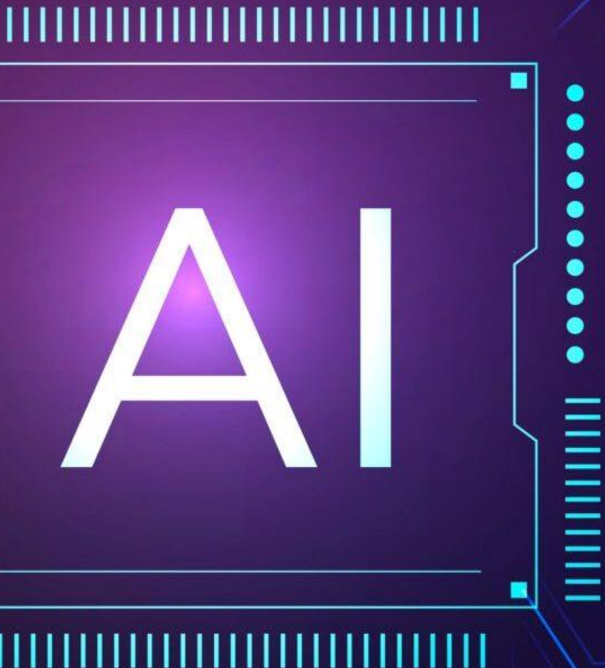


ID3 DECISION TREE

Nguyễn Ngọc Thảo – Nguyễn Hải Minh
{nnthao, nhminh}@fit.hcmus.edu.vn

Outline

- Supervised learning: A brief revision
- ID3 decision tree algorithm

Supervised learning: Training

- Consider a labeled training set of N examples.

$$(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$$

  - where each y_j was generated by an unknown function $y = f(x)$.
- The output $\boldsymbol{y_j}$ is called ground truth, i.e., the true answer that the model must predict.
- The training process finds a hypothesis h such that $\boldsymbol{h} \approx \boldsymbol{f}$.

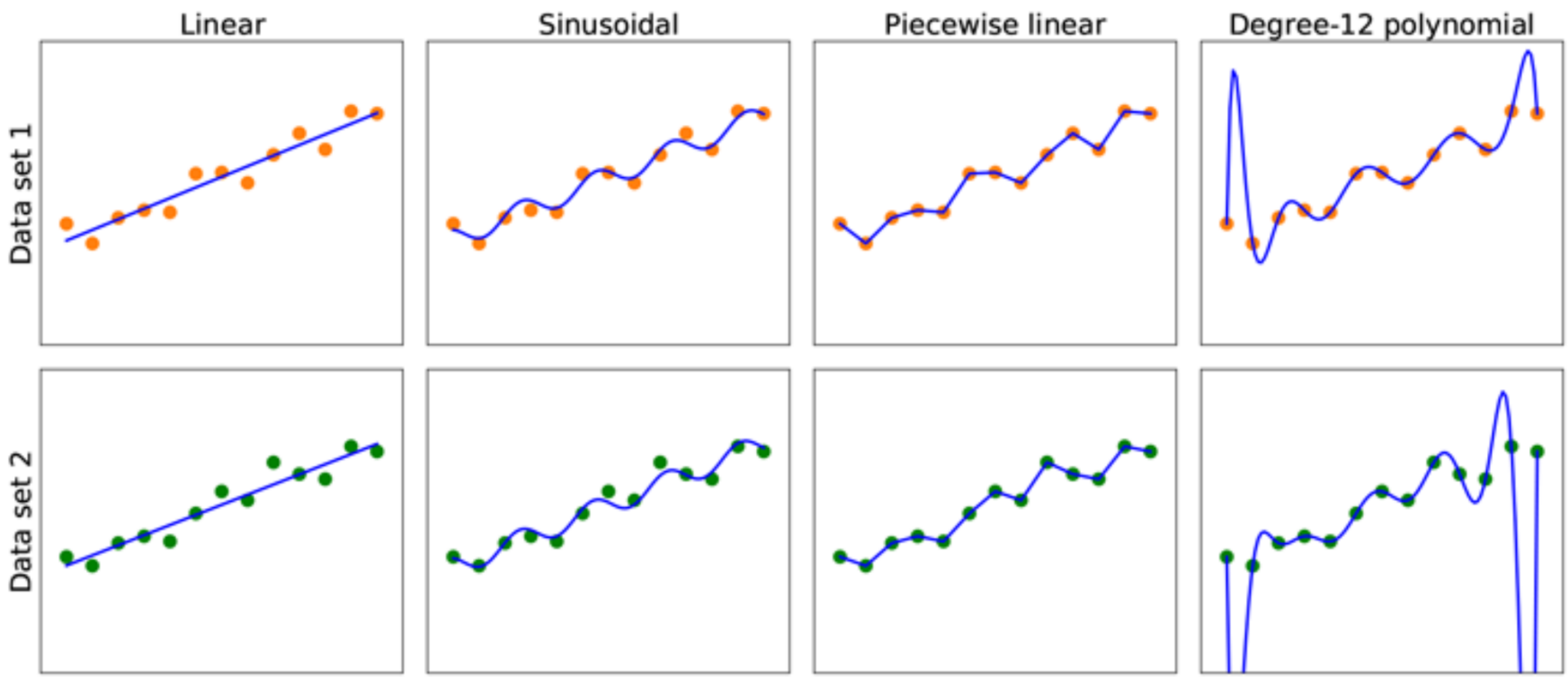
Supervised learning: Hypothesis space

- h is drawn from a hypothesis space H of possible functions.
  - E.g., H might be the set of polynomials of degree 3; or the set of 3-SAT Boolean logic formulas.
- Choose H by some prior knowledge about the process that generated the data or exploratory data analysis (EDA).
  - EDA examines the data with statistical tests and visualizations to get some insight into what hypothesis space might be appropriate.
- Or just try multiple hypothesis spaces and evaluate which one works best.

Supervised learning: Hypothesis

- The hypothesis h is consistent if it agrees with the true function f on all training observations, i.e., $\forall x_i \; h(x_i) = y_i$.
  - For continuous data, we instead look for a best-fit function for which each $h(x_i)$ is close to y_i.
- Ockham's razor: Select the simplest consistent hypothesis.

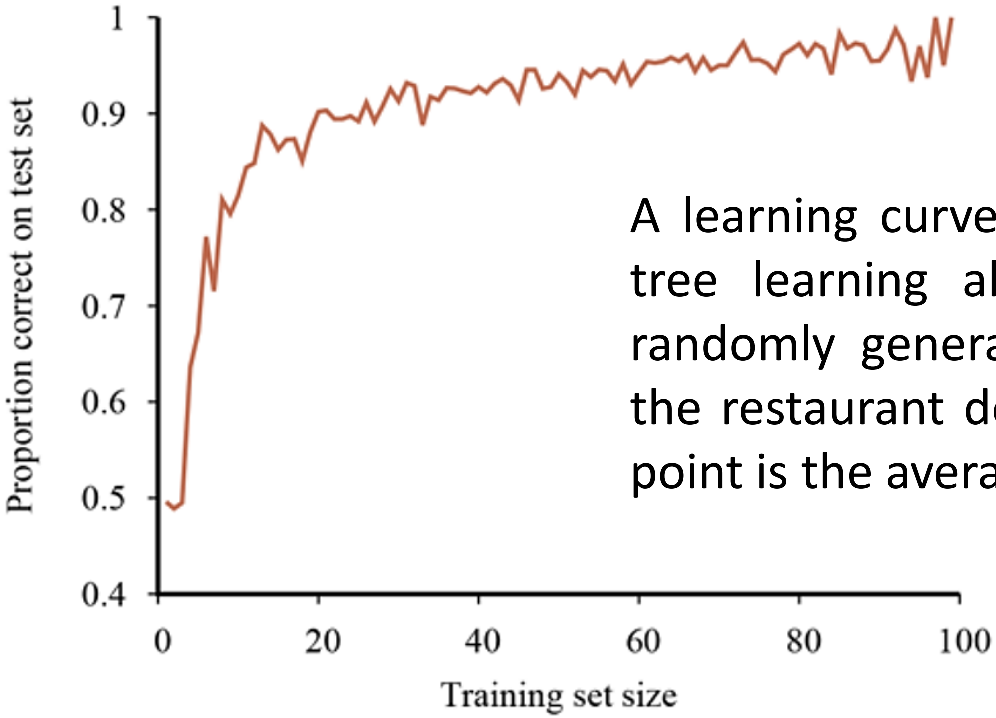
Supervised learning: Hypothesis



Finding hypotheses to fit data. Top row: four plots of best-fit functions from four different hypothesis spaces trained on data set 1. Bottom row: the same four functions, but trained on a slightly different data set (sampled from the same $f(x)$ function).

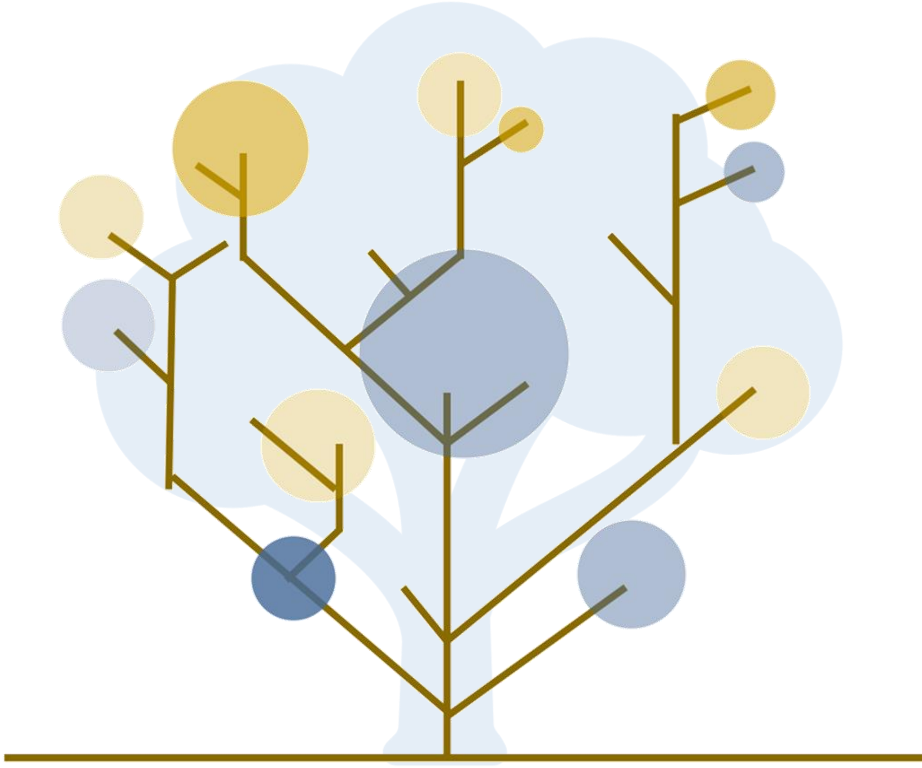
Supervised learning: Testing

- The quality of the hypothesis h depends on how accurately it predicts the observations in the test set $\rightarrow$ generalization.
  - The test set must use the same distribution over example space as training set.



