PROBLEM. 3.

Given an arbitary decision tree, it might have repeated queries splits (feature f, threshold t) on some paths root leaf. Prove that there exists an equivalent binary tree only with distinct splits on ~~of~~ each path.

Ans. let $a_1 \ldots a_n$ denote the attributes characterizing the data under consideration

let $D_1 \ldots D_n$ denote the corresponding domains (i.e D_i represents the set of values for attribute a_i)

Let $c_1, \ldots c_m$ represent the decision classes associated with dataset.

Now, two set of rules could result in the equivalent decision tree.

~~This is done assigning a~~.

There are decisions which solely depends on one attribute; equivalent tree could be designed by assigning a 'don't care' value to rest of the attributes.

Fore example, ~~w~~

we have attributes a_1, a_2 and a_3 with the same domain containing v_1, v_2 and v_3 as possible values.

Assuming that the following rules correspond to class C1 :

rule 1 : $C1 \leftarrow a_1 = V_1 \;\&\; a_2 = V_2$

rule 2 : $C1 \leftarrow a_1 = V_3$

Equivalent representations of these two rules will be following

rule 1 : $C_1 \leftarrow a_1 = V_1 \;\&\; a_2 = V_2 \;\&\; a_3 =$ don't care

rule 2 : $C_1 \leftarrow a_1 = V_3 \;\&\; a_2 =$ don't care $\;\&\; a_3 =$ don't care

