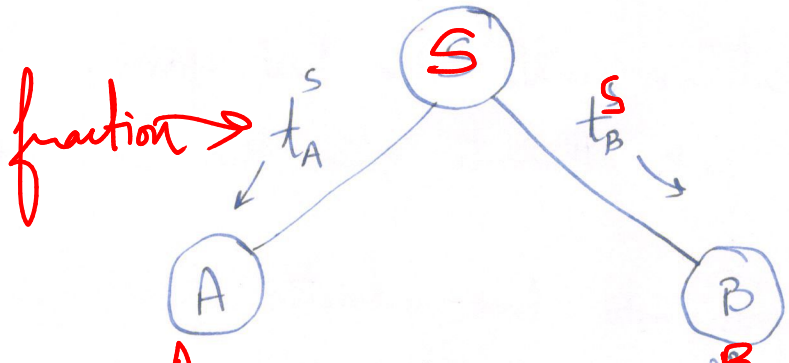


fraction $\rightarrow t_A^S$ $\quad t_B^S$

t_A^S: fraction of examples of S descending to node A.

$f_{\text{minority label}}^A$ leading to training error

$f_{\text{minority label}}^B$ leading to training error.

Reduction in training error:

$$f_{\text{minority label}}^S - \left(t_A^S \cdot f_{\text{minority label}}^A + t_B^S \cdot f_{\text{minority label}}^B \right)$$

error rate after the split

Gain as "Information Gain"

How much information gain (reduction in entropy) happens if the node is split?

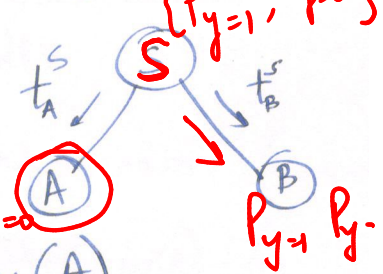
Entropy: Measure of randomness (Uncertainty)

Highly uncertain events carry more information

For a set of events $\{e_1\ e_2 \cdots e_N\}$ which occur with probability $\{p_{e_1}\ p_{e_2} \cdots p_{e_N}\}$ the entropy is given as