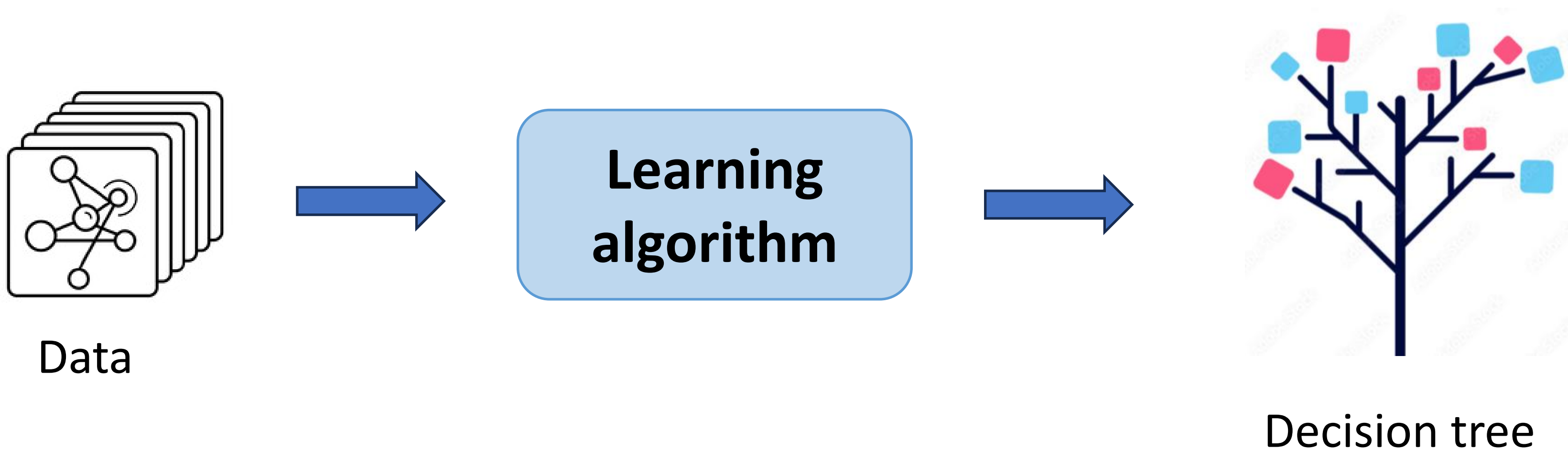
A learning curve for the decision tree learning algorithm on 100 randomly generated examples in the restaurant domain. Each data point is the average of 20 trials.

ID3
Decision Tree



What is a decision tree?

- A decision tree is a SL algorithm that predicts the output by learning decision rules inferred from the features in the data.



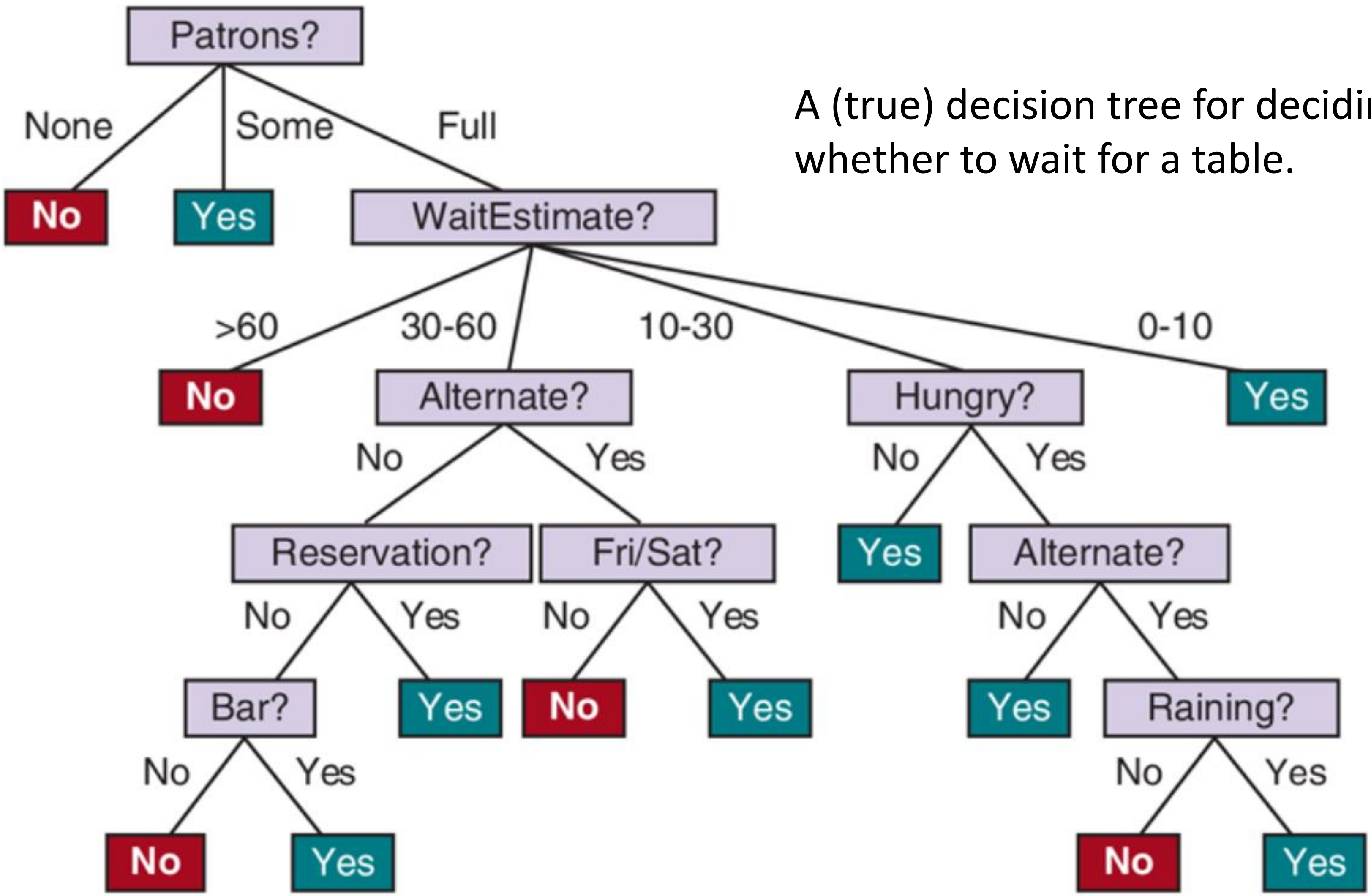
- It is offen the building blocks for more complex algorithms, such as random forests and gradient boosting machines.

Example problem: Restaurant waiting

Predicting whether a certain person will wait to have a seat in a restaurant.

1. **Alternate:** is there an alternative restaurant nearby?
2. **Bar:** is there a comfortable bar area to wait in?
3. **Fri/Sat:** is today Friday or Saturday?
4. **Hungry:** are we hungry?
5. **Patrons:** number of people in the restaurant (None, Some, Full)
6. **Price:** price range ($, $$, $$$)
7. **Raining:** is it raining outside?
8. **Reservation:** have we made a reservation?
9. **Type:** kind of restaurant (French, Italian, Thai, Burger)
10. **WaitEstimate:** estimated waiting time (0-10, 10-30, 30-60, >60)

Example problem: Restaurant waiting



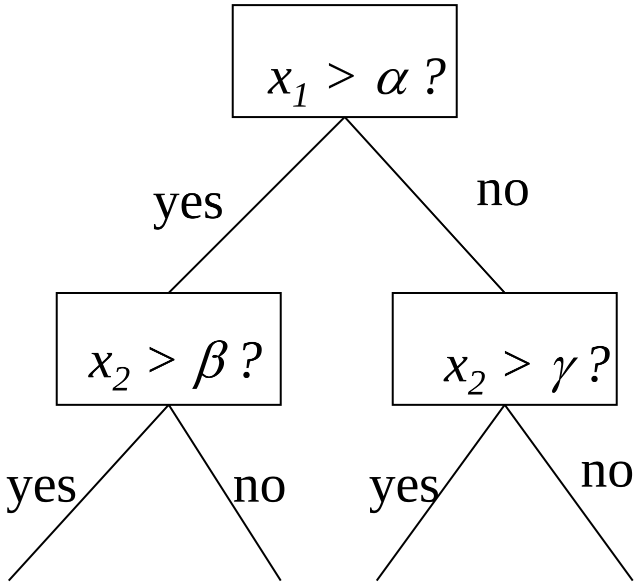
A (true) decision tree for deciding whether to wait for a table.