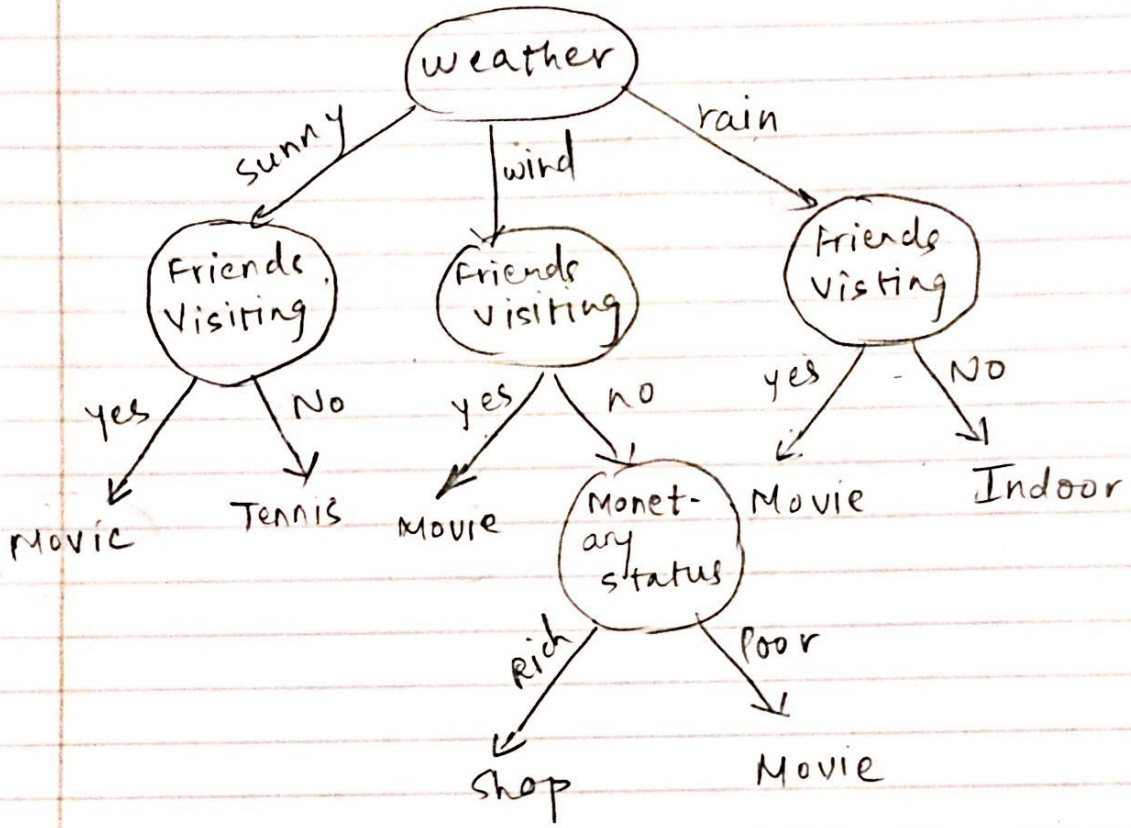
Considering real life example,

C_1 = Movie
C_2 = Tennis
C_3 = shop
C_4 = Indoors

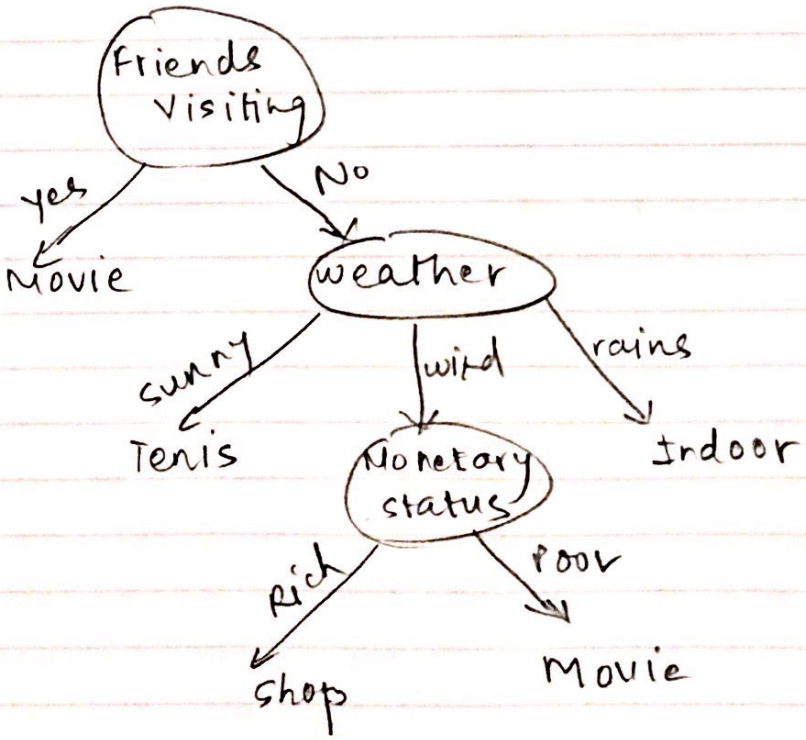
a_1 = Friends visiting : V_1 = yes
V_2 = no.

a_2 = weather : V_1 = sunny
V_2 = wind
V_3 = rain

a_3 = Monetary status : V_1 = rich
V_2 = poor.



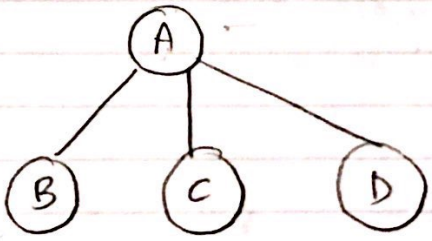
This tree would have an equivalent tree where the attribute won't be repeated as follows



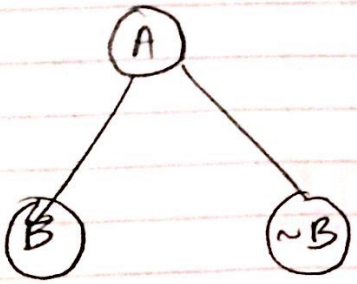
PROBLEM 3

(a) Prove that for any arbitary tree, with possible unequal branching ratios throughout, there exists a binary tree that implements the same classification functionality.