$\{1,0,0,0,0\}$ No information

$$\text{Entropy} = \sum_i p_{e_i} \log \frac{1}{p_{e_i}}$$

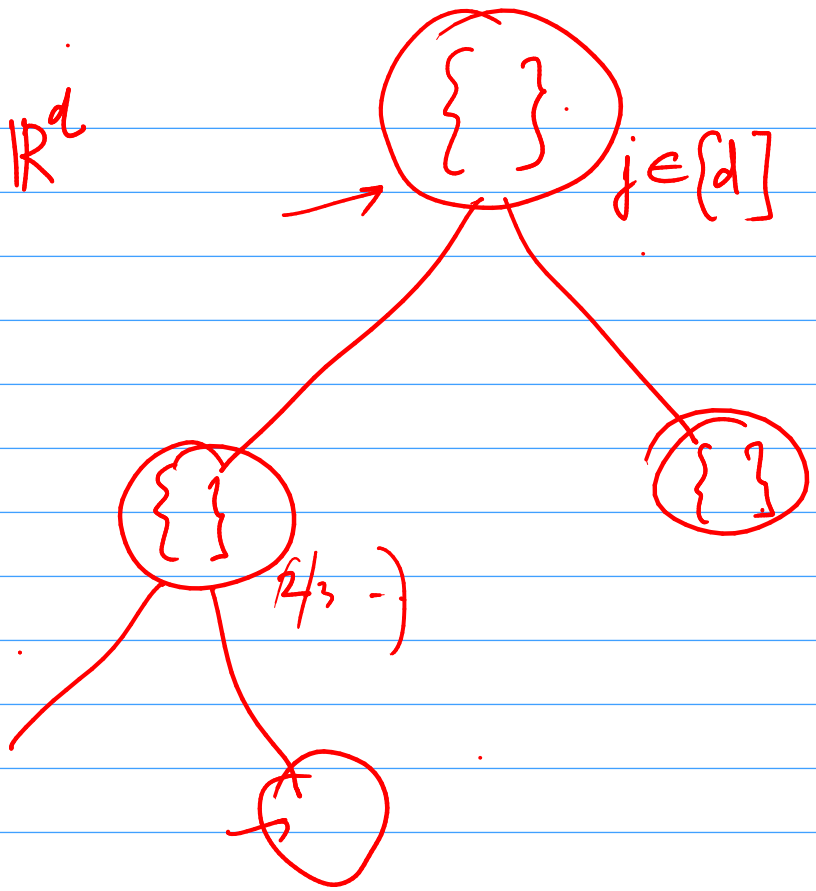


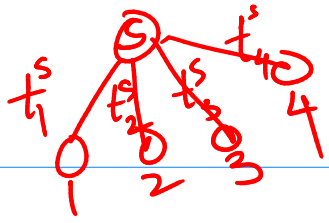
Information Gain: $\text{Entropy}(S) - t_A^S \text{ Entropy}(A) - t_B^S \cdot \text{Entropy}(B)$

Probability is computed as the frequency of target labels for the examples in the set.

Ground truth

$x \in \mathbb{R}^d$



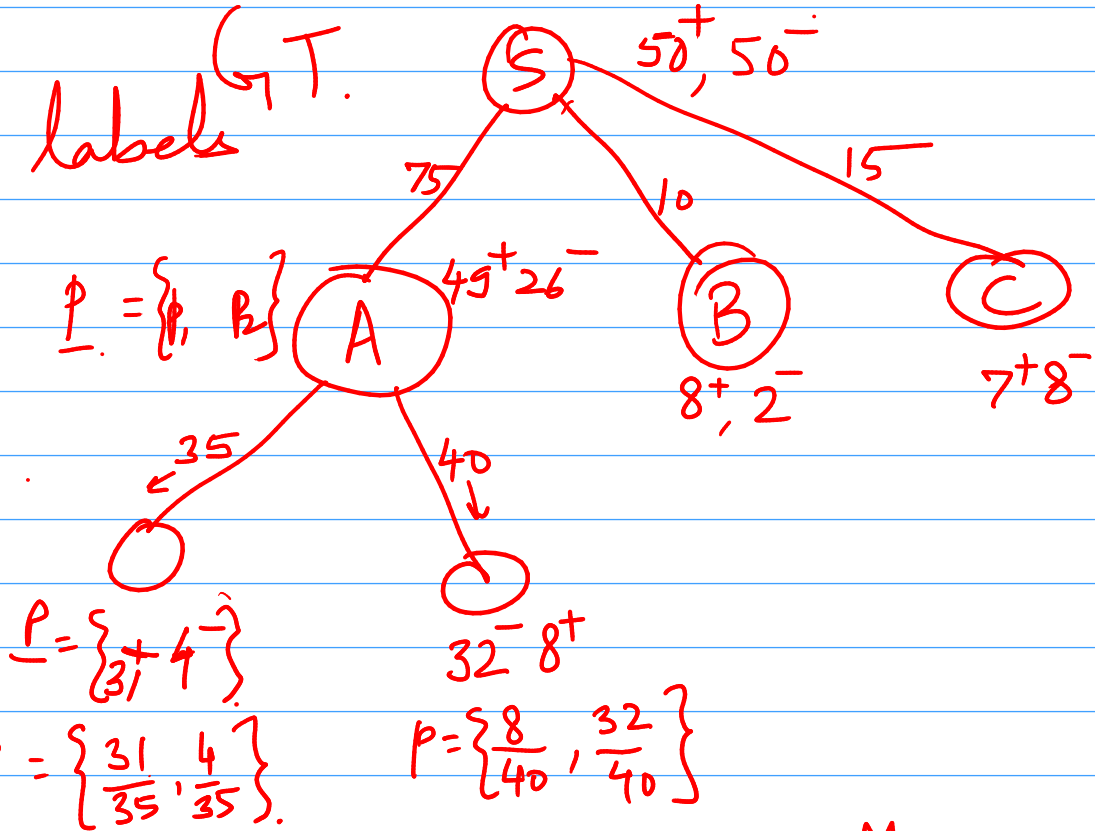


Entropy (S)

$$-\sum_{i=1}^{c} t_i^s \text{ Entropy (i)}$$

Probability of labels of examples in a node (set).

