	1	1
0	1	1	1
0	0	1	0
0	1	0	0
1	0	0	0

a single perceptron couldn't learn that for 3 inputs. Because there is not a single line in 3D space that can seperate these points.

A4:



the optimal seperating line is line*.

Support vector are (1, 0), (0, 1), (1, 1)

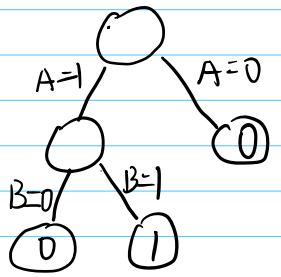
support vectors are the data points that lie closest to the decision boundry and they define the margin of the classifier.

A5. $\text{Entropy} = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{8}\times 4\times\log_2\frac{1}{8} = \frac{1}{2} + \frac{1}{2}\times 3 = 2$

Entropy measures the impurity of the dataset, the greater entropy is, the more uncertainty the dataset has.

A6.

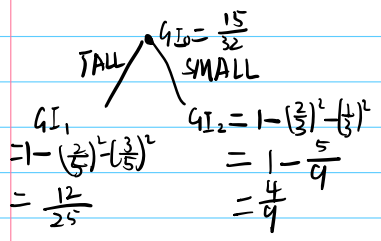
A	B	OUT
1	1	1
1	0	0
0	1	0
0	0	1



decision tree has less computing operations and only needs 2 equality operations.

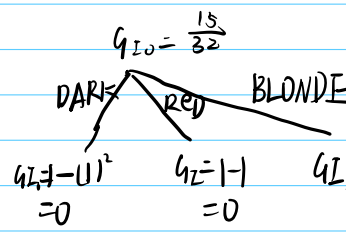
but perceptron needs 2 times multiplication, 1 time add and 1 time comparison operation.

A7 $GI_0 = 1-\left(\frac{3}{8}\right)^2-\left(\frac{5}{8}\right)^2 = \frac{64-9-25}{64} = \frac{30}{64} = \frac{15}{32}$



$IG = \frac{15}{32} - \frac{5}{8}\times\frac{12}{25} - \frac{3}{8}\times\frac{4}{9}$

$= 2.083 \times 10^{-3}$

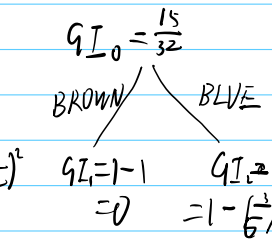


$IG = \frac{15}{32} - \frac{1}{2}\times\frac{1}{2}$

$= \frac{15-8}{32} = \frac{7}{32}$

$= 0.21875$

✓

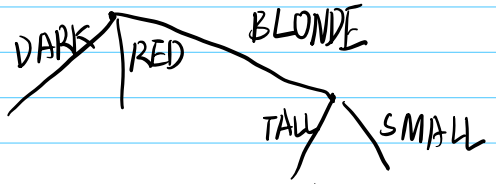


$IG = \frac{15}{32} - \frac{12}{25}\times\frac{5}{8}$

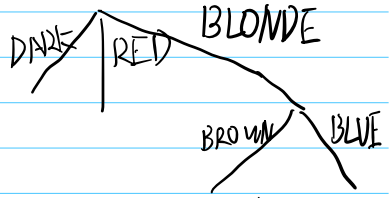
$= \frac{15}{32} - 0.3$

$= \frac{27}{160}$

$= 0.16875$



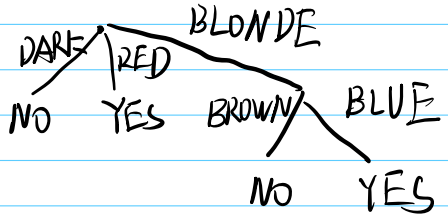
$IG = \frac{1}{2} - \left(\frac{1}{2}\right)\times\frac{1}{2} - \frac{1}{2}\times\frac{1}{2} = 0$



$IG = \frac{1}{2} - 0 = \frac{1}{2}$

✓

SO the final decision tree is



A8 (1) $p = \frac{e^{w^T x}}{1+e^{w^T x}} = \frac{e^{(-6, 0.05, 1)^T \cdot (1, 40, 35)}}{1+e^{(-6, 0.05, 1)^T (1, 40, 35)}}$

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

the estimated probability is $\frac{e^{-0.5}}{1+e^{-0.5}}$

(2) $\frac{e^{-6+0.05t+3.5}}{1+e^{-6+3.5+0.05t}} = \frac{1}{2} \Rightarrow \frac{e^{-2.5+0.05t}}{1+e^{-2.5+0.05t}} = \frac{1}{2}$

$\Rightarrow e^{-2.5+0.05t} = 1 \Rightarrow 0.05t - 2.5 = 0 \Rightarrow t = 50h.$

the student needs 50h.

A9. I think neither of them could be used for the new observations.

1. the logistic regression' model must overfit on the training data, based on the huge gap of error rate between training data and test dataset.
2. the KNN model doesn't give the error rate in the test dataset.

A10. SVM might suffer from different input scale., SVMs are sensitive to the features scale. The detailed reasons are as follow:

① SVM relys on the distance to determine the hyperplane.

② SVM needs regularity to avoid overfitting. If these features are not on the same scale, some of them receive heavy punishment and other receive weak regularization.

A11. ① the number of weak learners.
② higher learning rate
③ the model complexity of weak learners.

How:
1. decrease the learning rate when the model overfits the dataset.
2. change the loss function.
3. if use the decision tree, we could change the parameters of the decision tree, like depth.

A12. ① reduce overfitting by drawing random combinations of the training dataset with repetition.

② save the time and computerational resource from eliminating the need of cross-validation. process.

A13, Hard voting uses the mode, but soft voting uses the weighted majority vote classifier.
soft vote could avoid the situation that is the majority of "people" don't know but the expert know how to solve an issue.