Example problem: Restaurant waiting

Example	Input Attributes										Output
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
$\mathbf{x}_1$	*Yes*	*No*	*No*	*Yes*	*Some*	*$$$*	*No*	*Yes*	*French*	*0–10*	$y_1 = Yes$
$\mathbf{x}_2$	*Yes*	*No*	*No*	*Yes*	*Full*	*$*	*No*	*No*	*Thai*	*30–60*	$y_2 = No$
$\mathbf{x}_3$	*No*	*Yes*	*No*	*No*	*Some*	*$*	*No*	*No*	*Burger*	*0–10*	$y_3 = Yes$
$\mathbf{x}_4$	*Yes*	*No*	*Yes*	*Yes*	*Full*	*$*	*Yes*	*No*	*Thai*	*10–30*	$y_4 = Yes$
$\mathbf{x}_5$	*Yes*	*No*	*Yes*	*No*	*Full*	*$$$*	*No*	*Yes*	*French*	*>60*	$y_5 = No$
$\mathbf{x}_6$	*No*	*Yes*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Italian*	*0–10*	$y_6 = Yes$
$\mathbf{x}_7$	*No*	*Yes*	*No*	*No*	*None*	*$*	*Yes*	*No*	*Burger*	*0–10*	$y_7 = No$
$\mathbf{x}_8$	*No*	*No*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Thai*	*0–10*	$y_8 = Yes$
$\mathbf{x}_9$	*No*	*Yes*	*Yes*	*No*	*Full*	*$*	*Yes*	*No*	*Burger*	*>60*	$y_9 = No$
$\mathbf{x}_{10}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$$$*	*No*	*Yes*	*Italian*	*10–30*	$y_{10} = No$
$\mathbf{x}_{11}$	*No*	*No*	*No*	*No*	*None*	*$*	*No*	*No*	*Thai*	*0–10*	$y_{11} = No$
$\mathbf{x}_{12}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$*	*No*	*No*	*Burger*	*30–60*	$y_{12} = Yes$

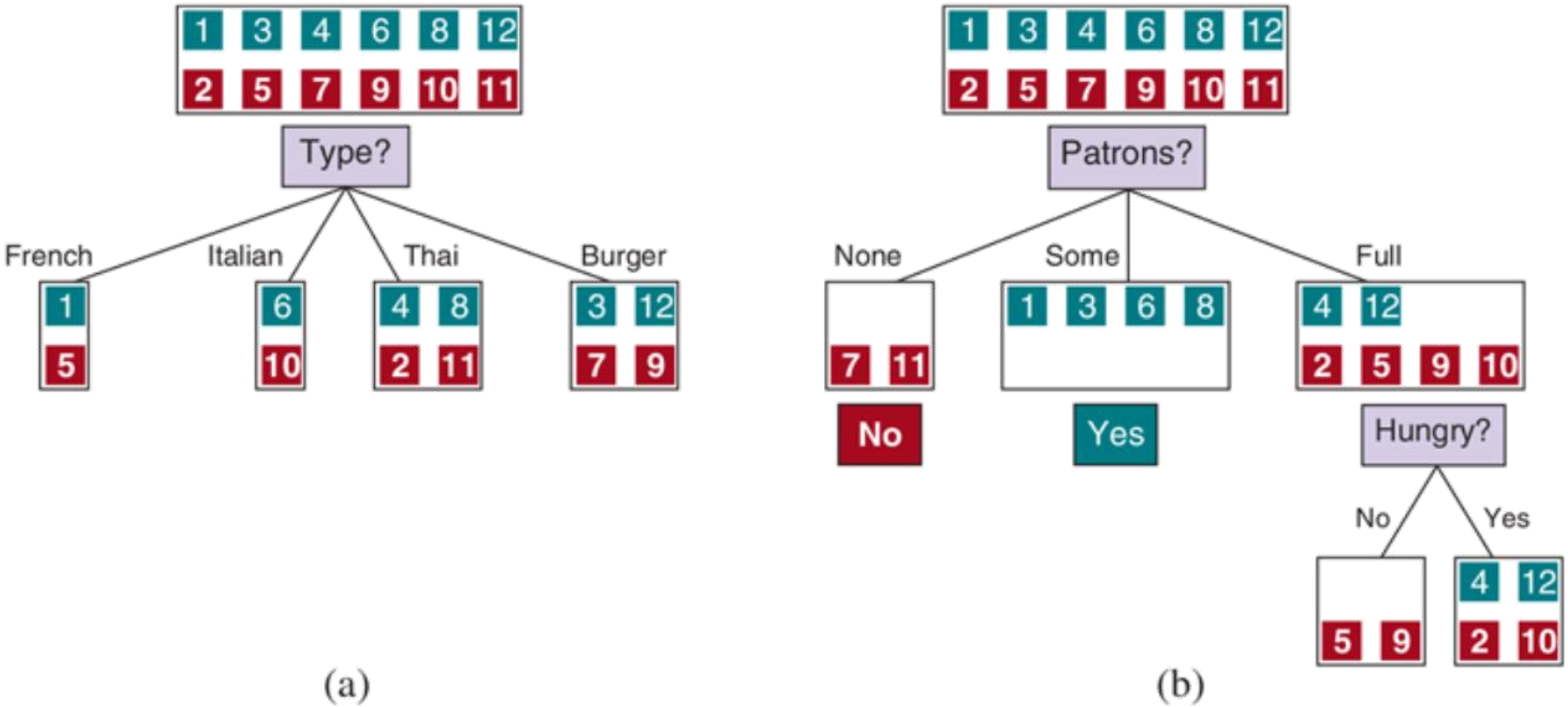
Learning decision trees

- Divide and conquer: Split data into smaller and smaller subsets
- Splits are usually on a single variable



- After splitting up, each outcome is a new decision tree learning problem with fewer examples and one less attribute.

Learning decision trees



Splitting the examples by testing on attributes. At each node we show the positive (light boxes) and negative (dark boxes) examples remaining. (a) Splitting on Type brings us no nearer to distinguishing between positive and negative examples. (b) Splitting on Patrons does a good job of separating positive and negative examples. After splitting on Patrons, Hungry is a fairly good second test.

ID3 Decision tree: Pseudo-code

function LEARN-ECISION-TREE(*examples*, *attributes*, *parent_examples*)
returns a tree
  if *examples* is empty
    then return PLURALITY-VALUE(*parent_examples*)
  else if all *examples* have the same classification
    then return the classification
  else if *attributes* is empty
    then return PLURALITY-VALUE(*examples*)
  else
    ...

No examples left 3

Remaining examples are all pos/neg 2

No attributes left but examples are still pos & neg 4

The decision tree learning algorithm. The function PLURALITY-VALUE selects the most common output value among a set of examples, breaking ties randomly.

ID3 Decision tree: Pseudo-code

function LEARN-DECISION-TREE($examples, attributes, parent\_examples$)
returns a tree

There are still attributes to split the examples

1

  ...
  else
    $A \leftarrow argmax_{a \in attributes}$ IMPORTANCE($a, examples$)
    $tree \leftarrow$ a new decision tree with root test A
    for each value v of A **do**
      $exs \leftarrow \{e : e \in examples$ **and** $e.A = v\}$
      $subtree \leftarrow$ LEARN-DECISION-TREE($exs, attributes - A, examples$)
      add a branch to $tree$ with label ($A = v$) and subtree $subtree$
    return $tree$

The decision tree learning algorithm. The function IMPORTANCE evaluates the profitability of attributes.

ID3 Decision tree algorithm

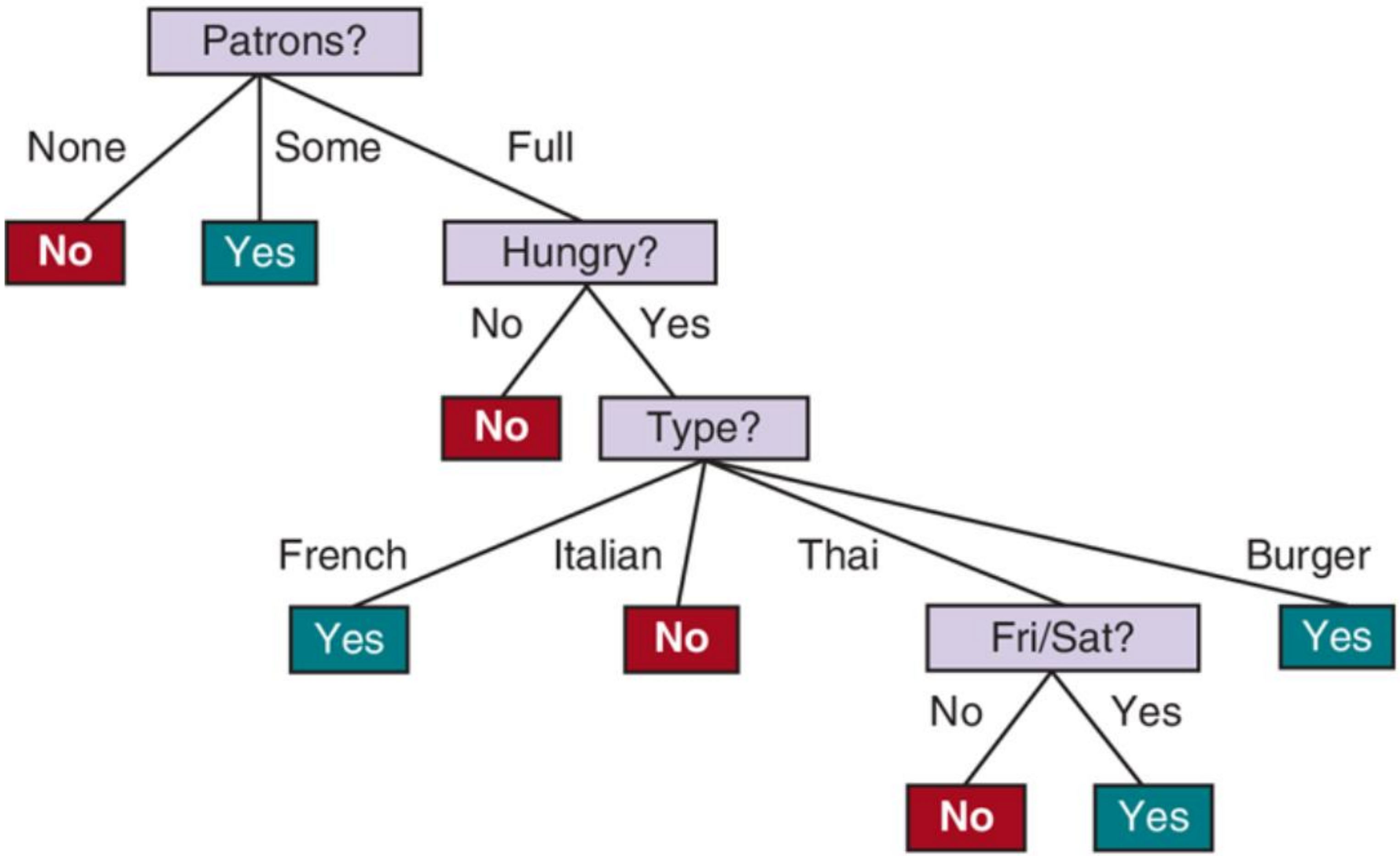
1. There are **some** positive and **some** negative examples → choose the **best** attribute to split them

2. The remaining examples are **all** positive (or **all** negative), → DONE, it is possible to answer Yes or No.

3. **No** examples left at a branch → return a default value.
   - No example has been observed for a combination of attribute values
   - The default value is calculated from the plurality classification of all the examples that were used in constructing the node's parent.