↳ T.

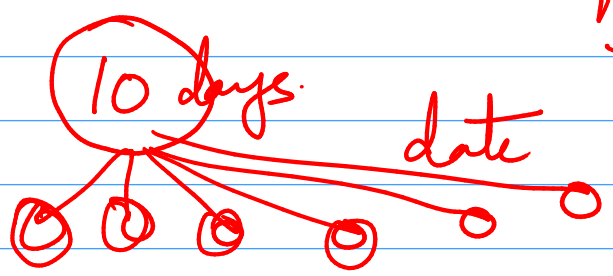


Weekend (Yes, No)

B'day (Yes, No)

Exams over (Yes, No)

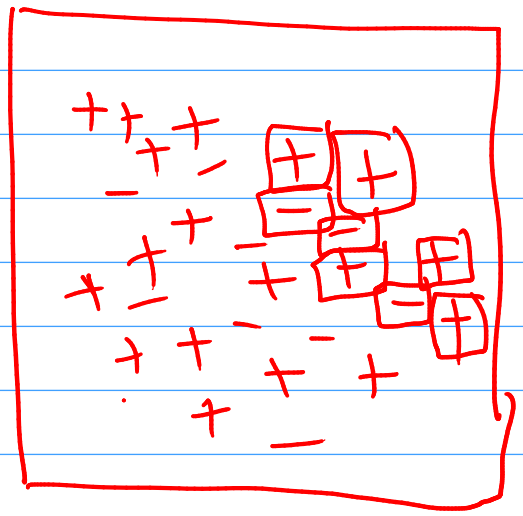
Date 1st/Sep (Unique value)



Memorized

{ dates is a not a good attribute to split

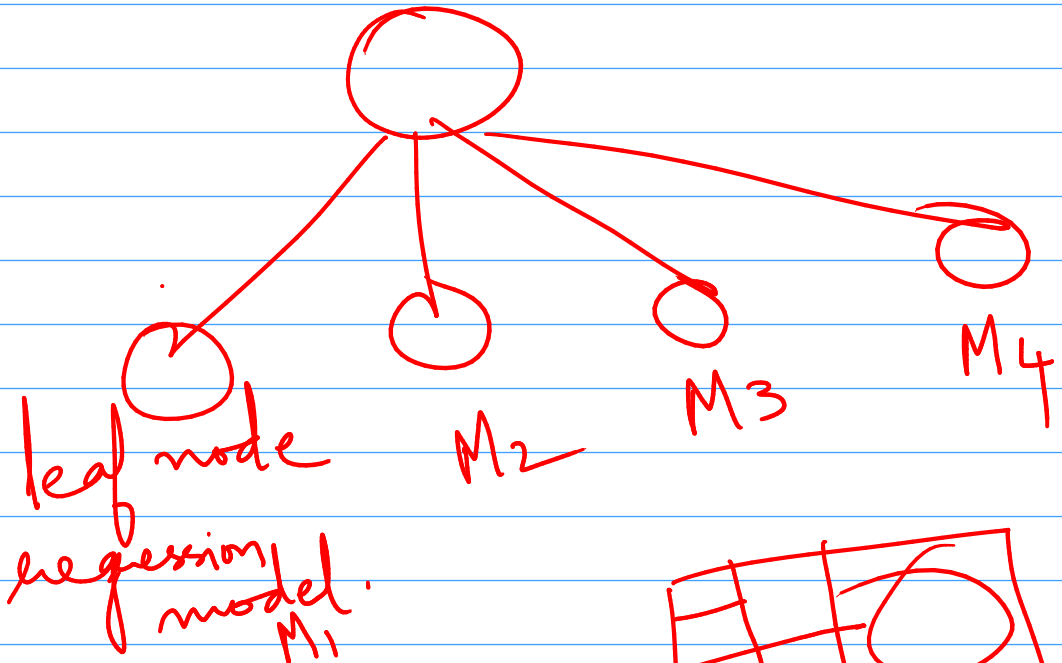
DTree creates a partitioning of the input space

