Problem 3

Whenever you see repeated queries on some root-leaf path, you can easily neglect the lower split and delete its corresponding node of decision tree.

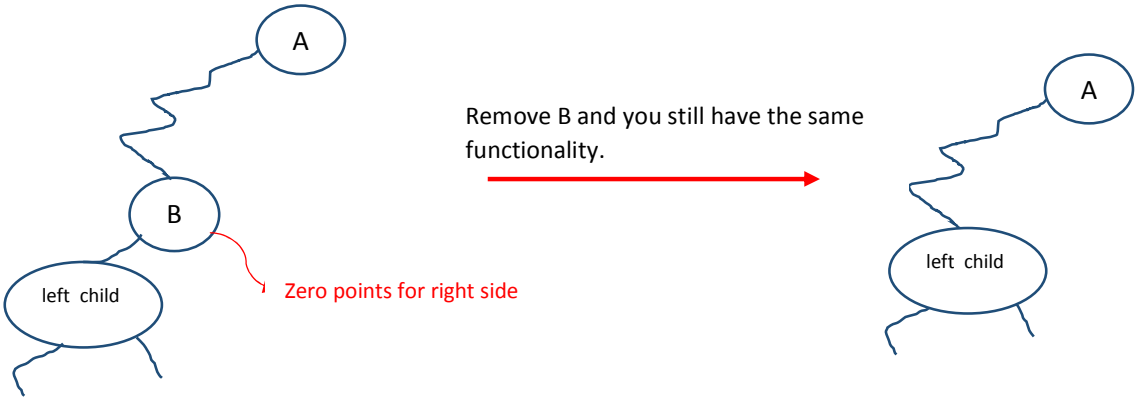
Proof. Let A and B be the two nodes with the same splitting query feature-threshold (f,t) in some root-leaf path. Take A as the upper node and B as the lower node. There are two possibilities for node B to be the left or right descendant of A. With loss of generality, assume that B is the left descendant of A. Points in Node A split by (f,t) in two classes left and right branches.

left. set of points that their feature f is less than or equal t.

right. set of points that their feature f is greater than t.

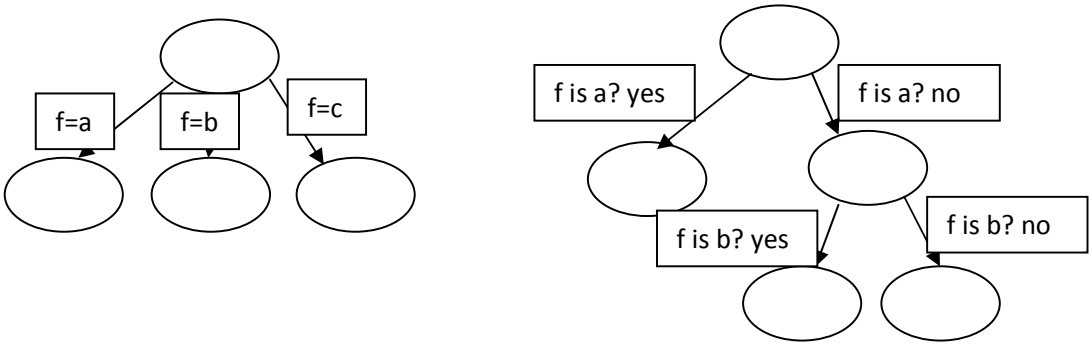
Thus, left branch contains points that their feature f is less than or equal t. Consequently, points which B includes are points which their feature f is less than or equal t. If we split B by query (f,t), its right branch has zero points to include since the feature f of all points in B is less than or equal to t. Thus, left branch of B includes all points from B and right branch of B has zero points, Which means:

$H(B) = H(B \mid f,t) \rightarrow \text{InfoGain}(B, (f,t)) = 0$



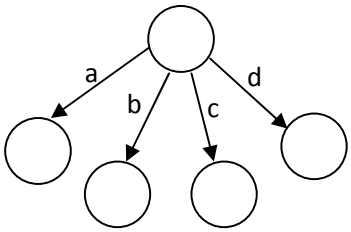
So the algorithm is whenever you see nodes with repeated splitting query in a root-leaf path, delete the lower node, and replace it with its only child.