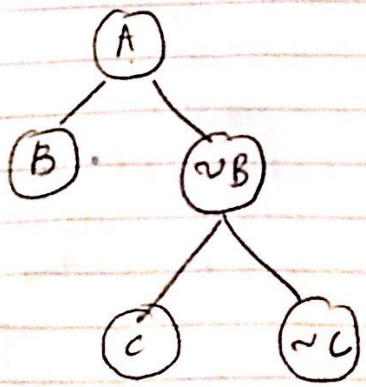
Consider a Node A,
Now, there is a choice of traversing to nodes B, C, & D from A based on the requirement



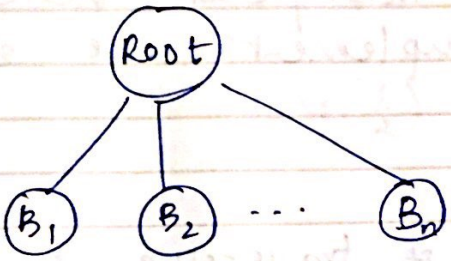
since they are mutually exclusive, this tree could be split into binary decision i.e B or not B



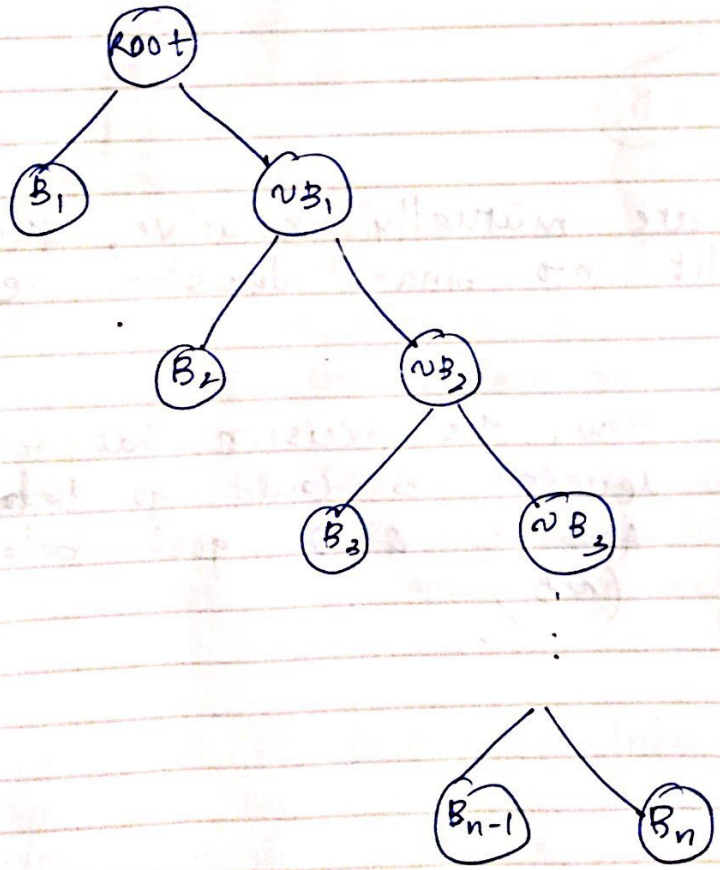
Now, the decision that follows the decision B could go below it, And C & D goes below not B (~B)



Thus, considering a root node and B_n leaf nodes in a level-2 decision tree;



there will be one extra node for each of the node starting from B_3 till B_n.



(b) Consider a tree with just two levels, — a root node connected to B leaf nodes (B >= 2). What are then upper and lower limits on the number of levels of functionally equivalent binary tree, as a function of B?