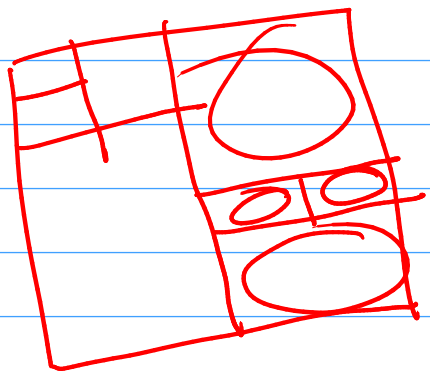


$VC_{dim} \infty$.

leaf nodes with single example

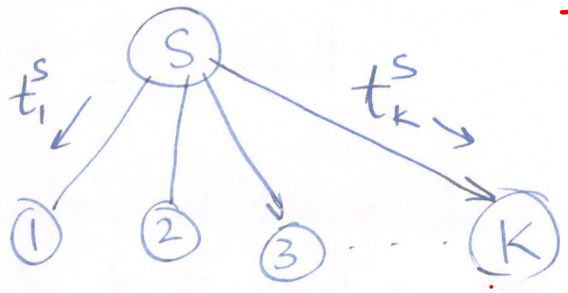
Useless machine



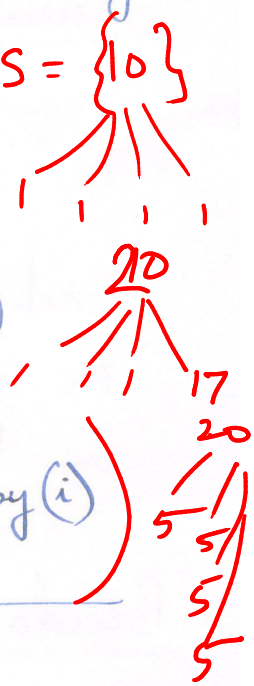


Information gain is more for attributes which have more number of values and therefore can lead to multiway splits.

— To avoid the bias in favour of attributes with a large number of values, we divide the information gain by a split gain.



$$\text{Split gain} = \sum_{i=1}^{k} t_i^S \log \frac{1}{t_i^S}$$



Information Gain Ratio:

$$\frac{\text{Entropy}(S) - \sum_i t_i^S \cdot \text{Entropy}(i)}{\left(\sum_{i=1}^{k} t_i^S \log \frac{1}{t_i^S}\right)}$$

Ratio

split gain

ID3 — Iterative Dichotomizer (Version 3) ↑ binary splits

Works for binary valued attributes

Uses information gain for choosing the split attribute.

recursion

C4.5

CHAID. CART.

VC dim of a decision tree?

The learned trees are usually very large.

This leads to low error on the training data.

But such large trees do not perform well on the test data.

i.e. they show poor generalization performance.

Solution to the overfitting problem: Stop growing the tree

Early stopping.

by limiting the number of iterations in the ID3 algorithm. depth of the tree.

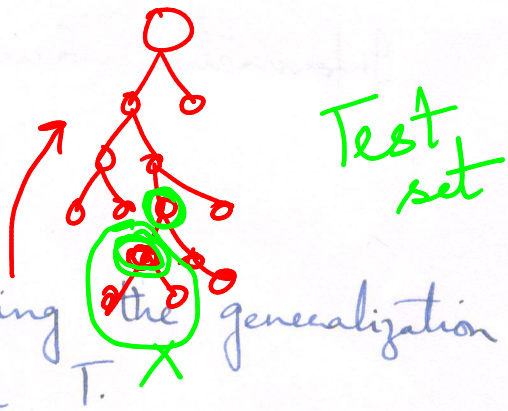
This results in a tree with bounded number of nodes.

Another way to address overfitting is to prune the tree after it is built.

Pruning: First construct a large decision tree.

Pruning is performed by a bottom-up walk on the large tree.

Each node might be replaced with one of its subtrees or with a leaf if it does not lead to much increase in the generalization error.



Pseudo Code for Pruning

Given: Some method $f(T)$ of estimating the generalization performance of the decision tree T.

For each node j in a bottom-up walk on T

find T' which minimizes $f(T')$ where T' is any one of the following: generalization error.

- the current tree after replacing node j with a leaf 1.
- the current tree after replacing node j with a leaf 0.
- the current tree after replacing node j with its left subtree.
- the current tree after replacing node j with its right subtree