LECTURER: TAI LE QUY

MACHINE LEARNING – SUPERVISED LEARNNG

UNIT 5.1

DECISION & REGRESSION TREES

STUDY GOALS

- explain the concept of decision and regression trees.
- define bagging and boosting.
- apply decision tree and regression tree models on your own with the use of Python.

EXPLAIN SIMPLY

1. How do tree-based algorithms generally work?
2. How do decision trees solve classification problems?
3. How do regression trees solve regression problems?

- One of the most popular classification methods
- DTs are included in many commercial systems nowadays
- Easy to interpret, human readable, intuitive
- Simple and fast methods.
- Many DT induction algorithms have been proposed
  - ID3 (Quinlan 1986)
  - C4.5 (Quinlan 1993)
  - CART (Breimanet al 1984)

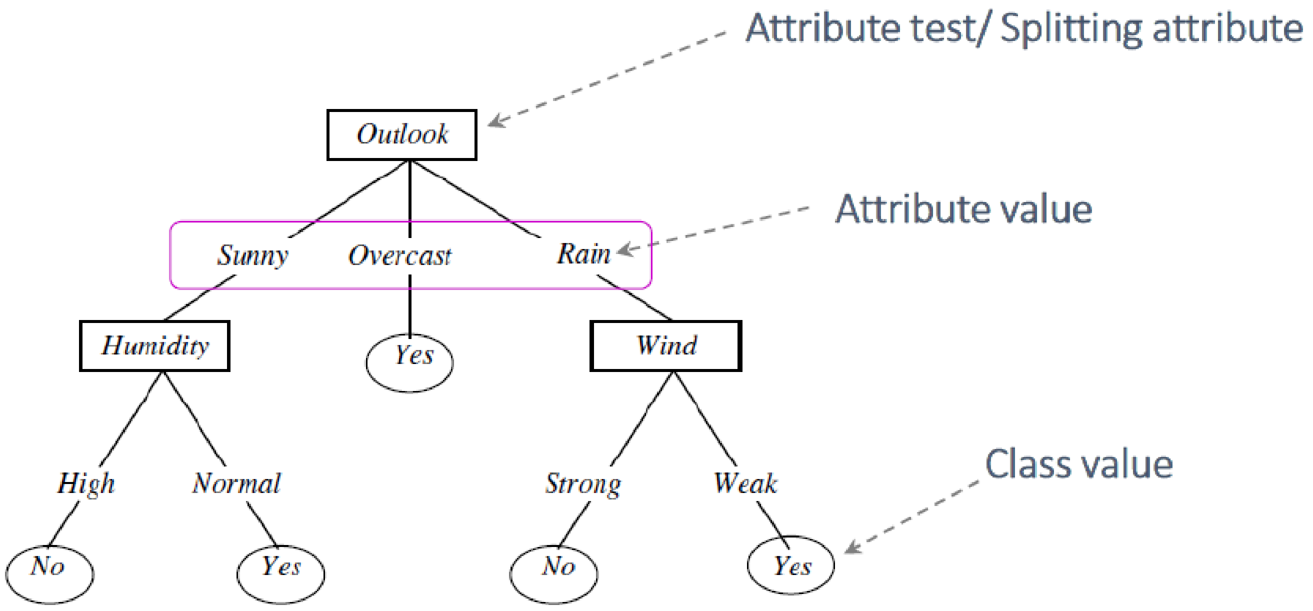
Source of image: Course book

– Representation

- Each internal node specifies a test of some predictive attribute
- Each branch descending from a node corresponds to one of the possible values for this attribute
- Each leaf node assigns a class label