4. **No** attributes left but both positive and negative examples → return the plurality classification of remaining ones.
   - Examples of the same description, but different classifications
   - It is due to an error or noise in the data, nondeterministic domain, or no observation of an attribute that would distinguish the examples.

Example problem: Restaurant waiting



The decision tree induced from the 12-example training set.

Example problem: Restaurant waiting

- The induced decision tree can classify all the examples without tests for Raining and Reservation.
- It can detect interesting and previously unsuspected pattern.
  - E.g., the customers will wait for Thai food on weekends.
- It is also bound to make some mistakes for cases where it has seen no examples.
  - E.g., how about a situation in which the wait is 0–10 minutes, the restaurant is full, yet the customer is not hungry?

Decision tree: Inductive learning

- **Simplest:** Construct a decision tree with one leaf for every example

  $\rightarrow$ memory based learning

  $\rightarrow$ *worse generalization*

- **Advanced:** Split on each variable so that the purity of each split increases (i.e. either only yes or only no).

A purity measure with entropy

- The Entropy measures the uncertainty of a random variable V with values v_k having probability $P(v_k)$ is defined as

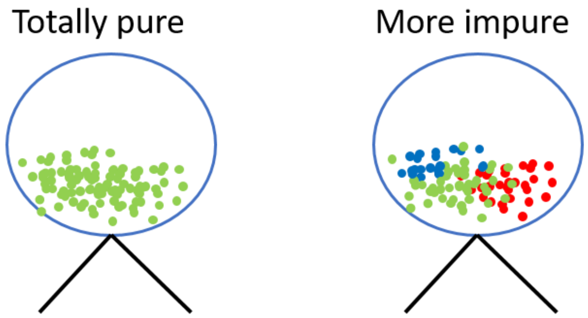
$$\boldsymbol{H(V) = \sum_k P(v_k)\, log_2 \frac{1}{P(v_k)} = -\sum_k P(v_k)\, log_2\, P(v_k)}$$

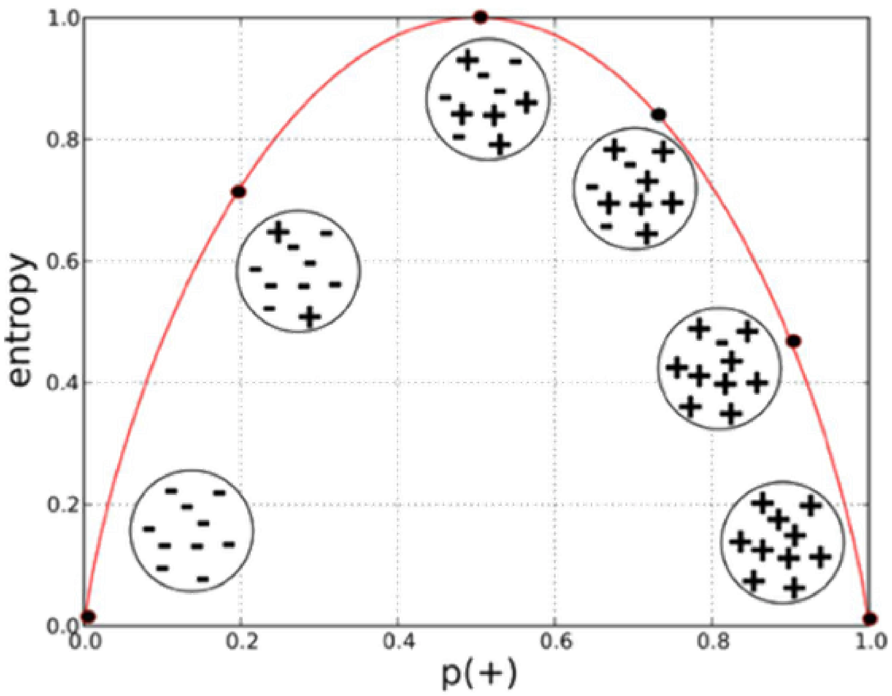
  - It is fundamental quantity in information theory.

- The information gain (IG) for an attribute A is the expected reduction in entropy from before to after splitting data on A.

A purity measure with entropy

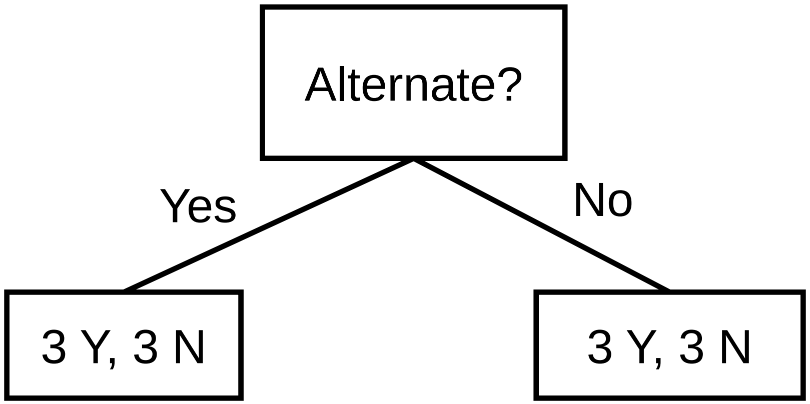
- Entropy is maximal when all possibilities are equally likely.
- Entropy is zero in a pure "yes" (or pure "no") node.





- Decision tree aims to decrease the entropy while increasing the information gain in each node.

Example problem: Restaurant waiting



Example	Input Attributes										Output
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
$\mathbf{x}_1$	*Yes*	*No*	*No*	*Yes*	*Some*	*$$$*	*No*	*Yes*	*French*	*0–10*	$y_1 = Yes$
$\mathbf{x}_2$	*Yes*	*No*	*No*	*Yes*	*Full*	*$*	*No*	*No*	*Thai*	*30–60*	$y_2 = No$
$\mathbf{x}_3$	*No*	*Yes*	*No*	*No*	*Some*	*$*	*No*	*No*	*Burger*	*0–10*	$y_3 = Yes$
$\mathbf{x}_4$	*Yes*	*No*	*Yes*	*Yes*	*Full*	*$*	*Yes*	*No*	*Thai*	*10–30*	$y_4 = Yes$
$\mathbf{x}_5$	*Yes*	*No*	*Yes*	*No*	*Full*	*$$$*	*No*	*Yes*	*French*	*>60*	$y_5 = No$
$\mathbf{x}_6$	*No*	*Yes*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Italian*	*0–10*	$y_6 = Yes$
$\mathbf{x}_7$	*No*	*Yes*	*No*	*No*	*None*	*$*	*Yes*	*No*	*Burger*	*0–10*	$y_7 = No$
$\mathbf{x}_8$	*No*	*No*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Thai*	*0–10*	$y_8 = Yes$
$\mathbf{x}_9$	*No*	*Yes*	*Yes*	*No*	*Full*	*$*	*Yes*	*No*	*Burger*	*>60*	$y_9 = No$
$\mathbf{x}_{10}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$$$*	*No*	*Yes*	*Italian*	*10–30*	$y_{10} = No$
$\mathbf{x}_{11}$	*No*	*No*	*No*	*No*	*None*	*$*	*No*	*No*	*Thai*	*0–10*	$y_{11} = No$
$\mathbf{x}_{12}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$*	*No*	*No*	*Burger*	*30–60*	$y_{12} = Yes$

- Calculate the Entropy of the whole data set

$$H(S) = -\left({}^{6}\!/_{12}\right)\log_2\left({}^{6}\!/_{12}\right) - \left({}^{6}\!/_{12}\right)\log_2\left({}^{6}\!/_{12}\right) = 1$$

- Calculate Average entropy of attribute Alternate?

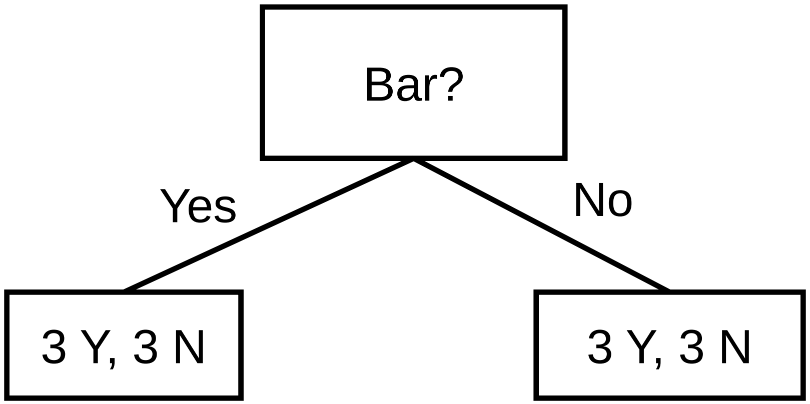
$$AE_{Alternate?} = P(Alt = Y) \times H(Alt = Y) + P(Alt = N) \times H(Alt = N) = 1$$

$$= \frac{6}{12}\left[-\left(\frac{3}{6}\log_2\frac{3}{6}\right) - \left(\frac{3}{6}\log_2\frac{3}{6}\right)\right] + \frac{6}{12}\left[-\left(\frac{3}{6}\log_2\frac{3}{6}\right) - \left(\frac{3}{6}\log_2\frac{3}{6}\right)\right]$$

- Calculate Information gain of attribute Alternate?

$$IG(Alternate?) = H(S) - AE_{Alternate?} = 1 - 1 = 0$$

Example problem: Restaurant waiting



Example	Input Attributes										Output
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
$\mathbf{x}_1$	Yes	No	No	Yes	Some	$$$	No	Yes	French	0–10	$y_1 = Yes$
$\mathbf{x}_2$	Yes	No	No	Yes	Full	$	No	No	Thai	30–60	$y_2 = No$
$\mathbf{x}_3$	No	Yes	No	No	Some	$	No	No	Burger	0–10	$y_3 = Yes$
$\mathbf{x}_4$	Yes	No	Yes	Yes	Full	$	Yes	No	Thai	10–30	$y_4 = Yes$
$\mathbf{x}_5$	Yes	No	Yes	No	Full	$$$	No	Yes	French	>60	$y_5 = No$
$\mathbf{x}_6$	No	Yes	No	Yes	Some	$$	Yes	Yes	Italian	0–10	$y_6 = Yes$
$\mathbf{x}_7$	No	Yes	No	No	None	$	Yes	No	Burger	0–10	$y_7 = No$
$\mathbf{x}_8$	No	No	No	Yes	Some	$$	Yes	Yes	Thai	0–10	$y_8 = Yes$
$\mathbf{x}_9$	No	Yes	Yes	No	Full	$	Yes	No	Burger	>60	$y_9 = No$
$\mathbf{x}_{10}$	Yes	Yes	Yes	Yes	Full	$$$	No	Yes	Italian	10–30	$y_{10} = No$
$\mathbf{x}_{11}$	No	No	No	No	None	$	No	No	Thai	0–10	$y_{11} = No$
$\mathbf{x}_{12}$	Yes	Yes	Yes	Yes	Full	$	No	No	Burger	30–60	$y_{12} = Yes$

