CS 725 (Aug 29, 2024)

Splitting criterion : Most popular is "INFORMATION GAIN"

[Other examples include Gini impurity, entropy, etc.]

Say we have an attribute "a" that takes values from $V(a)$.

Consider a dataset S. Let S_γ be the subset of S where all instances have the attribute value of "a" set to γ.

$$IG(S,a) = H(S) - \sum_{\gamma \in V(a)} \frac{|S_\gamma|}{|S|} H(S_\gamma)$$

Pick the attribute that maximizes information gain (IG)

$$a^* = \underset{a}{\operatorname{argmax}}\, IG(S,a)$$

$$IG(S,a) = H(S) - \sum_{v \in V(a)} \frac{|S_v|}{|S|} H(S_v)$$

What's the highest value that IG can take ?

IG in information-theoretic terms is called "MUTUAL INFORMATION"

$$I(X,Y) = H(Y) - H(Y|X)$$
$$= H(X) - H(X|Y)$$

Note that $I(X,Y)$ is non-negative

CONDITIONAL ENTROPY $H(Y|X) = \sum_x P(X=x) H(Y|X=x)$

Q. Consider a dataset with two Boolean attributes x_1 and x_2 & a binary label y. What do we split with at the root node? (Use IG)

x_1	x_2	y
1	1	1
1	0	1
1	1	1
1	0	1
0	1	1
0	0	0

Given $\log_2 \frac{1}{3} = -1.5$

$\log_2 \frac{2}{3} = -0.5$

Ans : x_1

$IG(S, x_1) = H(S) - \left[\frac{4}{6} \times 0 + \frac{2}{6} \times 1\right] = H(S) - \frac{1}{3}$

$IG(S, x_2) = H(S) - \left[\frac{3}{6} \times 0 + \frac{3}{6} \times \left[-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right]\right]$

$= H(S) - \frac{5}{12}$

STOPPING CRITERION

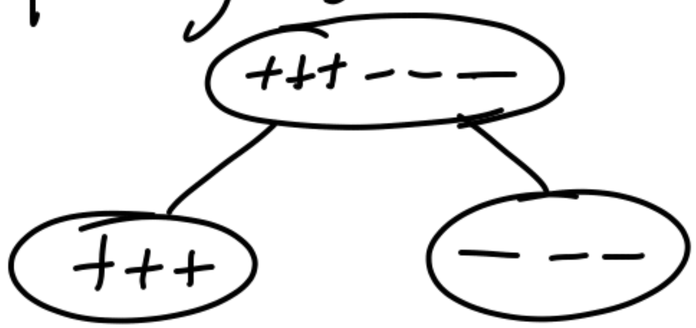
When do we stop building the tree?

Many stopping criteria possible for DT construction:

1. Stop splitting when the info. gain is below a threshold
2. Stop splitting when all instances have the same label
3. Stop splitting if the number of instances at a node falls below a threshold
4. Stop splitting when you hit a max depth

More about stopping criteria

① Always stop splitting if all the instances have the same label



Stop building further

② Don't split if the <u>remaining attributes</u> are identical across all the instances at a node

$a \in \{a_1, a_2, \ldots, a_n\}$

$b \in \{b_1, b_2, \ldots, b_m\}$

a	b	Y
a_2	b_3	0
a_2	b_3	1

IG of attribute $a = 0$

IG of attribute $b = 0$

The attributes don't distinguish the instances further

Should we split further if IG = 0 ?