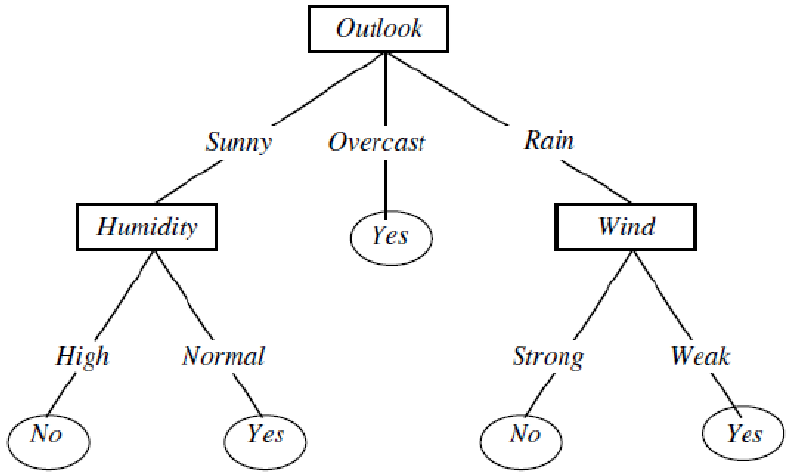
Training set

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



Source of the text: Ntoutsi, 2021

We can "translate" each path into IF-THEN rules (human readable)



IF ((Outlook = Sunny) ^ (Humidity = Normal)),
THEN (Play tennis=Yes)

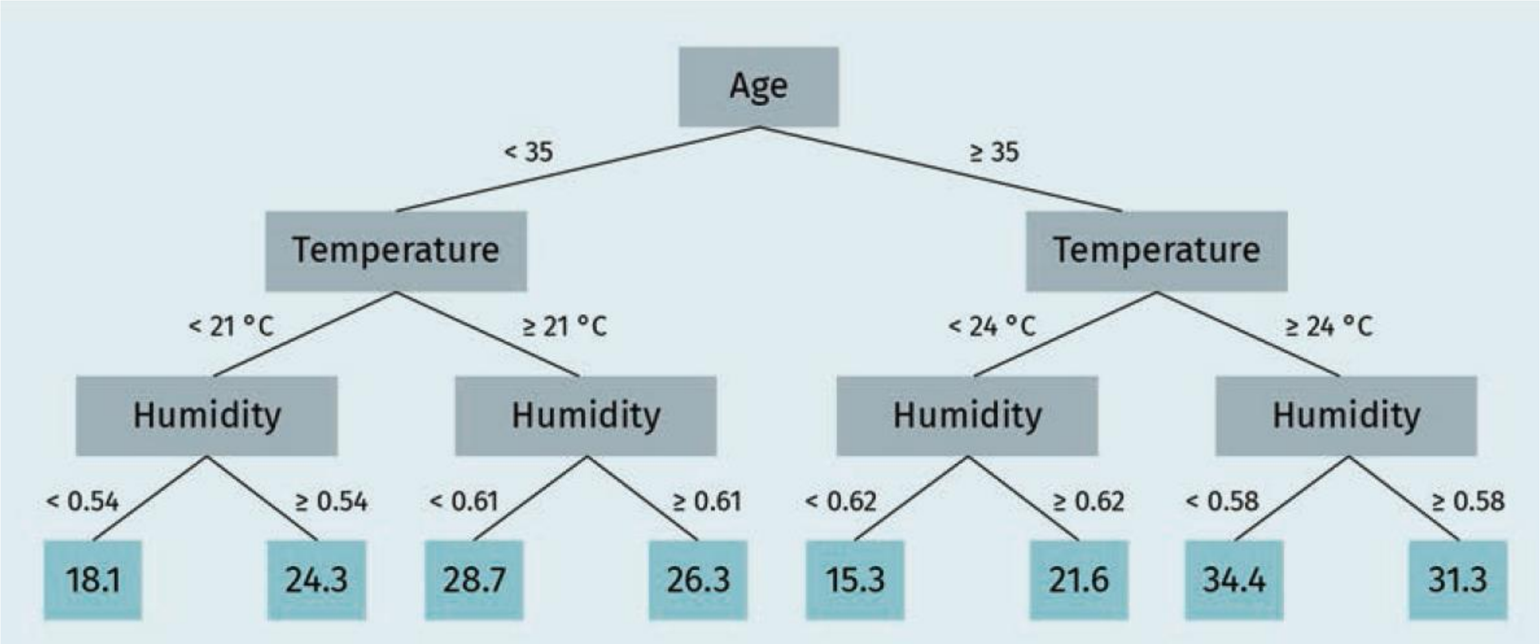
IF ((Outlook = Rain) ^ (Wind = Strong)),
THEN (Play tennis=No)

Should we play tennis?