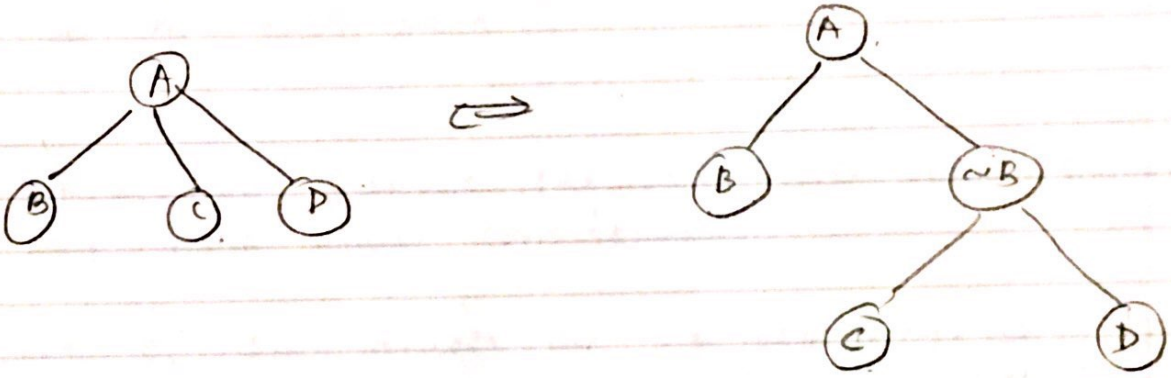
Ans. let levels of existing k-ary tree = Z = L
Number of branches = B

Best case (lower limit) = B + L − 1 = B + 2 − 1 = B + 1
Upperlimit = (L − 1) * × B + 1

Worstcase = (2 − 1) × B + 1 = B + 1

Here, since the level is 2, the Upperlimit & lower limit for the k equivalent binary tree is same.

(c) As in (b), what are the upper and lower limits on number of nodes in functionally equivalent binary tree?



lower limit : 1 Extra node for every node for (B-2) such Nodes.

If k-ary tree has a root node connected to B leaf nodes (B >= 2) then,

number of nodes in binary tree = (B-2) + B
= 2B - 2 + Root Node

Upper limit : When each node has an extra ~Node corresponding to it.

Thus, if k-ary tree has a root node connected to B leaf nodes (B >= 2) then,

number of nodes in binary tree = 2B + Rout node

PROBLEM 4

(a) Prove that the decrease in entropy by a split on a binary yes/no feature can never be greater than 1 bit.

Let number of items = N
Let these items fall into 2 categories i.e (Binary yes/no feature).

Number of items that has Label 1 = n
Number of items that has Label 2 = N − n = m

To get out data more ordered, we group them by labels.

Calculate the ratio,

$$p = \frac{n}{N} \quad \text{and} \quad q = \frac{m}{N} = 1 - p$$

The entropy of our set is given by the following equation:

$$E = -p\log_2(p) - q\log_2(q)$$

Now, consider a split when there is no item with label 1 in the set i.e $p = 0$ or set is full of items with Label 1 i.e $p = 1$.

$$E = -p\log_2(p) - q(\log_2 q)$$

$$= -p\log_2(p) - (1-p)[\log_2(1-p)]$$

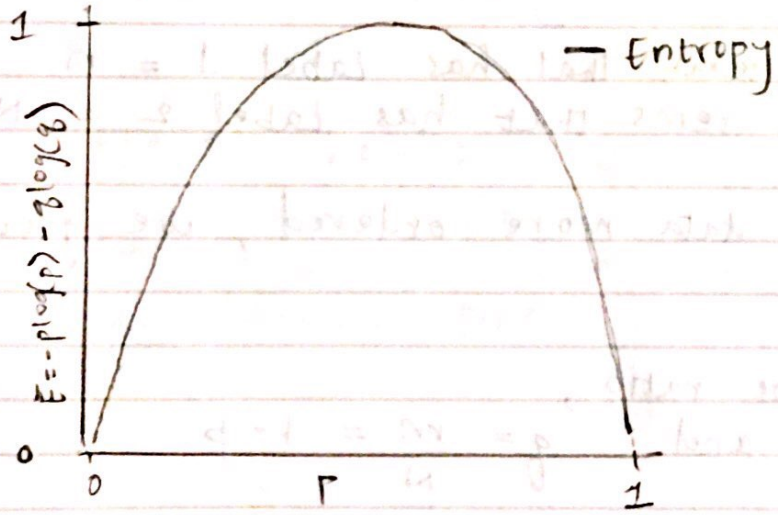
$$E = 0$$

Now, consider a split with set that has half Label 1 and half Label 2 i.e $p = \frac{1}{2}$

