Decision Trees

Dr. Mohamed Elshenawy
mmelshenawy@gmail.com

1

In Previous Lectures

- Linear Regression
- Polynomial Regression
- Training and test error
- Bias and Variance Tradeoff
- Classification
- KNN
- Logistic Regression
- SVM

2

In this session

- What is a decision tree?
- Expressiveness of decision trees (functions that can be represented by decision trees)
- Learning decision trees
- Decision Tree pruning

3

References

For further readings, check:

- Section 18.3 from the book: Artificial Intelligence – a modern approach (Third Edition) by Stuart Russell and Peter Norvig
- Section 8.1 from the book : An Introduction to Statistical Learning. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, ISBN: 978-1-461-47137-0

4

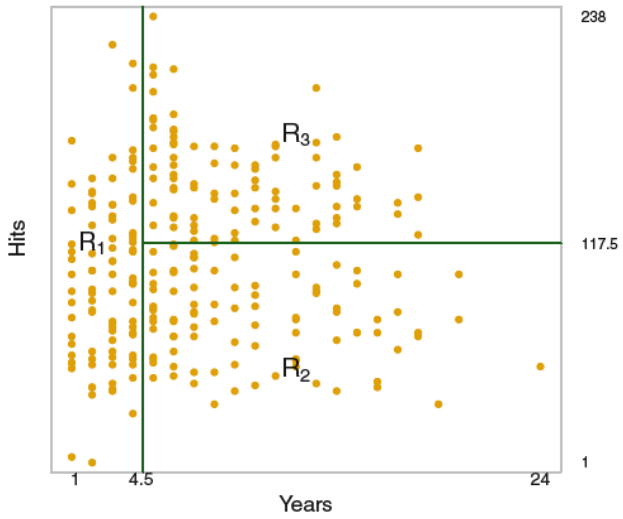
Decision Trees

- Non-parametric model
- One of the simplest and yet most successful forms of machine learning.
- Can be applied to both regression and classification problems.

5

Example

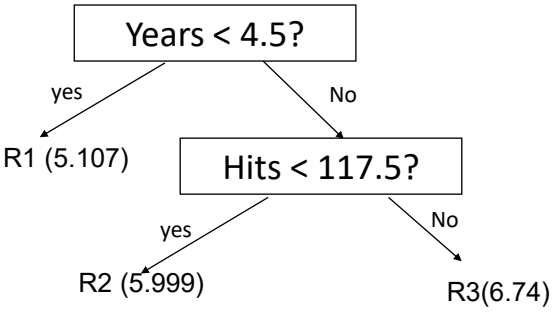
- A regression tree for predicting the log salary of a baseball player, based on the number of years that he has played in the major leagues and the number of hits that he made in the previous year.



6

We can predict the salary using a tree

- We can construct a tree as shown
- If (Years<4.5) then log salary =5.107
  - The predicted salaries for this group $1{,}000 \times e^{5.107} = \$165{,}174$
- If (Years>=4.5 and Hits<117.5} then log salary =5.999
  - The predicted salaries for this group $1{,}000 \times e^{5.999} = \$402{,}834$
- If (Years>=4.5 and Hits>=117.5) then log salary =6.74
  - The predicted salaries for this group $1{,}000 \times e^{6.740} = \$845{,}346$



