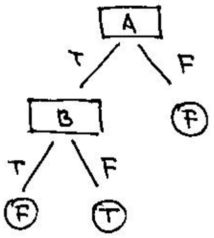
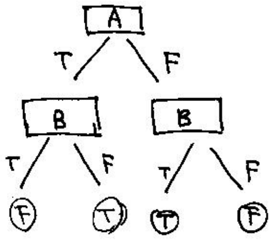
Homework #2.

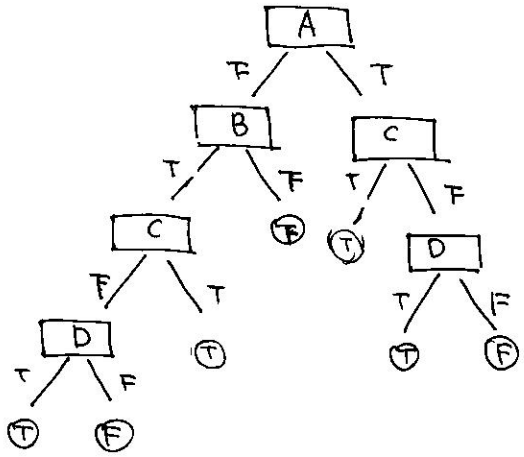
1. (a) $A \wedge \neg B$



(2) $A \veebar B$ (exclusive or)



(3) $[A \vee B] \wedge [C \vee D]$



2. (a) $\text{Entropy}(S) = \text{Entropy}([3+, 3-])$

$$= -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)$$

$$= 1.$$

(b). $\text{Entropy}(S_{a_1=T}) = \text{Entropy}([2+, 2-]) = 1$

$\text{Entropy}(S_{a_1=F}) = \text{Entropy}([1+, 1-]) = 1.$

$$\therefore \text{Gain}(S, a_1) = \text{Entropy}(S) - \left[\frac{4}{6}\cdot \text{Entropy}(S_{a_1=T}) + \frac{2}{6}\text{Entropy}(S_{a_1=F})\right]$$

$$= 1 - \left[\frac{4}{6} + \frac{2}{6}\right] = 0 \qquad -①$$

$$\text{Entropy}(S_{a_2=T}) = \text{Entropy}([3+,\ 1-])$$
$$= -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.8113$$

$$\text{Entropy}(S_{a_2=F}) = \text{Entropy}([0+,\ 2-])$$
$$= -0 - \frac{2}{2}\log_2\frac{2}{2} = 0$$

$$\therefore \text{Gain}(S, a_2) = \text{Entropy}(S) - \left[\frac{4}{6}\cdot 0.8113 + \frac{2}{6}\cdot 0\right] = 0.4591 \quad -②$$

$$\text{Entropy}(S, a_{3=T}) = \text{Entropy}([1+,\ 2-])$$
$$= -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) = 0.9183$$

$$\text{Entropy}(S, a_{3=F}) = \text{Entropy}([2+,\ 1-])$$
$$= -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) = 0.9183$$

$$\therefore \text{Gain}(S, a_3) = \text{Entropy}(S) - \left[\frac{3}{6}\cdot 0.1983 + \frac{3}{6}\cdot 0.9183\right]$$
$$= 1 - 0.9183 = 0.0817 \quad -③$$

By ①, ②, ③, $\text{Gain}(S, a_1) < \text{Gain}(S, a_3) < \text{Gain}(S, a_2)$

$\therefore$ a_2 is the best node.

(c). No!

reason ① : ID 3 algorithm perform no backtracking in its search. Therefore, it can converge to locally optimal solutions that are not globally optimal.

reason ② : ID3 algorithm can overfit the training examples when there is noise in the data on the number of training example is too small to produce a representative sample of the true target function.

3. To avoid the overfitting problem of ID3, we need the pruning method.

The ID3 algorithm grows each branch of the tree just deeply enough to perfectly classify the training examples. While this is sometimes a reasonable strategy, in fact it can lead to difficulties when there is noise in data, or when # of training example is too small to produce a representative sample of the true target function.

In either of these cases, this simple algorithm may produce trees that overfit the training examples.

And generally, after some instant, the training error will be reduced continuously, but the general error (or test-error) will be increased.

4. No!

(a) In this weight update rule, the learning rate $\rho_k = -1$.

So, this rule does not guarantee the convergence.

Let ω^* the optimal solution vector.

$$\omega_{k+1} - \omega^* = \omega_k - \omega^* + \rho_k x_k$$

$$\|\omega_{k+1} - \omega^*\|^2 = \|\omega_k - \omega^*\|^2 + 2\rho_k(\omega_k - \omega^*)^T x_k + \rho_k^2 \|x_k\|^2$$

Since $\omega_k^T x_k \le 0$

$$\|\omega_{k+1} - \omega^*\|^2 \le \|\omega_k - \omega^*\|^2 - 2\rho_k \omega^{*T} x_k + \rho_k^2 \|x_k\|^2$$

$$\therefore \quad 0 < \rho_k < \frac{2\omega^{*T} x_k}{\|x_k\|^2} \quad \cdots \quad (*)$$

Under the condition (*), this weight update rule guarantee the convergence to ω^*.

(b) $\|\underline{w}_{k+1} - \underline{w}^*\|^2 \le \|\underline{w}_k - \underline{w}^*\|^2 + \Delta$

$$= \|\underline{w}_k - w^*\|^2 - 2\underline{w}^{*T} x_k + \|x_k\|^2$$

$$\le \|\underline{w}_k - \underline{w}^*\|^2 - 2\alpha\cdot\eta + \beta^2$$

Let $\delta = 2\alpha\eta - \beta^2$ (>0, and fixed number)

$\Rightarrow$ $\therefore \|\underline{w}_{k+1} - w^*\|^2 \le \|\underline{w}_1 - w^*\|^2 - k\cdot\delta$

If $k > \dfrac{\|\underline{w} - \underline{w}^*\|^2}{\delta}$, then $\|w_{k+1} - \underline{w}^*\|^2 < 0$

This means $\underline{w}_k$ converges to $\underline{w}^*$

after no more than $\dfrac{\|w_1 - \underline{w}^*\|^2}{\delta}$ iterations.