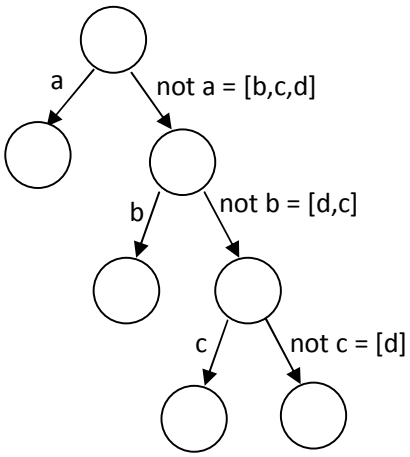
a) Every time, you have a node n with B>2 branches, you can convert its feature to a yes/no feature, just like example below, without losing functionality. This way you left node n with two branches, instead one of branch of n is replaced with B-1 branches. You can recursively update nodes with more than 2 branches. Using yes/no feature you can convert it to a binary tree.



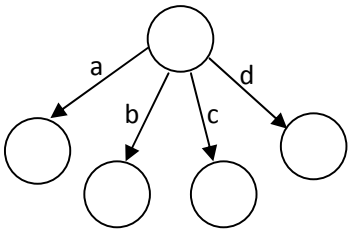
b.

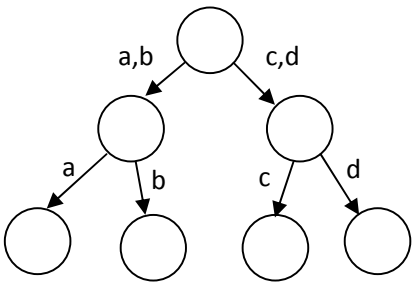
The upper limit is B level. Each time, if you pick one label and divide points whether they have that label or not, you will have a binary tree like right tree in below picture. Since each time you split based on only one label, you will have B levels.





The lower limit is $\lceil \log B \rceil + 1$. You can obtain this lower limit, if divide branches into almost equally two branches which each branch has about half of labels. For example, if a node has B branch, in the equivalent tree that node should be split in two branches each covers $\left\lfloor \frac{B}{2} \right\rfloor$ and $\left\lceil \frac{B}{2} \right\rceil$ labels. This way, the final number of levels would be $\lceil \log B \rceil + 1$.





c. In each of cases below, the number of leaves remains unchanged, and is B.

1. The case when the number of levels is B:

The number of internal nodes (including root) is B-1. Thus the whole number of nodes is 2B-1.

2. The case when the number of levels is $\lceil \log B \rceil + 1$:

The number of internal nodes (including root) is $\left\lfloor \frac{B}{2} \right\rfloor + \left\lfloor \frac{B}{2^2} \right\rfloor + \ldots + 1 = B - 1$. The total number of leaves is 2B-1.

To conclude, both of the cases are using the same number of internal nodes.

Problem 4

a) Think about the time when we have the maximum decrease in entropy, which implies the maximum increase in certainty. Consider X is a yes/no feature that we use to split points in Y.

max{ H(Y) – H(Y|X) } = ? The maximum is obtained when:

1. H(Y) has the maximum entropy or uncertainty which is 1, and this happens when all labels/classes have equal probability among elements of Y.
2. H(Y|X) is 0, which means in each of branches Y|X=yes and Y|X=no, elements have the same label/class, this implies that H(Y|X=yes) and H(Y|X=no) are both 0.
   In this case, in each branch we have a label/class which all elements of branch has that label, so the probability of that label is one and probability of other labels is 0. This lead to H(Y|X=yes) and H(Y|X=no) to be zero.