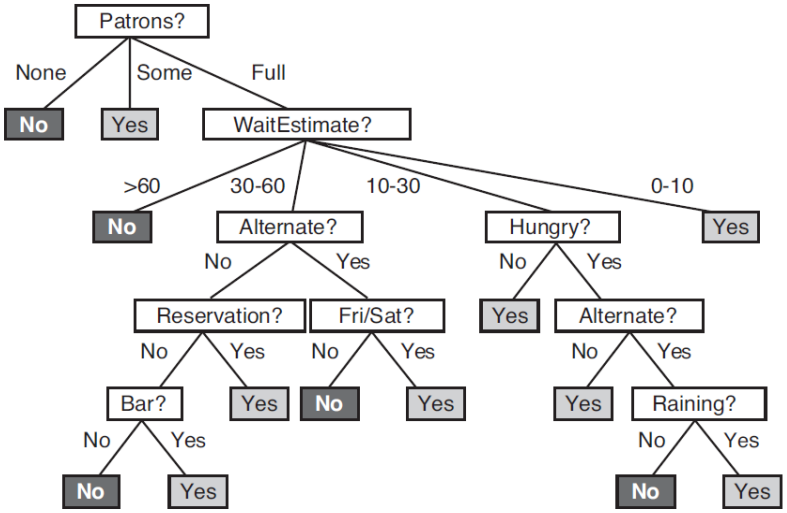
7

Another example (Classification)

- Goal: decide whether to wait for a table at a restaurant.
- Attributes
  1. Alternate: whether there is a suitable alternative restaurant nearby.
  2. Bar: whether the restaurant has a comfortable bar area to wait in.
  3. Fri/Sat: true on Fridays and Saturdays.
  4. Hungry: whether we are hungry.
  5. Patrons: how many people are in the restaurant (values are None, Some, and Full).
  6. Price: the restaurant's price range ($, $$, $$$).
  7. Raining: whether it is raining outside.
  8. Reservation: whether we made a reservation.
  9. Type: the kind of restaurant (French, Italian, Thai, or burger).
  10. WaitEstimate: the wait estimated by the host (0–10 minutes, 10–30, 30–60, or >60).

8

Possible Solution



9

What is a decision tree?

- A decision tree represents a function that takes as input a vector of attribute values and returns a "decision"—a single output value
- A decision tree reaches its decision by performing a sequence of tests.
  - Each internal node in the tree corresponds to a test of the value of one of the input attributes A_i,
  - the branches from the node are labeled with the possible values of the attribute $f(A_i) = v_{ik}$
  - Each leaf node in the tree specifies a value to be returned by the decision tree
- Natural to humans- think of "how to" manuals, troubleshooting questions, etc

10

How can we construct such trees?

11

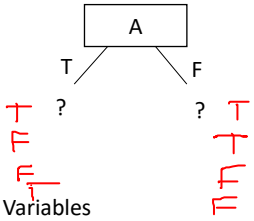
How many functions a tree can represent?

- Consider p Boolean attributes
- How many possible leaf nodes in that set?
  - $(2 \times 2 \ldots p\ times\ ) \ i.e. 2^p$, p: the number of attributes
- Using these leaf nodes, how many different models (functions) can we represent for each combination of input values?

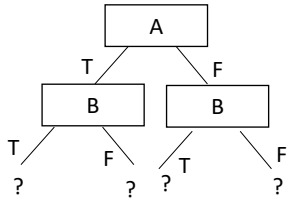
12

Possible functions (tree models)

One Variable



Two Variables



Possible functions

A	f_1	f_2	f_3	f_4
T	T	F	T	F
F	F	T	T	F

4 Possible functions

Possible functions

16 Possible functions: $2^{2^2} = 2^4 = 16$

13

Expressiveness of decision trees

- How many possible functions (trees) can the model fit?
  - We have more than 2^{2^n} possible trees.
- For 10 attribute(such as the previous example), we can get 2^{1024}or about 10^{308} models
- How can the model learn the right tree using the training data?

