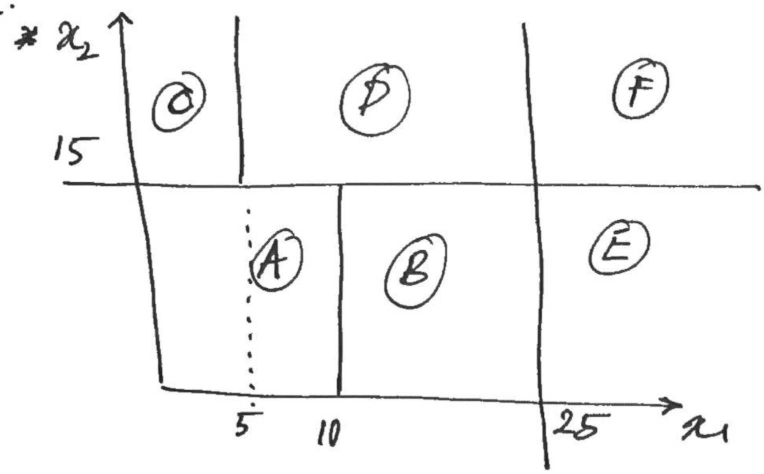
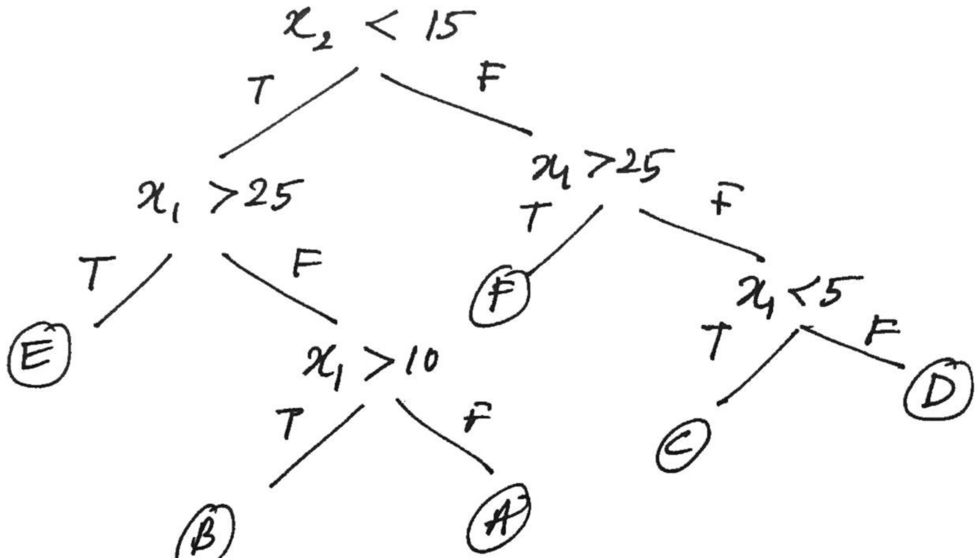
AI 534 Written Assignment 4

1. a.



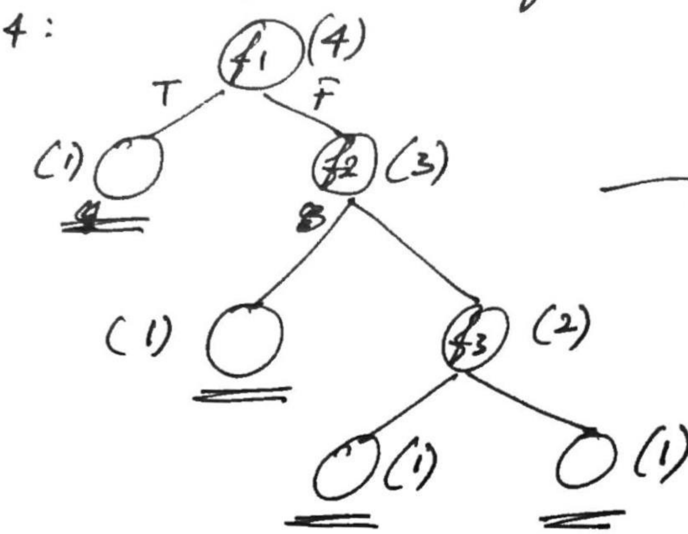
b.



c. This redundancy in the space of decision trees makes it easier to find decision boundaries. This is because independent of the feature selected to split the data in the previous node, there is a high possibility that there still exists a feature that can lead to the optimal decision boundary. For example, irrespective of whether $x_2 < 15$ or $x_1 < 25$ was the better choice, there is still a possibility to reach the optimal boundary through both of them. Therefore it makes it easier to find boundaries.