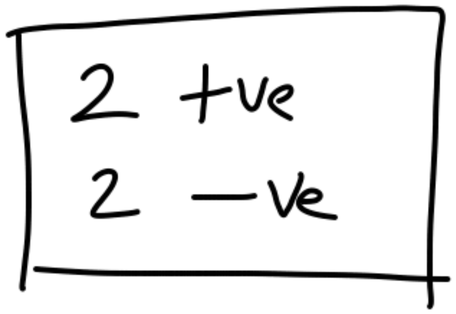
Consider the XOR dataset

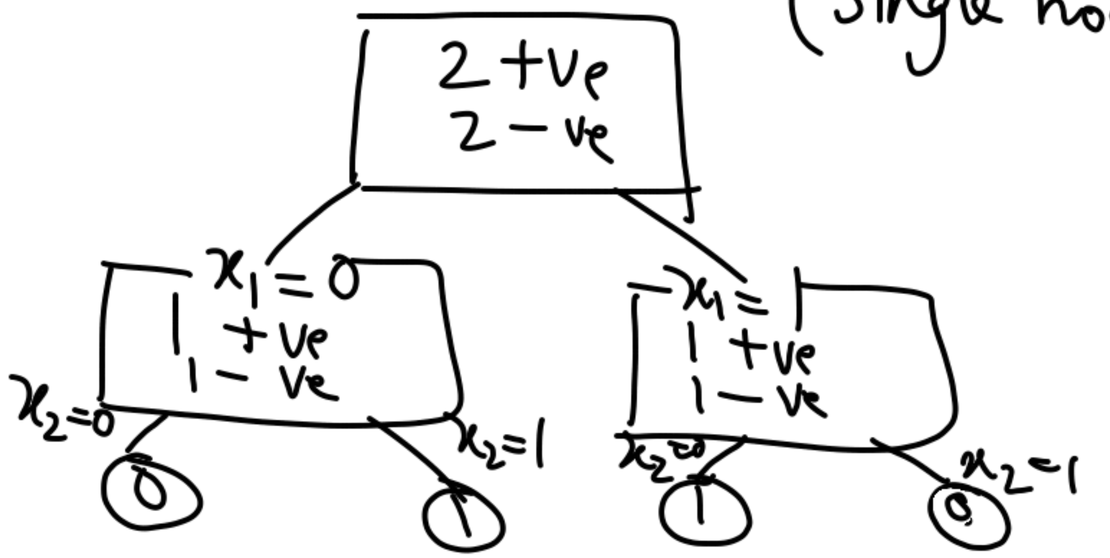
x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0

IG of $x_1 = 0$

IG of $x_2 = 0$

if we don't split, we get a decision Stump (Single node)





Continuous Attributes

If we have a continuous attribute $a \in \mathbb{R}$, then we need to define thresholds τ so that we can pose questions with binary responses like "Is $a \leq \tau$?".

How do we find these thresholds?

Procedure to find thresholds

① Let the values of a continuous attribute a across n instances be $v_1, \ldots, v_n$

② Sort these values $v_1, \ldots, v_n$ in increasing order

③ Compute midpoints $m_j = \frac{v_j + v_{j+1}}{2}$

④ Thresholds of interest are those m_j's whose surrounding instances have different labels

Example: Let the attribute a take values 10, 20, 30, 50, 100.
across 5 training instances with labels 0, 0, 0, 1, 0.

The midpoints will be 15, 25, 40, 75

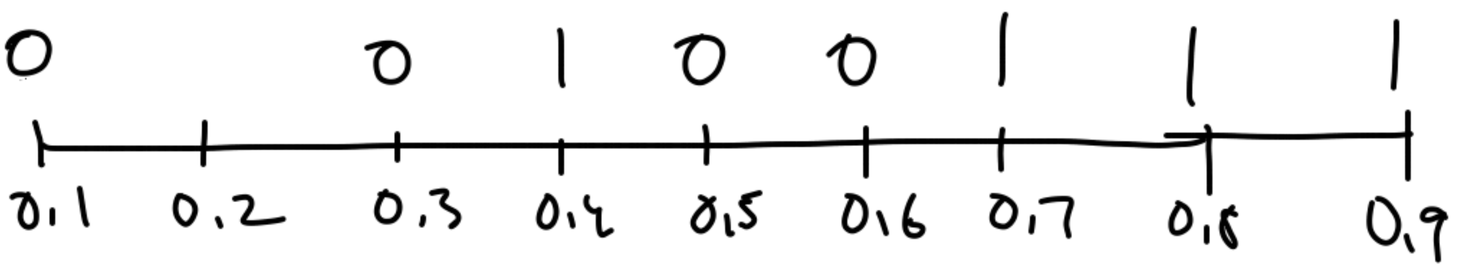
these can be used as thresholds
to create binary questions of the form
$a \leq 40$? or $a > 75$?