$$E = -p\log(p) - (1-p)\left[\log_2(1-p)\right]$$

$$= -\frac{1}{2}\log\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right)\left[\log_2\left(\frac{1}{2}\right)\right]$$

$$= 1.$$



The figure represents the Entropy for a binary-state variable with probabilities p & q where p+q = 1.

Thus entropy can only vary between 0 and 1 as probability varies between 0 and 1. Hence, decrease in entropy by split can never be greater than 1.

From the perspective of information theory, it specifies the minimum number of bits of information needed to encode the classification of an arbitary member of set S. (i.e an ~~me~~ item drawn from set at random with uniform probability

For example, if $p = 1$, the receiver knows that the drawn sample will be positive, so no message needs to be sent and the entropy is zero.

Whereas, if $p = 0.5$, 1 bit is required to indicate whether the sample is +ve or negative.

(b) Generalize this result to the case of arbitary branching $B > 1$

Ans. Now, consider a case where the target can take up b possible (different) values. Then entropy,

$$E = \sum_{i=1}^{b} -p_i \log_2 p_i$$

where p_i is the probability of the i^{th} class.

Now, if the target attribute can take up b possible values, the Entropy could be as large as

$$\log_2 b$$

PROBLEM 5.

Derive explicit formulas for normal equations solution presented in class for the case of one input dimension.

Ans. The data is (x_i, y_i) where $i = 1, 2, \ldots m$ and we are looking at $h(x) = ax + b$ that realizes the minimum mean square error.

The equation $h(x_i) = ax_i + b$ can be put into matrix form as :

$$\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

So the normal equations are :

~~$[X^T X]\theta$~~

$$X^T X \theta = X^T y$$

This could be further simplified as :

$$\begin{bmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_m \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_m \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

which could be simplified as :

$$\underbrace{\begin{bmatrix} n & \Sigma_i x_i \\ \Sigma_i x_i & \Sigma_i x_i^2 \end{bmatrix}}_{X^T X} \underbrace{\begin{bmatrix} a \\ b \end{bmatrix}}_{\Theta} = \underbrace{\begin{bmatrix} \Sigma_i y_i \\ \Sigma_i x_i y_i \end{bmatrix}}_{X^T Y}$$

Now solving the 2x2 linear system,

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} n & \Sigma_i x_i \\ \Sigma_i x_i & \Sigma_i x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_i y_i \\ \Sigma_i x_i y_i \end{bmatrix}$$

$$= \frac{1}{n\Sigma_i x_i^2 - (\Sigma_i x_i)^2} \begin{bmatrix} \Sigma_i x_i^2 & -\Sigma_i x_i \\ -\Sigma_i x_i & n \end{bmatrix} \begin{bmatrix} \Sigma_i y_i \\ \Sigma_i x_i y_i \end{bmatrix}$$

$$= \frac{1}{n\Sigma_i x_i^2 - (\Sigma_i x_i)^2} \begin{bmatrix} (\Sigma_i x_i^2)(\Sigma_i y_i) - (\Sigma_i x_i)(\Sigma_i x_i y_i) \\ -(\Sigma_i x_i)(\Sigma_i y_i) + n(\Sigma_i x_i y_i) \end{bmatrix}$$

$$= \frac{1}{\frac{1}{n}\Sigma_i x_i^2 - \left(\frac{1}{n}\Sigma_i x_i\right)^2} \begin{bmatrix} \left(\frac{1}{n}\Sigma_i x_i^2\right)\left(\frac{1}{n}\Sigma_i y_i\right) - \left(\frac{1}{n}\Sigma_i x_i\right)\left(\frac{1}{n}\Sigma_i x_i y_i\right) \\ -\left(\frac{1}{n}\Sigma_i x_i\right)\left(\frac{1}{n}\Sigma_i y_i\right) + \left(\frac{1}{n}\Sigma_i x_i y_i\right) \end{bmatrix}$$