Why the decrease in entropy by a split on a binary yes/no feature can never be greater than 1 bit? Using only one yes/no feature, we will split elements to two branches, which in best case (when we have maximum decrease in entropy) elements in each branch have only one label. This means we spend one bit two separate the elements and finally determine their label.

b) For an arbitrary branching B>1, decrease in entropy can never be greater than $\lceil \log_2 B \rceil$.
   In best case, H(Y)=1 and H(Y|X=1) = 0 and … and H(Y|X=B)=0 and we have maximum decrease here, which elements of each branch have the same label. When num of branches is B, we can assign at most one label to each branch in best case. Since in best case, elements of each branch has one label. Thus, using $\lceil \log_2 B \rceil$ bits, we can distinguish at most B class of with pure element having the same label.

Problem 5

MSE = $\frac{\sum_{i=1}^{m}(y_i - ax_i - b)^2}{m}$

To minimize MSE, we should set first derivative in term of a and b equal to 0. Then based on these two equations find a and b:

$$\frac{\partial \text{MSE}}{\partial \text{a}} = 0, \frac{\partial \text{MSE}}{\partial \text{b}} = 0$$

$\frac{\partial \text{MSE}}{\partial \text{a}} = \frac{2}{m}.\sum_{i=1}^{m}((y_i - ax_i - b).(-x_i)) = 0$ → $\sum_{i=1}^{m}(y_i x_i - ax_i{}^2 - bx_i) = 0$

→ b. $\sum_{i=1}^{m} x_i$ + a. $\sum_{i=1}^{m} x_i{}^2 = \sum_{i=1}^{m} y_i x_i$ (eq1)

$\frac{\partial \text{MSE}}{\partial \text{b}} = \frac{2}{m}.\sum_{i=1}^{m}((y_i - ax_i - b).(-1)) = 0$ → $\sum_{i=1}^{m}(y_i - ax_i - b) = 0$

➔ $b.m + a.\sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i$ (eq2)

In terms of matrix:

A . $\vec{b}$= x ➔ $\vec{b} = A^{-1}$.x

$$\begin{bmatrix} \sum_{i=1}^{m} x_i & \sum_{i=1}^{m} {x_i}^2 \\ m & \sum_{i=1}^{m} x_i \end{bmatrix} . \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{m} y_i x_i \\ \sum_{i=1}^{m} y_i \end{bmatrix}$$

$$\text{A} = \begin{bmatrix} \sum_{i=1}^{m} x_i & \sum_{i=1}^{m} {x_i}^2 \\ m & \sum_{i=1}^{m} x_i \end{bmatrix} \quad \rightarrow \quad A^{-1} = \frac{1}{((\sum_{i=1}^{m} x_i)^2 - m.(\sum_{i=1}^{m} {x_i}^2))} . \begin{bmatrix} \sum_{i=1}^{m} x_i & -m \\ -\sum_{i=1}^{m} {x_i}^2 & \sum_{i=1}^{m} x_i \end{bmatrix}$$

$$\text{x} = \begin{bmatrix} \sum_{i=1}^{m} y_i x_i \\ \sum_{i=1}^{m} y_i \end{bmatrix}$$

$$\vec{b} = A^{-1}.\text{x} = \frac{1}{((\sum_{i=1}^{m} x_i)^2 - m.(\sum_{i=1}^{m} {x_i}^2))} . \begin{bmatrix} \sum_{i=1}^{m} x_i & -m \\ -\sum_{i=1}^{m} {x_i}^2 & \sum_{i=1}^{m} x_i \end{bmatrix} . \begin{bmatrix} \sum_{i=1}^{m} y_i x_i \\ \sum_{i=1}^{m} y_i \end{bmatrix} = \begin{bmatrix} b \\ a \end{bmatrix}$$

Problem 7

Let the two convex hull be and x_2, and suppose that both statements are true:

1. they are linearly separable, based on definition of DHS, chapter 5, there is a weight vector **a,** and there is a discriminate function g = $a^T.y$, which without loss of generality, when you put a sample point from x_1 into g, g>0 and when you put a sample point from x_1 into g, g<0.
2. now consider that x_1 and x_2 intersect each other, this means there is a point p, which belongs to both x_1 and x_2.

Now put p into function g. what should be the value of g(p). g(p) could not be less than and greater than zero at the same time. This contradicts the first assumption that both statements are true at the same time.