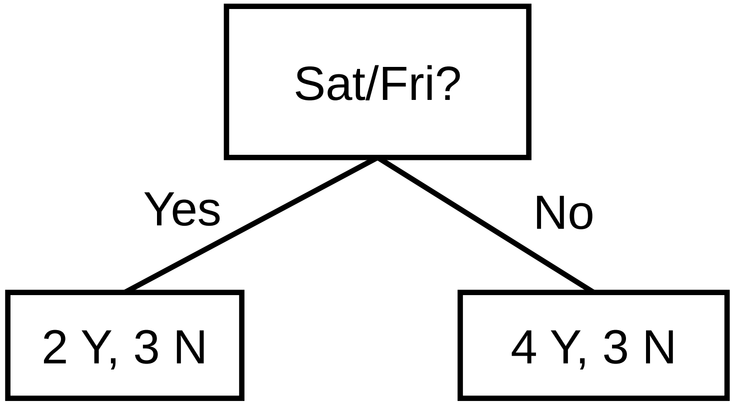
- Calculate Average entropy of attribute Bar?

$$AE_{Bar?} = \frac{6}{12}\left[-\left(\frac{3}{6}\log_2\frac{3}{6}\right) - \left(\frac{3}{6}\log_2\frac{3}{6}\right)\right] + \frac{6}{12}\left[-\left(\frac{3}{6}\log_2\frac{3}{6}\right) - \left(\frac{3}{6}\log_2\frac{3}{6}\right)\right] = 1$$

- Calculate Information gain of attribute Bar?

$$IG(Bar?) = H(S) - AE_{Bar?} = 1 - 1 = 0$$

Example problem: Restaurant waiting



Example	Input Attributes										Output
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
$\mathbf{x}_1$	*Yes*	*No*	*No*	*Yes*	*Some*	*$$$*	*No*	*Yes*	*French*	*0–10*	$y_1 = Yes$
$\mathbf{x}_2$	*Yes*	*No*	*No*	*Yes*	*Full*	*$*	*No*	*No*	*Thai*	*30–60*	$y_2 = No$
$\mathbf{x}_3$	*No*	*Yes*	*No*	*No*	*Some*	*$*	*No*	*No*	*Burger*	*0–10*	$y_3 = Yes$
$\mathbf{x}_4$	*Yes*	*No*	*Yes*	*Yes*	*Full*	*$*	*Yes*	*No*	*Thai*	*10–30*	$y_4 = Yes$
$\mathbf{x}_5$	*Yes*	*No*	*Yes*	*No*	*Full*	*$$$*	*No*	*Yes*	*French*	*>60*	$y_5 = No$
$\mathbf{x}_6$	*No*	*Yes*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Italian*	*0–10*	$y_6 = Yes$
$\mathbf{x}_7$	*No*	*Yes*	*No*	*No*	*None*	*$*	*Yes*	*No*	*Burger*	*0–10*	$y_7 = No$
$\mathbf{x}_8$	*No*	*No*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Thai*	*0–10*	$y_8 = Yes$
$\mathbf{x}_9$	*No*	*Yes*	*Yes*	*No*	*Full*	*$*	*Yes*	*No*	*Burger*	*>60*	$y_9 = No$
$\mathbf{x}_{10}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$$$*	*No*	*Yes*	*Italian*	*10–30*	$y_{10} = No$
$\mathbf{x}_{11}$	*No*	*No*	*No*	*No*	*None*	*$*	*No*	*No*	*Thai*	*0–10*	$y_{11} = No$
$\mathbf{x}_{12}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$*	*No*	*No*	*Burger*	*30–60*	$y_{12} = Yes$

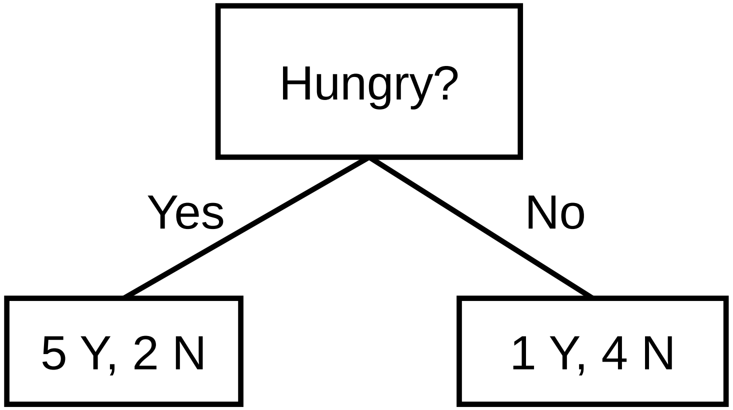
- Calculate Average entropy of attribute Sat/Fri?

$$AE_{Sat/Fri?} = \frac{5}{12}\left[-\left(\frac{2}{5}\log_2\frac{2}{5}\right)-\left(\frac{3}{5}\log_2\frac{3}{5}\right)\right]+\frac{7}{12}\left[-\left(\frac{4}{7}\log_2\frac{4}{7}\right)-\left(\frac{3}{7}\log_2\frac{3}{7}\right)\right] = 0.979$$

- Calculate Information gain of attribute Sat/Fri?

$$IG(Sat/Fri?) = H(S) - AE_{Sat/Fri?} = 1 - 0.979 = 0.021$$

Example problem: Restaurant waiting



Example	Input Attributes										Output
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
$\mathbf{x}_1$	Yes	No	No	Yes	Some	$$$	No	Yes	French	0–10	$y_1 = Yes$
$\mathbf{x}_2$	Yes	No	No	Yes	Full	$	No	No	Thai	30–60	$y_2 = No$
$\mathbf{x}_3$	No	Yes	No	No	Some	$	No	No	Burger	0–10	$y_3 = Yes$
$\mathbf{x}_4$	Yes	No	Yes	Yes	Full	$	Yes	No	Thai	10–30	$y_4 = Yes$
$\mathbf{x}_5$	Yes	No	Yes	No	Full	$$$	No	Yes	French	>60	$y_5 = No$
$\mathbf{x}_6$	No	Yes	No	Yes	Some	$$	Yes	Yes	Italian	0–10	$y_6 = Yes$
$\mathbf{x}_7$	No	Yes	No	No	None	$	Yes	No	Burger	0–10	$y_7 = No$
$\mathbf{x}_8$	No	No	No	Yes	Some	$$	Yes	Yes	Thai	0–10	$y_8 = Yes$
$\mathbf{x}_9$	No	Yes	Yes	No	Full	$	Yes	No	Burger	>60	$y_9 = No$
$\mathbf{x}_{10}$	Yes	Yes	Yes	Yes	Full	$$$	No	Yes	Italian	10–30	$y_{10} = No$
$\mathbf{x}_{11}$	No	No	No	No	None	$	No	No	Thai	0–10	$y_{11} = No$
$\mathbf{x}_{12}$	Yes	Yes	Yes	Yes	Full	$	No	No	Burger	30–60	$y_{12} = Yes$

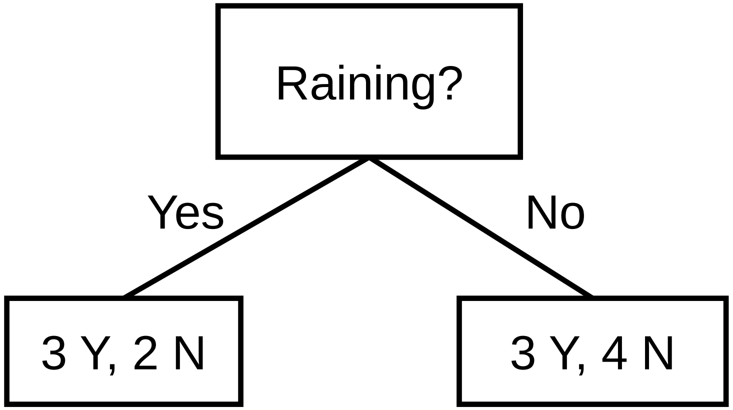
- Calculate Average entropy of attribute Hungry?

$$AE_{Hungry?} = \frac{7}{12}\left[-\left(\frac{5}{7}\log_2\frac{5}{7}\right)-\left(\frac{2}{7}\log_2\frac{2}{7}\right)\right]+\frac{5}{12}\left[-\left(\frac{1}{5}\log_2\frac{1}{5}\right)-\left(\frac{4}{5}\log_2\frac{4}{5}\right)\right] = 0.804$$

- Calculate Information gain of attribute Hungry?

$$IG(Hungry?) = H(S) - AE_{Hungry?} = 1 - 0.804 = 0.196$$

Example problem: Restaurant waiting



Example	Input Attributes										Output
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
$\mathbf{x}_1$	*Yes*	*No*	*No*	*Yes*	*Some*	*\$\$\$*	*No*	*Yes*	*French*	*0–10*	$y_1 = Yes$
$\mathbf{x}_2$	*Yes*	*No*	*No*	*Yes*	*Full*	*\$*	*No*	*No*	*Thai*	*30–60*	$y_2 = No$
$\mathbf{x}_3$	*No*	*Yes*	*No*	*No*	*Some*	*\$*	*No*	*No*	*Burger*	*0–10*	$y_3 = Yes$
$\mathbf{x}_4$	*Yes*	*No*	*Yes*	*Yes*	*Full*	*\$*	*Yes*	*No*	*Thai*	*10–30*	$y_4 = Yes$
$\mathbf{x}_5$	*Yes*	*No*	*Yes*	*No*	*Full*	*\$\$\$*	*No*	*Yes*	*French*	*>60*	$y_5 = No$
$\mathbf{x}_6$	*No*	*Yes*	*No*	*Yes*	*Some*	*\$\$*	*Yes*	*Yes*	*Italian*	*0–10*	$y_6 = Yes$
$\mathbf{x}_7$	*No*	*Yes*	*No*	*No*	*None*	*\$*	*Yes*	*No*	*Burger*	*0–10*	$y_7 = No$
$\mathbf{x}_8$	*No*	*No*	*No*	*Yes*	*Some*	*\$\$*	*Yes*	*Yes*	*Thai*	*0–10*	$y_8 = Yes$
$\mathbf{x}_9$	*No*	*Yes*	*Yes*	*No*	*Full*	*\$*	*Yes*	*No*	*Burger*	*>60*	$y_9 = No$
$\mathbf{x}_{10}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*\$\$\$*	*No*	*Yes*	*Italian*	*10–30*	$y_{10} = No$
$\mathbf{x}_{11}$	*No*	*No*	*No*	*No*	*None*	*\$*	*No*	*No*	*Thai*	*0–10*	$y_{11} = No$
$\mathbf{x}_{12}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*\$*	*No*	*No*	*Burger*	*30–60*	$y_{12} = Yes$