14

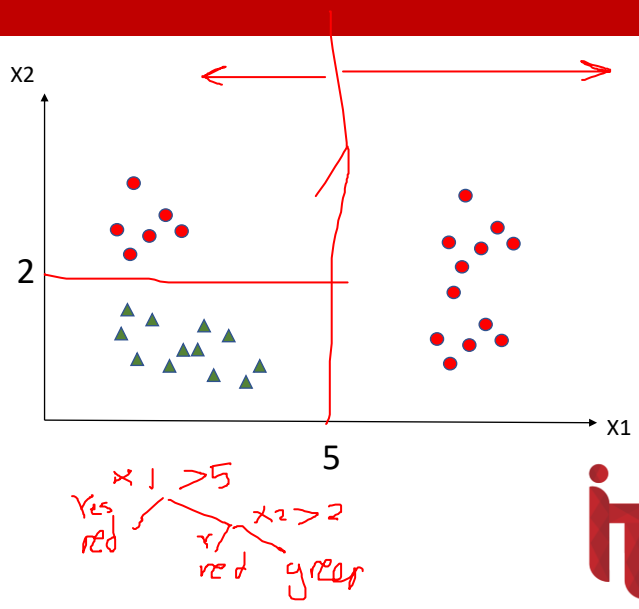
How can we choose a tree that produces the correct output?

- We use the training data
- Best solution: a tree that is consistent with the training examples and is as small as possible.
- This is an intractable problem. There is no efficient way to search through the 2^{2^d} trees
- We need to find near best (suboptimal) solution

15

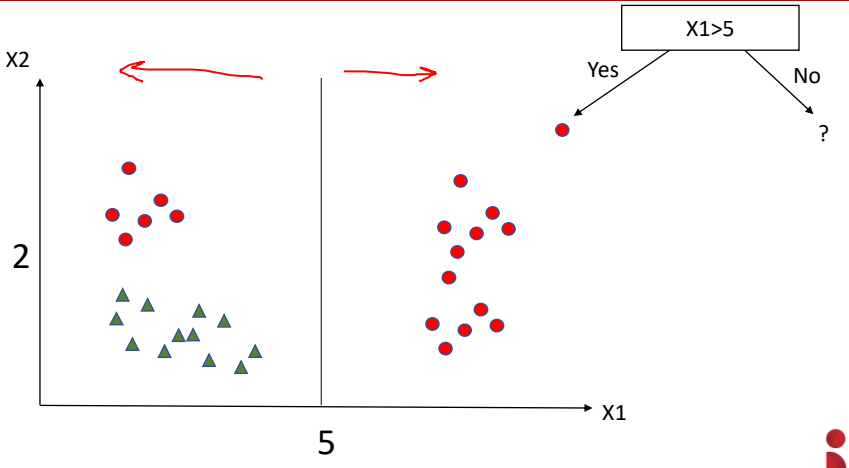
Example

- How can we split these classes?