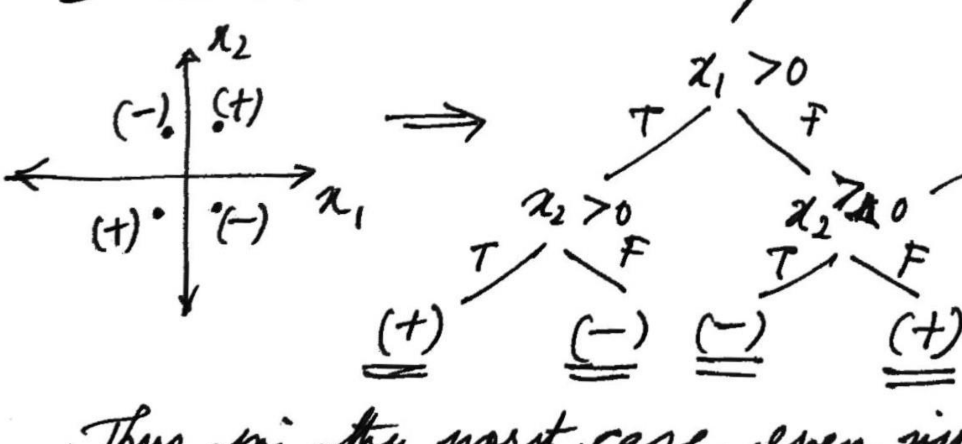
2a. Consider the training set has m examples. In the worst case, we would select a feature (with non-zero gain), that splits the data into $m-1$ elements on one side (for example True) and one element on the other side (False). Therefore, in the worst case, each feature splits the data into 1 and $i-1$ examples, where i is the number of training data examples belonging to that particular node. Thus in total we would require $\underline{m}$ leaf nodes.

For $m = 4$:



4 leaf nodes.

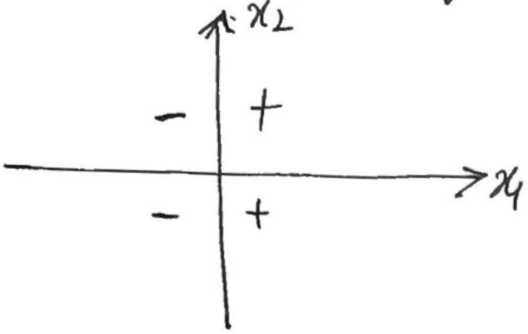
b. In the case of the mutual information scenario as well, there might only be features that split the data set into 1 and $i-1$ sets. For example consider



4 leaf nodes $= m$.

Thus in the worst case, even with information gain we would get m leaf nodes = random feature selection.

c. Using maximum information gain would lead to better results on average in comparison to using random splits. Consider the following example:



Using maximum information gain we would select $(x_1 > 0)$ as the first decision which would give the best split. Using random splits, we might get two conditions $(x_2 > 0,\ x_1 > 0)$.

Thus as the number of data elements grows, random splits has a higher probability of using larger trees to split the data as compared to max information gain. As a result random splits are thus more prone to overfitting although they can get good training accuracies. Thus maximum information gain leads to more generalizable splits and lesser overfitting than random ~~split~~ splits.