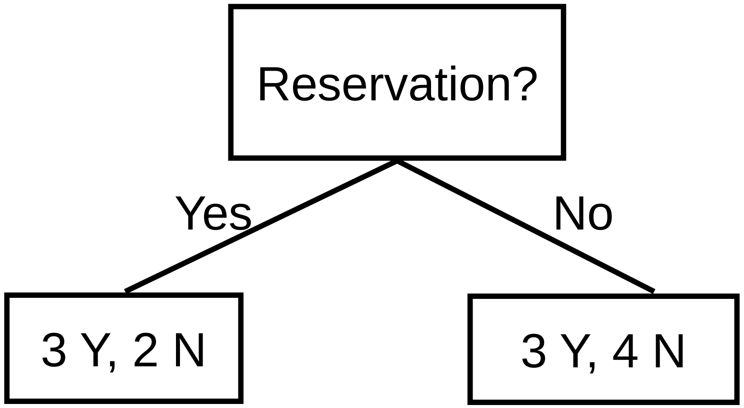
- Calculate Average entropy of attribute Raining?

$$AE_{Raining?} = \frac{5}{12}\left[-\left(\frac{3}{5}\log_2\frac{3}{5}\right) - \left(\frac{2}{5}\log_2\frac{2}{5}\right)\right] + \frac{7}{12}\left[-\left(\frac{3}{7}\log_2\frac{3}{7}\right) - \left(\frac{4}{7}\log_2\frac{4}{7}\right)\right] = 0.979$$

- Calculate Information gain of attribute Raining?

$$IG(Raining?) = H(S) - AE_{Raining?} = 1 - 0.979 = 0.021$$

Example problem: Restaurant waiting



Example	Input Attributes										Output
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
$\mathbf{x}_1$	Yes	No	No	Yes	Some	$$$	No	Yes	French	0–10	$y_1 = Yes$
$\mathbf{x}_2$	Yes	No	No	Yes	Full	$	No	No	Thai	30–60	$y_2 = No$
$\mathbf{x}_3$	No	Yes	No	No	Some	$	No	No	Burger	0–10	$y_3 = Yes$
$\mathbf{x}_4$	Yes	No	Yes	Yes	Full	$	Yes	No	Thai	10–30	$y_4 = Yes$
$\mathbf{x}_5$	Yes	No	Yes	No	Full	$$$	No	Yes	French	>60	$y_5 = No$
$\mathbf{x}_6$	No	Yes	No	Yes	Some	$$	Yes	Yes	Italian	0–10	$y_6 = Yes$
$\mathbf{x}_7$	No	Yes	No	No	None	$	Yes	No	Burger	0–10	$y_7 = No$
$\mathbf{x}_8$	No	No	No	Yes	Some	$$	Yes	Yes	Thai	0–10	$y_8 = Yes$
$\mathbf{x}_9$	No	Yes	Yes	No	Full	$	Yes	No	Burger	>60	$y_9 = No$
$\mathbf{x}_{10}$	Yes	Yes	Yes	Yes	Full	$$$	No	Yes	Italian	10–30	$y_{10} = No$
$\mathbf{x}_{11}$	No	No	No	No	None	$	No	No	Thai	0–10	$y_{11} = No$
$\mathbf{x}_{12}$	Yes	Yes	Yes	Yes	Full	$	No	No	Burger	30–60	$y_{12} = Yes$

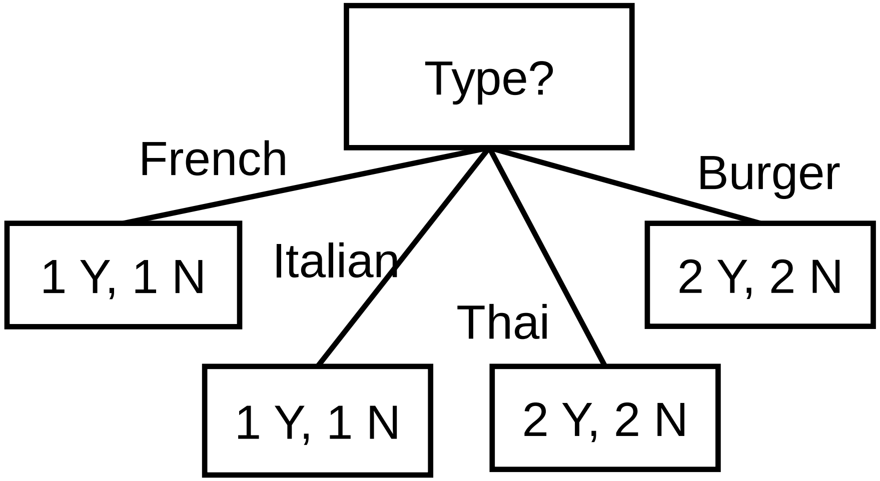
- Calculate Average entropy of attribute Reservation?

$$AE_{Reservation?} = \frac{5}{12}\left[-\left(\frac{3}{5}\log_2\frac{3}{5}\right)-\left(\frac{2}{5}\log_2\frac{2}{5}\right)\right]+\frac{7}{12}\left[-\left(\frac{3}{7}\log_2\frac{3}{7}\right)-\left(\frac{4}{7}\log_2\frac{4}{7}\right)\right] = 0.979$$

- Calculate Information gain of attribute Reservation?

$$IG(Reservation?) = H(S) - AE_{Reservation?} = 1 - 0.979 = 0.021$$

Example problem: Restaurant waiting



Example	Input Attributes										Output
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
$\mathbf{x}_1$	*Yes*	*No*	*No*	*Yes*	*Some*	*$$$*	*No*	*Yes*	*French*	*0–10*	$y_1 = Yes$
$\mathbf{x}_2$	*Yes*	*No*	*No*	*Yes*	*Full*	*$*	*No*	*No*	*Thai*	*30–60*	$y_2 = No$
$\mathbf{x}_3$	*No*	*Yes*	*No*	*No*	*Some*	*$*	*No*	*No*	*Burger*	*0–10*	$y_3 = Yes$
$\mathbf{x}_4$	*Yes*	*No*	*Yes*	*Yes*	*Full*	*$*	*Yes*	*No*	*Thai*	*10–30*	$y_4 = Yes$
$\mathbf{x}_5$	*Yes*	*No*	*Yes*	*No*	*Full*	*$$$*	*No*	*Yes*	*French*	*>60*	$y_5 = No$
$\mathbf{x}_6$	*No*	*Yes*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Italian*	*0–10*	$y_6 = Yes$
$\mathbf{x}_7$	*No*	*Yes*	*No*	*No*	*None*	*$*	*Yes*	*No*	*Burger*	*0–10*	$y_7 = No$
$\mathbf{x}_8$	*No*	*No*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Thai*	*0–10*	$y_8 = Yes$
$\mathbf{x}_9$	*No*	*Yes*	*Yes*	*No*	*Full*	*$*	*Yes*	*No*	*Burger*	*>60*	$y_9 = No$
$\mathbf{x}_{10}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$$$*	*No*	*Yes*	*Italian*	*10–30*	$y_{10} = No$
$\mathbf{x}_{11}$	*No*	*No*	*No*	*No*	*None*	*$*	*No*	*No*	*Thai*	*0–10*	$y_{11} = No$
$\mathbf{x}_{12}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$*	*No*	*No*	*Burger*	*30–60*	$y_{12} = Yes$

- Calculate Average entropy of attribute Type?

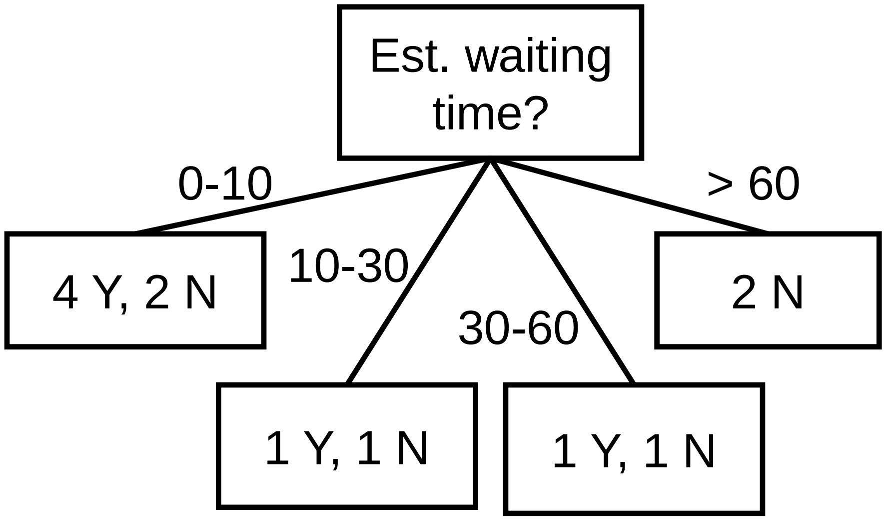
$$AE_{Type?} = \frac{2}{12}\left[-\left(\frac{1}{2}\log_2\frac{1}{2}\right)-\left(\frac{1}{2}\log_2\frac{1}{2}\right)\right]+\frac{2}{12}\left[-\left(\frac{1}{2}\log_2\frac{1}{2}\right)-\left(\frac{1}{2}\log_2\frac{1}{2}\right)\right]$$

$$+\frac{4}{12}\left[-\left(\frac{2}{4}\log_2\frac{2}{4}\right)-\left(\frac{2}{4}\log_2\frac{2}{4}\right)\right]+\frac{4}{12}\left[-\left(\frac{2}{4}\log_2\frac{2}{4}\right)-\left(\frac{2}{4}\log_2\frac{2}{4}\right)\right]=1$$

- Calculate Information gain of attribute Type?