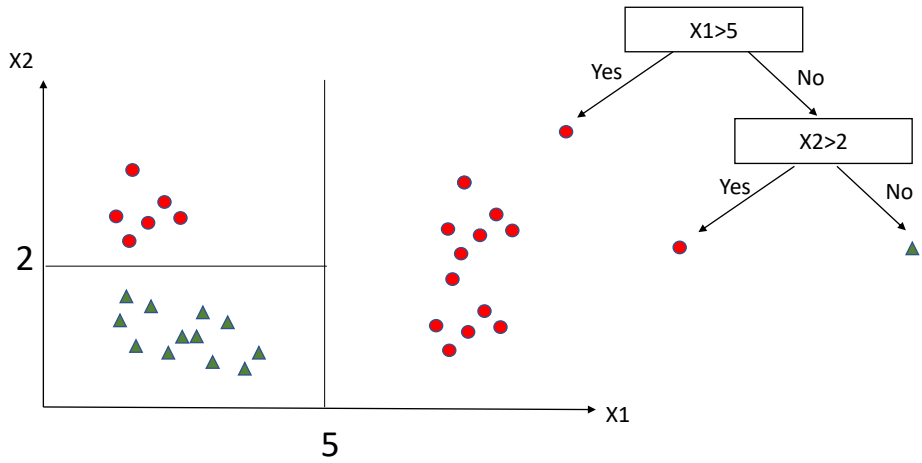
16

First Split



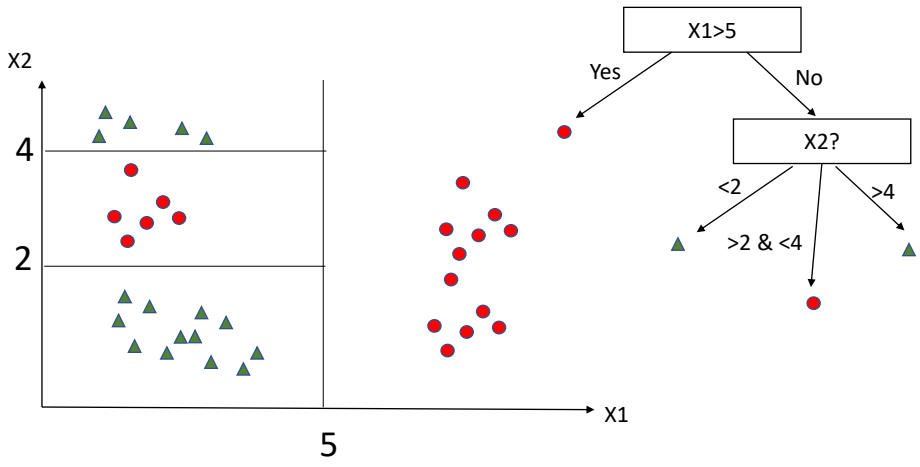
17

Second Split



18

How about



19

Recursive Splitting

- We use a top-down greedy approach (also known as recursive splitting)- a good approximation that result in a small (not the smallest) tree
- Starting with an empty tree:
  - begins at the top of the tree (at which point all observations belong to a single region)
  - Test the most important attribute first and use it to split the predictor space into sub-regions (best split).
  - Use recursion till you reach the leaf nodes
- It is *greedy* because at each step of the tree-building process, the *best* split is made at that particular step
- How can we choose the most important attribute?

20

The most important attribute

- Which attribute is better to split on?
- The best attribute will offer more information about the output (reduce the uncertainty) better than any other attribute
- How can we measure uncertainty (classification problems)?

21

Review: Information Theory

- We learn from the information theory that:
  - learning that an unlikely event has occurred is more informative than learning that a likely event has occurred.
- Which of the following statements has more information
  - The sun rose this morning
  - There was a solar eclipse this morning
- We define the **self-information** of an event $X = x$

$$I(X = x) = -logp(X = x)$$

- -ve sign to make $I(X = x)$ a positive value

22

self-information

- If $p(X = x) = 1$ then $I(X = x) = 0$
- If $p(X = x) = 0.75$ then $I(X = x) = -(-0.415) = 0.415$
- If $p(X = x) = 0.5$ then $I(X = x) = -(-1) = 1$
- If $p(X = x) = 0.25$ then $I(X = x) = -(-2) = 2$
- If $p(X = x) = 0.05$ then $I(X = x) = -(-4.32) = 4.32$

23

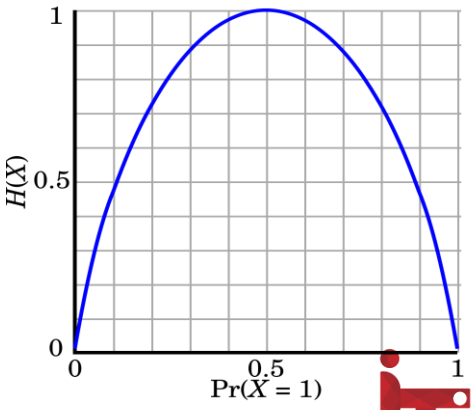
Entropy

- Self-information deals only with a single outcome. We can quantify the amount of uncertainty in an entire probability distribution using the Shannon entropy:

$$H(X) = \mathbb{E}_{X\sim P}[I(X = x)] = -\mathbb{E}_{X\sim P}[\log P(X = x)]$$

$$H(X) = -\sum_{x\in X} p(X = x) log\ p(X = x)$$

- It gives a lower bound on the number of bits (if the logarithm is base
- Distributions that are nearly deterministic (where the outcome is nearly certain) have low entropy.
- Distributions that are closer to uniform have high entropy.