3. $H(Y|A=0) = -\left(\frac{2}{3}\log\frac{2}{3} + \frac{1}{3}\log\frac{1}{3}\right)\times\frac{1}{2} = 0.4591$

$H(Y|A=1) = -\left(\frac{1}{3}\log\frac{2}{3} + \frac{2}{3}\log\frac{1}{3}\right)\times\frac{1}{2} = 0.4591$

$H(Y|B=0) = -\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right)\times\frac{1}{3} = 0.333$

$H(Y|B=1) = -\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right)\times\frac{2}{3} = 0.666$

$H(Y|C=0) = -\left(\frac{1}{3}\log\frac{1}{3} + \frac{2}{3}\log\frac{2}{3}\right)\times\frac{1}{2} = 0.4591$

$H(Y|C=1) = -\left(\frac{2}{3}\log\frac{2}{3} + \frac{1}{3}\log\frac{1}{3}\right)\times\frac{1}{2} = 0.4591$

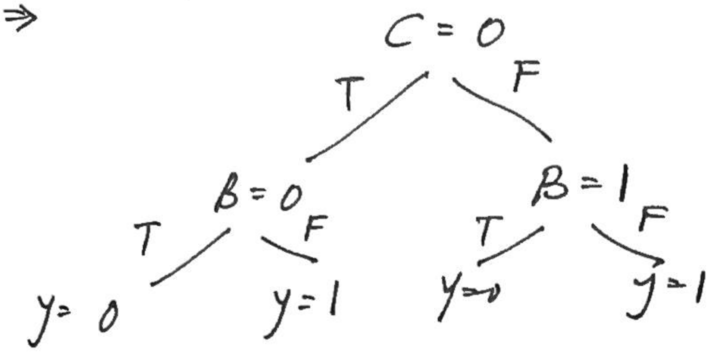
We have a tie for A and C.

But we can see that:

$$H(Y|B=0, C=0) = H(Y|B=1, C=0) = H(Y|B=0, C=1) = H(Y|B=1, C=1) = 0$$

Therefore selecting feature C results in a better split later on:

$\Rightarrow$



4. $\alpha_l = \frac{1}{2} \ln\left(\frac{1-\varepsilon_l}{\varepsilon_l}\right)$

$$D_{l+1} = D_l(i) \times \begin{cases} e^{\alpha_l} & h_l(x_i) \neq y_i \\ e^{-\alpha_l} & h_l(x_i) = y_i \end{cases}$$

$$\varepsilon_l = \text{error} = \frac{\#\text{mistakes}}{\#\text{training examples}} = \frac{m}{N}$$

$\Rightarrow$ If $h_l(x_i) \neq y_i$: $D_{l+1}(i) = D_l(i) . e^{\frac{1}{2}\ln\left(\frac{1-\varepsilon_l}{\varepsilon_l}\right)}$

$$= D_l(i) \cdot \left(\frac{1-\varepsilon_l}{\varepsilon_l}\right)^{1/2}$$

$$= D_l(i) \left(\frac{1-m/N}{m/N}\right)^{1/2} = D_l(i)\left(\frac{N-m}{m}\right)^{1/2}$$

$\Rightarrow$ If $h_l(x_i) = y_i$: $D_{l+1}(i) = D_l(i) . e^{-\frac{1}{2}\ln\left(\frac{1-\varepsilon_l}{\varepsilon_l}\right)}$

$$= D_l(i) \cdot \left(\frac{1-\varepsilon_l}{\varepsilon_l}\right)^{-1/2} = D_l(i) \cdot \left(\frac{m}{N-m}\right)^{1/2}.$$

After normalizing by $\sum_{\text{mistake}} D_{\ell+1} + \sum_{\substack{\text{not}\\ \text{mistake}}} D_{\ell+1}$

$$\Rightarrow \text{normalizing factor} = N.F = \sum_{h_\ell(x_i)=y_i} D_\ell(i)\left(\frac{m}{N-m}\right)^{1/2} + \sum_{h_\ell(x_i)\neq y_i} D_\ell(i)\left(\frac{N-m}{m}\right)^{1/2}$$

$\Rightarrow$ After normalization :

$$D_{\ell+1}(i) = \begin{cases} \dfrac{D_\ell(i)\cdot\left(\frac{N-m}{m}\right)^{1/2}}{NF} & \text{if } h_\ell(x_i)\neq y_i \\[2ex] \dfrac{D_\ell(i)\cdot\left(\frac{m}{N-m}\right)^{1/2}}{NF} & \text{if } h_\ell(x_i) = y_i \end{cases}$$