PROBLEM - 7

The convex hull of a set of vectors x_i, where $i = 1, 2, \ldots n$ is the set of all vectors of the form

$$x = \sum_{i=1}^{n} \alpha_i x_i \qquad \text{——— (1)}$$

where the coefficients α_i are non-negative and sum to one. Given two sets of vectors, show that either they are linearly separable or their convex hulls intersect.

Ans. Given: $\alpha_i \geq 0$; $\sum_{i=1}^{n} \alpha_i = 1$

Now, consider 2nd set of points corresponding 2nd hull. z_j, where $j = 1, 2 \ldots m$

The two set of points will be linearly separable if there exists a vector $\hat{w}$ and a scalar w_0 such that

$\hat{w}^T x_i + w_0 > 0$ for all x_i, and

$\hat{w}^T z_j + w_0 < 0$ for all z_j

Now, calculate the linear discriminant for the points belonging to the two convex hulls.

Linear discriminant for points in set I,

$$y(x) = \hat{w}^T x^{(n)} + w_0 \qquad \text{——— (2)}$$

Substituting (1) in (2), we get

$$y(x) = \hat{w}^T \left(\sum_{i=1}^{n} \alpha_i x_i \right) + w_0 \quad \text{——— (3)}$$

Bringing the α outside,

$$y(x) = \sum_{i=1}^{n} \alpha_i (\hat{w}^T x_i) + w_0$$

$$= \sum_{i=1}^{n} \alpha_i (\hat{w}^T x_i + w_0) \quad \text{——— (4)}$$

Give: $\sum_{i=1}^{n} \alpha_i = 1$.

similarly, we find linear discriminant for points belonging to 2nd set

$$y(z) = \sum_{i=1}^{m} \beta_i (\hat{w}^T z_i + w_0) \quad \text{——— (5)}$$

Given: $\beta_i \geq 0$ & $\sum_{i=1}^{m} \beta_i = 1$.

consider the scenario where convex hull intersects.
This means there must be atleast one point in common between $\{x\}$ & $\{z\}$

Let the common point = a

since a belong to both convex hulls, there must be set of α & β that give rise to a.

linear discriminant for a from equation (4) & (5),

$$y(a) = \sum_{i=1}^{n} \alpha_i \left(\hat{w}^T x_i + w_0 \right)$$

$$= \sum_{i=1}^{m} \beta_i \left(\hat{w}^T z_i + w_0 \right) \quad \text{——— (6)}$$

For linear separability, we must have

$$\therefore y(x) = \hat{w}^T x + w_0 > 0 \quad \& \quad y(z) = \hat{w}^T z + w_0 < 0 \quad \text{——— (7)}$$

From the non-negativity and simplex constraints on α & β and equations (6) & (7); we have a contradiction.

The linear discriminant $y(a)$ has to be simultaneously greater than and less than zero which is impossible.

Now, consider the scenario where patterns are linearly separable,

$$y(x) = \hat{w}^T x + w_0 > 0$$

$$\text{i.e. } y(z) = \hat{w}^T z + w_0 < 0 \quad \text{——— (8)}$$

Assuming that there is a point a lying in the intersection of the convex hulls.

From equation (6), we have

$$y(a) = \sum_{i=1}^{n} \alpha_i (\hat{w}^T x_i + w_0) = \sum_{i=1}^{m} \beta_i (\hat{w}^T z_i + w_0)$$

Equation is not posible since from equation (8), we have the fact that contradicts the above equation.