24

Conditional Entropy

$$H(Y|X) = -\sum_{y \in Y, x \in X} p(Y = y, X = x) \log \frac{p(Y = y, X = x)}{p(X = x)}$$

$$H(Y|X) = 0 \ (Y \ is \ completely \ determined \ by \ the \ value \ of \ X)$$

$$H(Y|X) = H(Y) \ (Y \ and \ X \ are \ independent)$$

25

Information Gain

- The information gain of an attribute A_i is the reduction in entropy due to the use of attribute to split on

$$IG(Y|X) = H(Y) - H(Y|X)$$

- $H(Y)$: Uncertainty about the output
- $H(Y|X)$: Uncertainty about the output if we know the value of input X
- We can use information gain to determine the best attribute to split on.

26

Another metric

- Gini impurity
- Measure of node *purity*—a small value indicates that a node contains predominant observations from a single class
- We will not discuss it in this session

27

Algorithm (Recursive Splitting)

- Also known as ID3 (Iterative Dichotomiser 3)
- Key idea

1. Pick an attribute A_i to split at a non-terminal node
2. Use the attribute value to split examples into groups
3. For each group:
   - If all examples in same class then we are done {return class}
   - if No examples { return majority from parent as default}, no examples have been observed for this combination.
   - Else loop to step 1

28

Example

Example	Input Attributes										Goal
	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*	$y_{12} = Yes$

29

How to solve

- Calculate the conditional entropy for all attributes
- Choose the attribute with the maximum information Gain.
- Do the splitting

30

Example calculation

$$H(Y) = -\mathbb{E}_{Y\sim P}[\log P(Y=y)] = -\sum_{y\in Y} P(Y=y)\log P(Y=y)$$

For the output, Two possible Y values: 'Yes', 'No'

$$P(Y='Yes') = \frac{6}{12} = 0.5 \quad P(Y=' No') = \frac{6}{12} = 0.5$$

$$H(Y) = -0.5\ log_2 0.5 - 0.5\ log_2 0.5 = 0.5 + 0.5 = 1$$

31

Calculate the IG from an input 'patron'

- Three possible values: 'Some', 'Full', and 'None'

$$H(Y|X) = -\sum_{y\in Y, x\in X} p(Y=y, X=x)\log\frac{p(Y=y, X=x)}{p(X=x)}$$

$$P(Y=' Yes', X='some') = \frac{4}{12} \qquad P(Y=' No', X='some') = 0$$

$$P(Y=' Yes', X=' full') = \frac{2}{12} \qquad P(Y=' No', X=' full') = \frac{4}{12}$$

$$P(Y=' Yes', X=' None') = 0 \qquad P(Y=' No', X=' None') = \frac{2}{12}$$

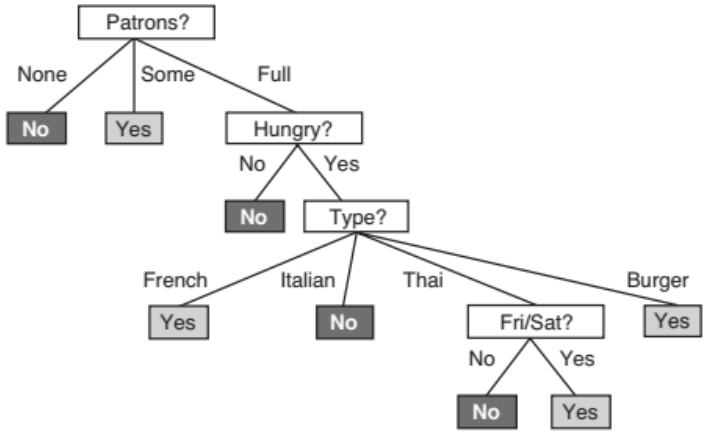
$$p(X=' some') = \frac{4}{12}, p(X=' full') = \frac{6}{12}, p(X=' none') = \frac{2}{12}$$