$$IG(Type?) = H(S) - AE_{Type?} = 1 - 1 = 0$$

Example problem: Restaurant waiting



Example	Input Attributes										Output
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
$\mathbf{x}_1$	Yes	No	No	Yes	Some	$$$	No	Yes	French	0–10	$y_1 = Yes$
$\mathbf{x}_2$	Yes	No	No	Yes	Full	$	No	No	Thai	30–60	$y_2 = No$
$\mathbf{x}_3$	No	Yes	No	No	Some	$	No	No	Burger	0–10	$y_3 = Yes$
$\mathbf{x}_4$	Yes	No	Yes	Yes	Full	$	Yes	No	Thai	10–30	$y_4 = Yes$
$\mathbf{x}_5$	Yes	No	Yes	No	Full	$$$	No	Yes	French	>60	$y_5 = No$
$\mathbf{x}_6$	No	Yes	No	Yes	Some	$$	Yes	Yes	Italian	0–10	$y_6 = Yes$
$\mathbf{x}_7$	No	Yes	No	No	None	$	Yes	No	Burger	0–10	$y_7 = No$
$\mathbf{x}_8$	No	No	No	Yes	Some	$$	Yes	Yes	Thai	0–10	$y_8 = Yes$
$\mathbf{x}_9$	No	Yes	Yes	No	Full	$	Yes	No	Burger	>60	$y_9 = No$
$\mathbf{x}_{10}$	Yes	Yes	Yes	Yes	Full	$$$	No	Yes	Italian	10–30	$y_{10} = No$
$\mathbf{x}_{11}$	No	No	No	No	None	$	No	No	Thai	0–10	$y_{11} = No$
$\mathbf{x}_{12}$	Yes	Yes	Yes	Yes	Full	$	No	No	Burger	30–60	$y_{12} = Yes$

- Calculate Average entropy of attribute Est. waiting time?

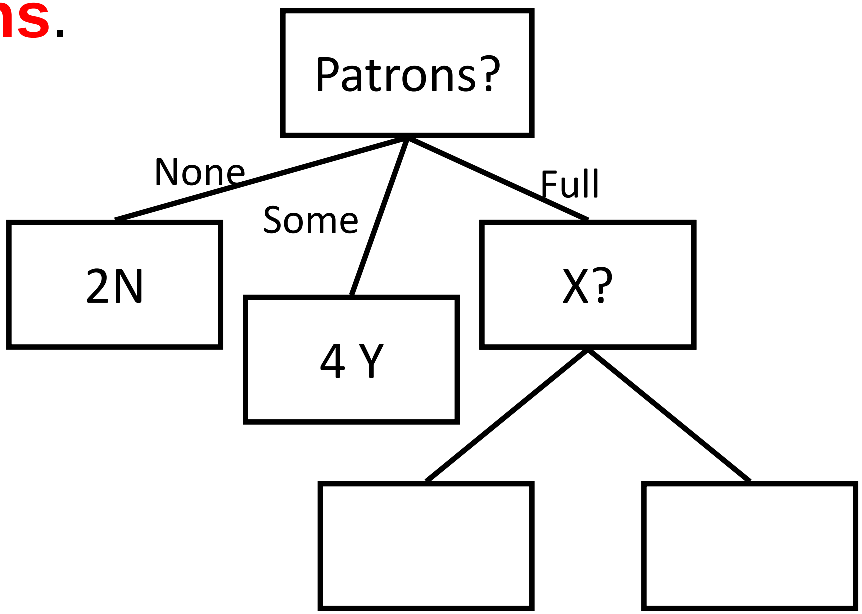
$$AE_{Est.waiting\ time?} = \frac{6}{12}\left[-\left(\frac{4}{6}\log_2\frac{4}{6}\right)-\left(\frac{2}{6}\log_2\frac{2}{6}\right)\right]+\frac{2}{12}\left[-\left(\frac{1}{2}\log_2\frac{1}{2}\right)-\left(\frac{1}{2}\log_2\frac{1}{2}\right)\right]$$
$$+\frac{2}{12}\left[-\left(\frac{1}{2}\log_2\frac{1}{2}\right)-\left(\frac{1}{2}\log_2\frac{1}{2}\right)\right]+\frac{2}{12}\left[-\left(\frac{0}{2}\log_2\frac{0}{2}\right)-\left(\frac{2}{2}\log_2\frac{2}{2}\right)\right] = 0.792$$

- Calculate Information gain of attribute Est. waiting time?

$$IG(Est.waiting\ time?\,, S) = H(S) - AE_{Est.waiting\ time?} = 1 - 0.792 = 0.208$$

Example problem: Restaurant waiting

- Largest Information Gain (0.459) / Smallest Entropy (0.541) achieved by splitting on **Patrons**.



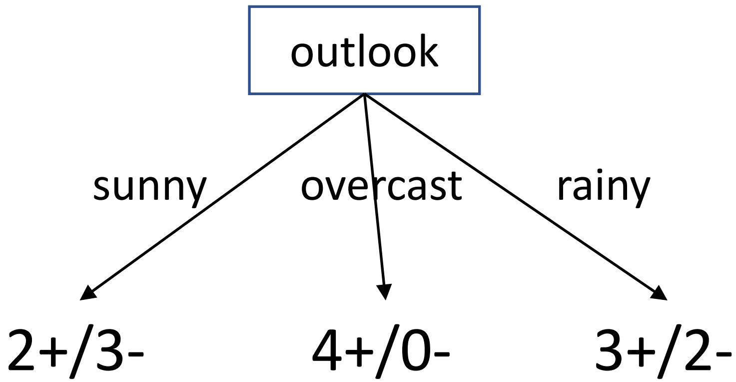
- Continue making new splits, always purifying nodes

Another numerical example

Example data set: Weather data

outlook	temperature	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Numerical example: Choose the root

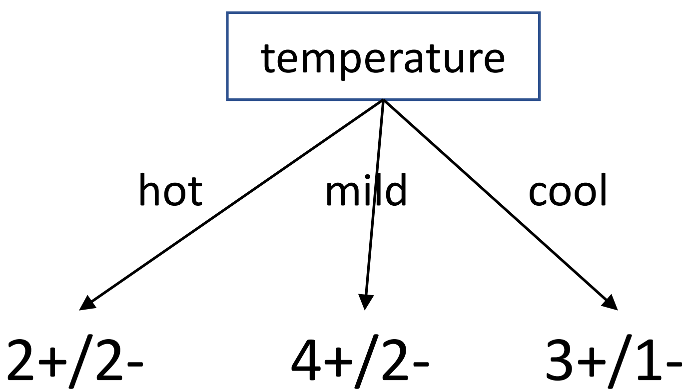


H_{sunny} = $-\ 2/5 \cdot \log_2 2/5 - 3/5 \cdot \log_2 3/5 = 0.971$

$H_{overcast}$ = $-\ 4/4 \cdot \log_2 4/4 - 0/4 \cdot \log_2 0/4 = 0$

H_{rainy} = $-\ 3/5 \cdot \log_2 3/5 - 2/5 \cdot \log_2 2/5 = 0.971$

$AE = 5/14 \cdot 0.971 + 4/14 \cdot 0 + 5/14 \cdot 0.971 = 0.694$



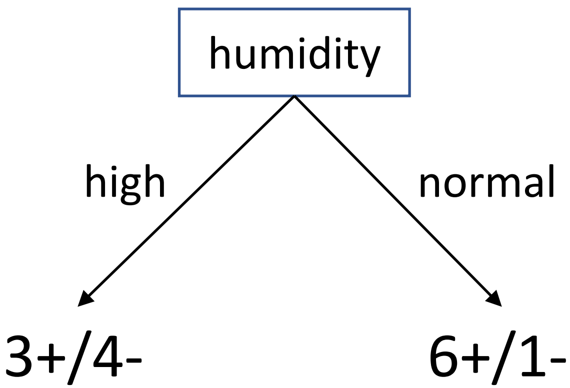
H_{hot} = $-\ 2/4 \cdot \log_2 2/4 - 2/4 \cdot \log_2 2/4 = 1$