Summing up the weights of all errors

$$\frac{\sum_{h_\ell(x_i)\neq y_i} D_\ell(i)\cdot\left(\frac{N-m}{m}\right)^{1/2}}{\sum_{h_\ell(x_i)\neq y_i} D_\ell(i)\left(\frac{N-m}{m}\right)^{1/2} + \sum_{h_\ell(x_i)=y_i} D_\ell(i)\left(\frac{m}{N-m}\right)^{1/2}} = S.$$

Basis

At the ~~stat~~ start of the first iteration $D_1(0) = \frac{1}{N}$, no. of mistakes $= m_0$

$$\Rightarrow S_0 = \frac{\frac{1}{N}\cdot\left(\frac{N-m_0}{m_0}\right)^{1/2}\cdot m_0}{\frac{1}{N}\left(\frac{N-m_0}{m_0}\right)^{1/2}\cdot m_0 + \frac{1}{N}\left(\frac{m_0}{N-m_0}\right)^{1/2}\cdot(N-m_0)}$$

$$= \frac{((N-m)m_0)^{1/2}}{((N-m_0)m_0)^{1/2} + ((N-m_0)m_0)^{1/2}} = \underline{\underline{0.5}}$$

Assume that this is true for iteration ~~i~~, (i.e) the sum of weights after iteration (i) of ~~errors~~ mistakes = 1.5 (Hypothesis)

Proof: For iteration i+1

$$S_{i+1} = \frac{\sum\limits_{h_l(x_i)\neq y_i} D_l(i)\cdot\left(\frac{N-m_i}{m_i}\right)^{1/2}}{\sum\limits_{h_l\neq y_i} D_l(i)\left(\frac{N-m_i}{m_i}\right)^{1/2} + \sum\limits_{h_l(x_i)=y_i} D_l(i)\left(\frac{m_i}{N-m_i}\right)^{1/2}}$$

$$= \frac{\left(\frac{N-m_i}{m_i}\right)^{1/2}\cdot \sum\limits_{h_l(x_i)\neq y_i} D_l(i)}{\left(\frac{N-m_i}{m_i}\right)^{1/2}\cdot \sum\limits_{h_l(x_i)\neq y_i} D_l(i) + \left(\frac{m_i}{N-m_i}\right)^{1/2}\cdot \sum\limits_{h_l(x_i)=y_i} D_l(i)}$$

$$= \frac{\left(\frac{N-m_i}{m_i}\right)^{1/2}\cdot (0.5)}{\left(\frac{N-m_i}{m_i}\right)^{1/2}\cdot(0.5) + \left(\frac{m_i}{N-m_i}\right)^{1/2}\cdot(0.5)}$$

Multiplying Numerator and Denominator by $(N-m_i)m_i$

$$S_i = \frac{(m_i(N-m_i))^{1/2}}{(m_i(N-m_i))^{1/2} + ((N-m_i)m_i)^{1/2}} = \underline{\underline{0.5}}$$

5. <u>Basis</u> : $D_1 = \frac{1}{N}$

Each example is weighted equally

$$\epsilon_1 = \sum_{i=1}^{N} \epsilon_1^{(i)} \quad \text{where} \quad \epsilon_1^{(i)} = D_1^{(i)} \cdot I[y_i \neq h_1(x_i)].$$

$\Rightarrow$ When minimising the total error, each example is weighted by $D_1^{(i)}$

<u>Hypothesis</u> : Assume this to be true for iteration $l-1$

<u>Proof</u> : $D_l^{(i)} = D_{l-1}^{(i)} \exp(-y_i \alpha_{l-1}^{(i)} h_{l-1}(x_i)).$

Weight of example in iteration $l = D_l^{(i)}$

$$\epsilon_l^{(i)} = D_l^{(i)} I(y_i \neq h_l(x_i)).$$

$$\Rightarrow \underline{\underline{w_l^i \propto D_l^{(i)}}}$$