2. $H(D) = \sum_{i=1}^{k} P_i \log_2\left(\frac{1}{P_i}\right) =$

$$\frac{3}{5}\log_2\left(\frac{5}{2}\right) + \frac{3}{5}\log_2\left(\frac{5}{3}\right) = 0.97$$

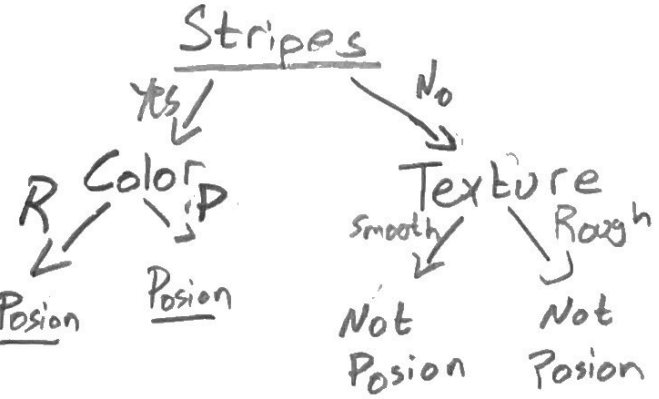
$$Gain(D, Color) = H(D) - \sum_i \frac{|D_i|}{|D|} H(D_i)$$
$$= H(D) - \frac{4}{5}H(Purple) - \frac{1}{5}H(Red)$$
$$H(Purple) = \frac{2}{4}\log_2\left(\frac{4}{2}\right) + \frac{2}{4}\log_2\left(\frac{4}{2}\right) = 1$$
$$H(Red) = 0\log_2(0) + 1\log(1) = 0$$
$$= H(D) - \frac{4}{5}(1) = \boxed{0.1709}$$

$$Gain(D, Texture) = H(D) - \sum_i \frac{|D_i|}{|D|} H(D_i)$$
$$= H(D) - \frac{3}{5}H(smooth) - \frac{2}{5}H(Rough)$$
$$H(smooth) = \frac{2}{3}\log_2\left(\frac{3}{2}\right) + \frac{1}{3}\log(3) = 0.918$$
$$H(rough) = \frac{1}{2}\log_2(2) + \frac{1}{2}\log(2) = 1$$
$$= H(D) - \frac{3}{5}(0.91) - \frac{2}{5}(1) = \boxed{0.0199}$$

$$Gain(D, Stripes) = H(D) - \sum_i \frac{|D_i|}{|D|} H(D_i)$$
$$= H(D) - \frac{2}{5}H(No\ Stripes) - \frac{3}{5}H(stripes)$$
$$H(No\ Stripes) = \frac{0}{2}\log_2 0 + \ldots = 0$$
$$H(stripes) = \frac{1}{3}\log_2(3/1) + \frac{2}{3}\log_2 3/2 = 0.9183$$
$$= H(D) - 0 - \frac{3}{5}(0.9182) = \boxed{0.419}$$

(a) Root is Stripes b/c most Information gain.

(b)



3. Depth 0 decision tree

Starting: $D_{t=0}(i)$ $\frac{1}{m}, \frac{1}{m}, \frac{1}{m}, \frac{1}{m}, \ldots, \frac{1}{m}$

y_t $+1, +1, +1, +1, \ldots, -1$

$g_t(i)$ $+1, +1, +1, +1, \ldots +1$

This will return +1 for everything and have $E_{in}^{(D_t)} = 0.2$

For D_{t+1}, we want $E_{in}^{(D_{t+1})}(g_t) = 0.5$.

Reweight: $\alpha_0 = \frac{1}{2}\ln\left(\frac{1-\varepsilon_0}{\varepsilon_0}\right) = \frac{1}{2}\ln\left(\frac{1-0.2}{0.2}\right)$

$= 0.6931$

$\gamma = \sqrt{\frac{1-\varepsilon_0}{\varepsilon_0}} = 2$

$Z_0 = \gamma\varepsilon_0 + \frac{1}{\gamma}(1-\varepsilon_0) = 2(0.2) + \frac{0.8}{2} = 0.8$

$$D_1(n) = \frac{1}{Z_0} D_0(n) e^{-\gamma_0 g_0(x_n) y_n}$$

- For negative (misclassified) points

$D_1 = \frac{1}{0.8}\frac{1}{m} e^{-0.6931(1)(-1)}$

$= \frac{2}{0.8m} = \frac{5}{2m}$

- For positive points

$D_1 = \frac{1}{0.8}\frac{1}{m} e^{-0.6931(1)(1)}$

$= \frac{1}{0.8m}\frac{1}{2} = \frac{5}{4m}\left(\frac{1}{2}\right) = \frac{5}{8m}$

The cumulative weight for any m is

Positive: $\frac{1}{2}$

Negative: $\frac{1}{2}$

after one update

- This will make the weighted sum of the points 0 for the next iteration. The learned hypothesis from this will be a random guess, (since the weighted sum is 0) and does not even learn anything about the distribution. Since this process will repeat, a depth 0 decision tree is not a good choice for a weak learner.

4. - Bagging works best on high variance, low bias models like the decision tree. By combining and weighting each tree, bagging is able to reduce the variance.
- Linear regression is a low variance, high bias model. Thus, there are not many gains to be made - the variance is already low. It does not really make sense to try and improve linear regression by decreasing its variance. Further, adding many high bias models together could even add different biases together in the ensamble prediction.