$$H(Y|X) = -\frac{4}{12} * \log\frac{4}{4} - \frac{2}{12} * log\frac{2}{6} - \frac{4}{12} * \log\frac{4}{6} - \frac{2}{12} * \log\frac{2}{2} = 0 + 0.264 + 0.195 + 0 = 0.459$$

$$IG(Patrons) = 1 - 0.459 = 0.541$$

32

Solution



33

Continuous Attribute

- How do we generate braches from a continuous attribute?
  - You need a split point (e.g. height>160)
  - Choose a point that gives the highest information gain.

34

Decision Tree pruning

- As the size of the tree grows, it is more likely that the model will overfit the data.
- A smaller tree with fewer splits (that is, fewer regions) might lead to lower variance and better interpretation at the cost of a little bias.
- How to solve the overfitting in decision tress:

1. First approach: split the tree only if the decrease in training error exceeds certain threshold (early stopping).
   Problem: a seemingly worthless split early on in the tree might be followed by a very good split.
2. Second approach: grow a large tree T_0, and then prune it back in order to obtain a subtree. How do we determine the best prune way to prune the tree?

35

Decision Tree pruning

- Prune the tree such that the cost of the resultant subtree T $\subset T_0$

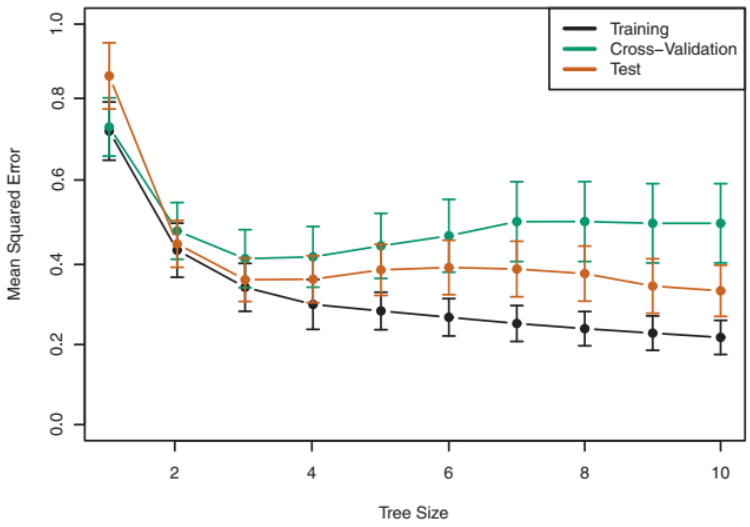
$Cost(T)$ *is small as possible*, where

$$Cost(T) = Error(T) + \alpha|T|$$

- $|T|$ indicates the number of terminal nodes of the tree T
- The tuning parameter α controls a trade-off between the subtree's complexity and its fit to the training data.
- When α = 0, then the subtree T will simply equal T_0
- as we increase α from zero, branches get pruned from the tree

36

Pruning



37

Advantages of Decision Trees

- Trees are very easy to explain to people. In fact, they are even easier to explain than linear regression.
- Some people believe that decision trees more closely mirror human decision-making than do the regression and classification approaches seen in previous chapters.
- Trees can be displayed graphically and are easily interpreted even by a non-expert (especially if they are small).

38

Disadvantages

- Trees generally do not have the same level of predictive accuracy compared to other models. This can be improved using methods ensemble learning techniques (e.g. random forests).
- As the branching goes deeper, you get exponentially less data.
- Overfitting: more likely as the hypothesis space and the number of input attributes grows (large trees), and less likely as we increase the number of training examples. (a way to solve overfitting is decision tree pruning)
- Greedy algorithms don't yield the global optimum tree structure.

39