Q. Consider the dataset below with $x_1, x_2 \in \mathbb{R}$ and binary label y.

x_1	x_2	y
0.1	0.4	1
0.2	0.7	1
0.3	0.3	0
0.4	0.9	1
0.5	0.8	1
0.7	0.1	0
0.8	0.6	0
1.0	0.5	0

Which of x_1, x_2 should we split on? (Use IG)

For x_2



Both x_1, x_2 are equally good options

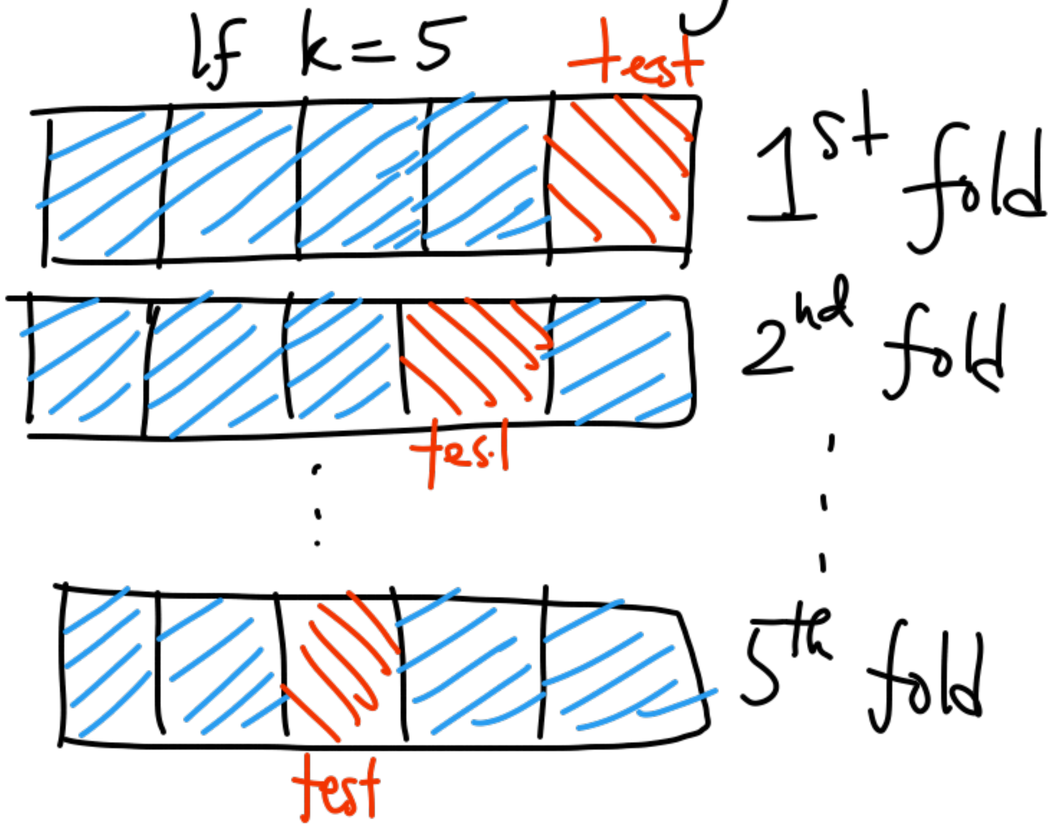
Random Forests : An ensemble of DTs

[E.g. of an aggregation technique called "bagging"
BOOTSTRAP AGGREGATION]

k-fold CROSS VALIDATION (CV)

CV is invoked typically for small-data settings

In k-fold cross-validation:



With CV, guaranteed that every sample appears as a test instance

Train a predictor each for each of the k folds.
Compute the average test error (loss) across the
k test splits.

Typically, $k=5$, $k=10$ are popular

Leave-one out validation → Test split in each fold is a single instance