H_{mild} = $-\ 4/6 \cdot \log_2 4/6 - 2/6 \cdot \log_2 2/6 = 0.918$

H_{cool} = $-\ 2/4 \cdot \log_2 2/4 - 3/4 \cdot \log_2 3/4 = 0.811$

$AE = 4/14 \cdot 1 + 6/14 \cdot 0.918 + 4/14 \cdot 0.811 = 0.911$

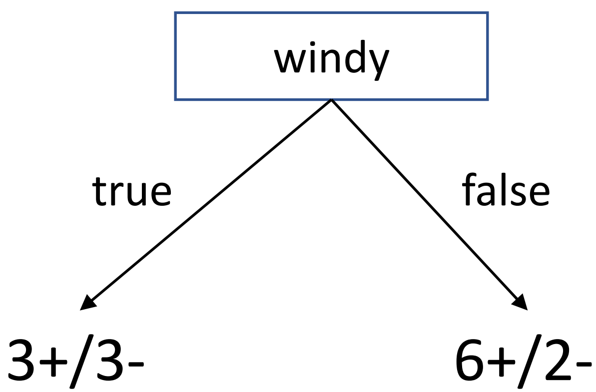
Numerical example: Choose the root



$H_{high} = -3/7 \cdot \log_2 3/7 - 4/7 \cdot \log_2 4/7 = 0.985$

$H_{normal} = -6/7 \cdot \log_2 6/7 - 1/7 \cdot \log_2 1/7 = 0.592$

$AE = 7/14 \cdot 0.985 + 7/14 \cdot 0.592 = 0.789$

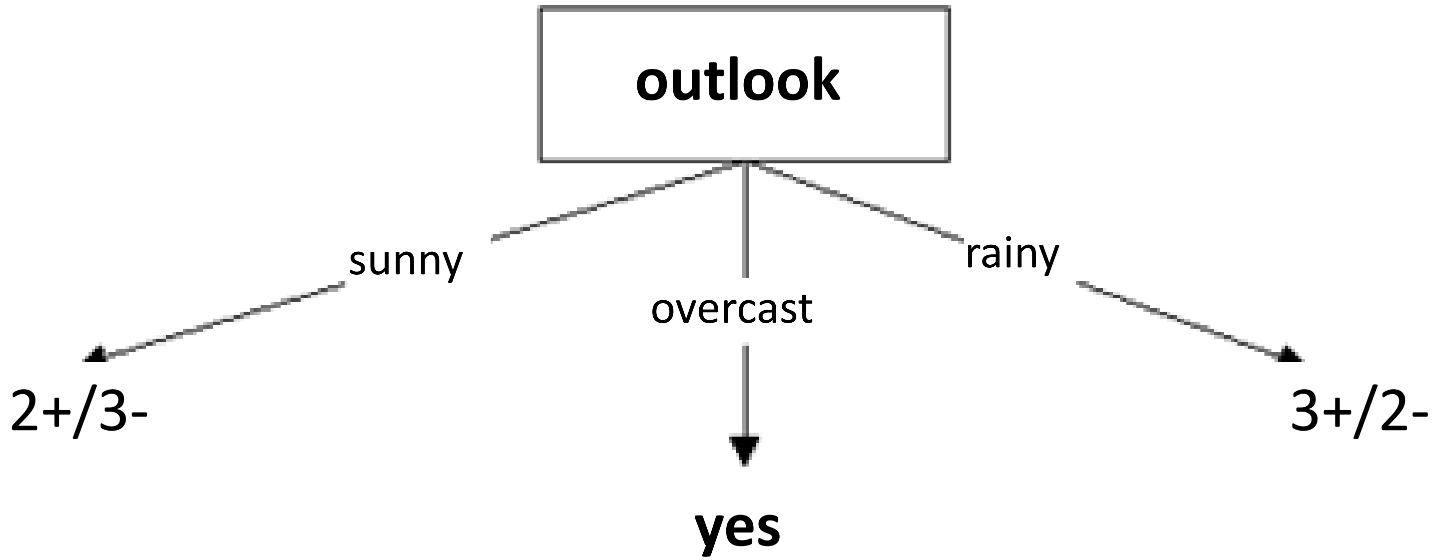


$H_{true} = -3/6 \cdot \log_2 3/6 - 3/6 \cdot \log_2 3/6 = 1$

$H_{false} = -6/8 \cdot \log_2 6/8 - 2/8 \cdot \log_2 2/8 = 0.811$

$AE = 6/14 \cdot 1 + 8/14 \cdot 0.811 = 0.892$

Numerical example: The partial tree

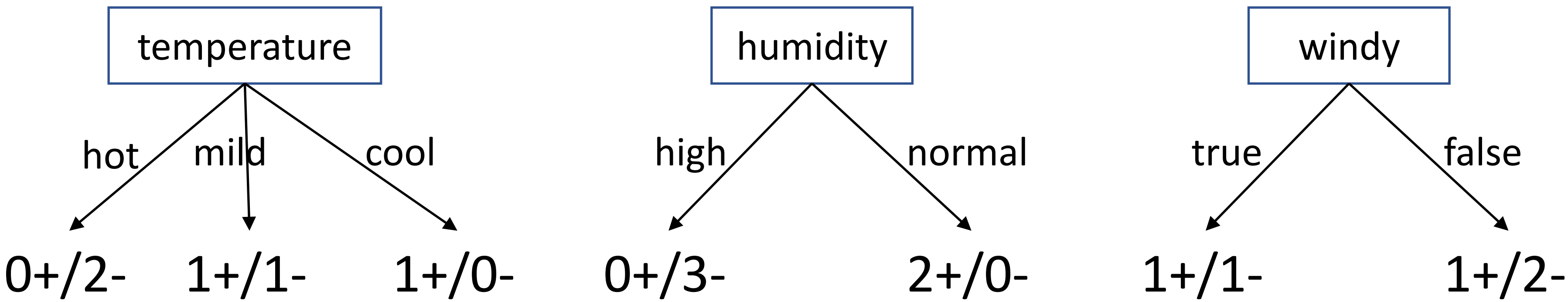


- Which attributes are chosen for the next splits?
- Continue splitting…

Numerical example: The second level

- Choose an attribute for the branch outlook = sunny.

outlook	**temperature**	**humidity**	**windy**	**play**
sunny	hot	high	false	no
sunny	hot	high	true	no
sunny	mild	high	false	no
sunny	cool	normal	false	yes
sunny	mild	normal	true	yes



$H_{hot} = 0$, $H_{mild} = 1$, $H_{cool} = 0$

AE = 0.4

$H_{high} = 0$, $H_{normal} = 0$

AE = 0

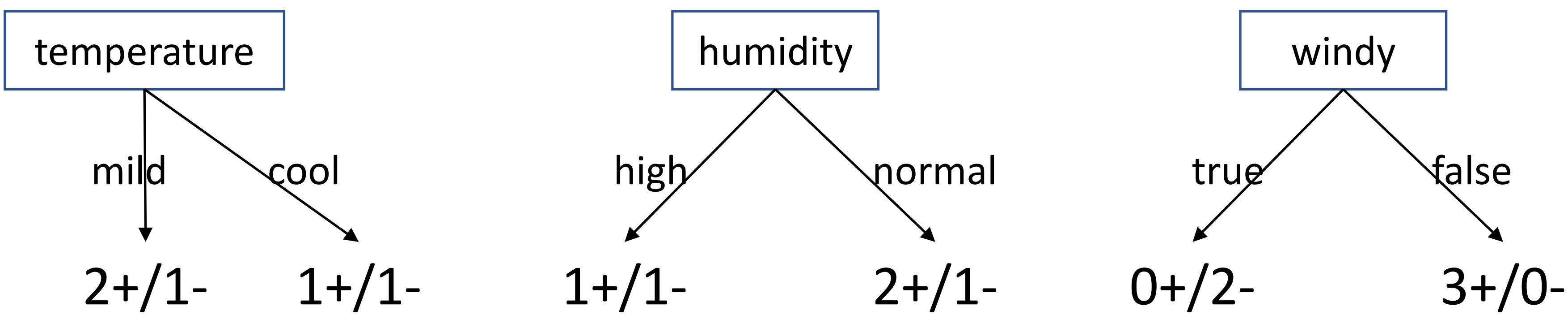
$H_{TRUE} = 1$, $H_{FALSE} = 0.918$

AE = $3/3 \cdot 0.918 = 0.951$

Numerical example: The second level

- Choose an attribute for the branch outlook = rainy

outlook	temperature	humidity	windy	play
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
rainy	mild	normal	false	yes
rainy	mild	high	true	no



H_{mild} = 0.918, H_{cool} = 1

AE = 0.951

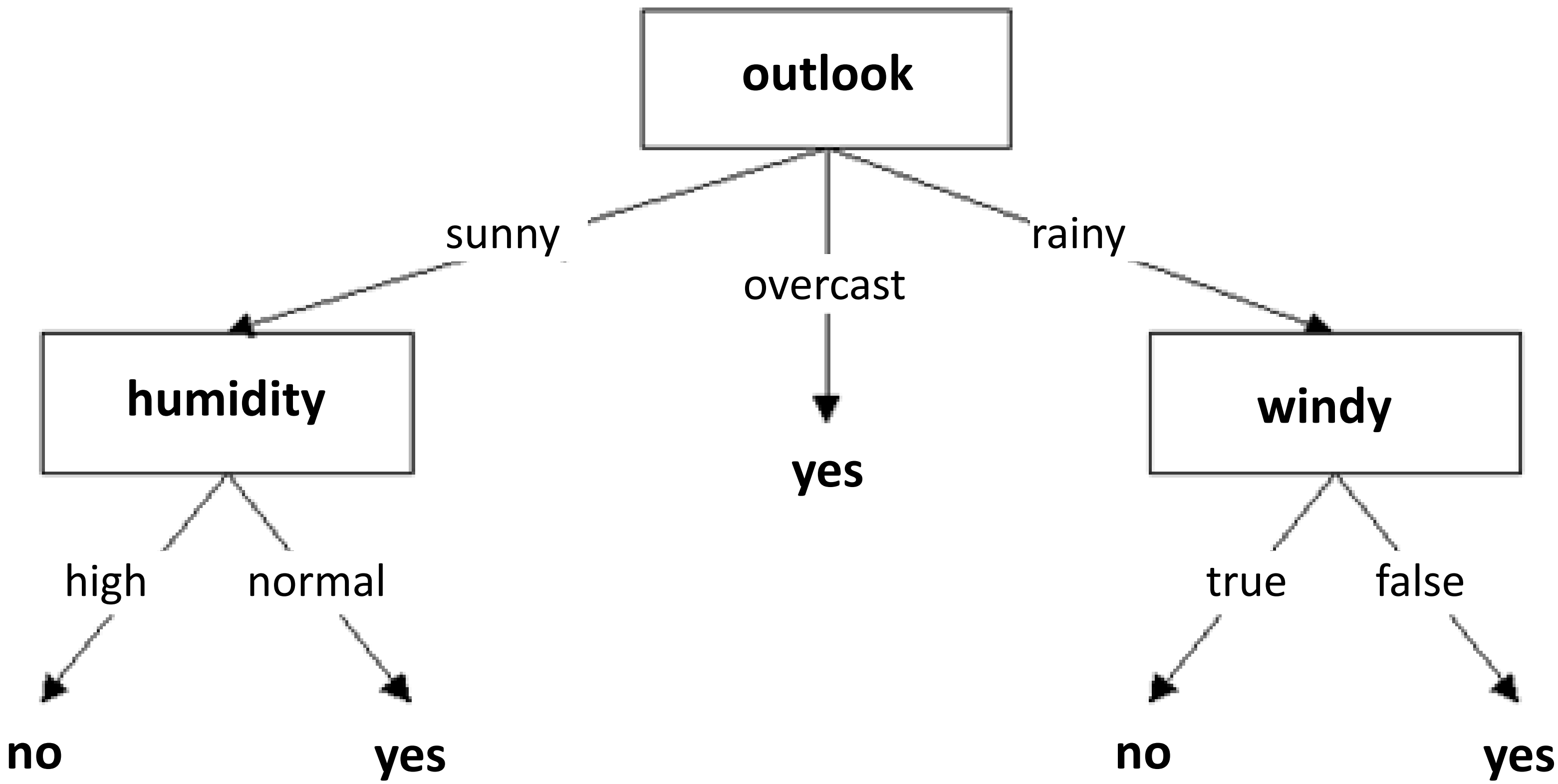
H_{high} = 1, H_{normal} = 0.918

AE = 0.951

H_{TRUE} = 0, H_{FALSE} = 0

AE = 0

Numberical example: The final tree



Quiz 01: ID3 decision tree

- The data represent files on a computer system. Possible values of the class variable are “infected”, which implies the file has a virus infection, or “clean” if it doesn't.
- Derive decision tree for virus identification.

No.	Writable	Updated	Size	Class
1	Yes	No	Small	Infected
2	Yes	Yes	Large	Infected
3	No	Yes	Med	Infected
4	No	No	Med	Clean
5	Yes	No	Large	Clean
6	No	No	Large	Clean


...the end.