X1=((Outlook=sunny) (Temperature=hot)(Humidity=high)(Wind=Weak))
X2=((Outlook=overcast) (Temperature=hot)(Humidity=high)(Wind=Weak))

Source of the text: Ntoutsi, 2021

- Tree-based structures can also be used on numerical features and building regression models
- Numerical features become more manageable through a discretization process, i.e., by assigning threshold values
  - These numerical features can be treated like categorical features



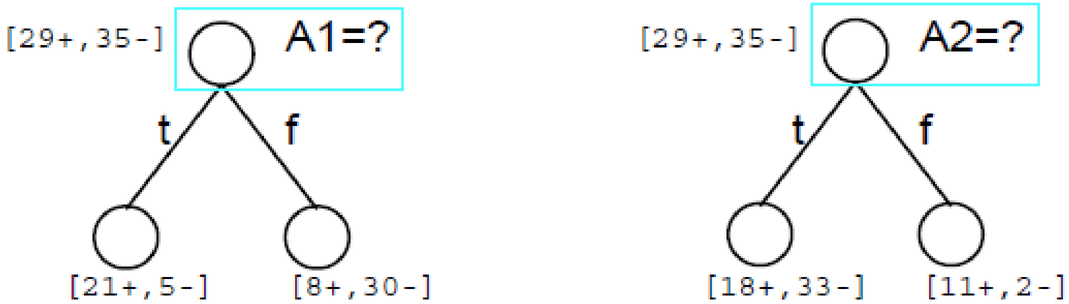
A regression tree predicting the duration of a walk

Source of image: Course book

DECISION TREE - BASIC METHOD

- The tree is constructed in a top-down recursive divide-and-conquer manner
- At start, all the training examples are at the root node
- The question is "Which attribute should be tested/ selected for split?"
  - Attributes are evaluated using some statistical measure, which determines how well each attribute alone classifies the training examples.
  - The best attribute is selected and used as the splitting attribute at the root.
- For each possible value of the splitting attribute, a descendant of the root node is created and the instances are mapped to the appropriate descendant node.
- The procedure is repeated for each descendant node, so instances are partitioned recursively.
- "When do we stop partitioning?"
  - All samples for a given node belong to the same class
  - There are no remaining attributes for further partitioning

- Which attribute to choose for splitting: A_1 or A_2?



- Different split attribute selection measures
  - Information gain
  - Gini impurity (Gini index)
  - Sum of squared errors (SSE), with regards to regression tree

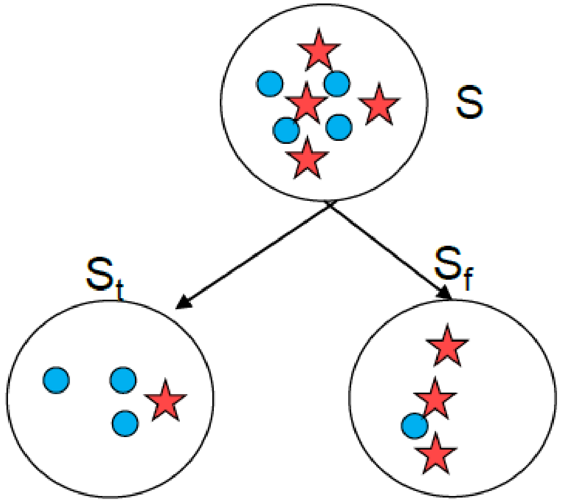
DECISION TREE – INFORMATION GAIN

- Used in ID3 (Quinlan, 1986)
- It uses entropy, a measure of pureness of the data
- The Information Gain *Gain(S, A)* of an attribute A relative to a collection of examples S measures the entropy reduction in S due to splitting on A:

$$G(S, A) = \boxed{\mathrm{Entropy}\Big(S\Big)} - \boxed{\sum\nolimits_{v\,\in\,\mathrm{values}(A)} \frac{|S_v|}{|S|}\mathrm{Entropy}\Big(S_v\Big)}$$

Before splitting

After splitting on A



- Information Gain measures the expected reduction in entropy due to splitting on A
- The attribute with the higher entropy reduction is chosen for splitting

Source of the text: Ntoutsi, 2021

ENTROPY FOR MEASURING IMPURITY OF A SET OF INSTANCES

- Entropy comes from information theory.
  - It represents the average amount of information needed to identify the class label of an instance in S
  - The higher the entropy the more the information content
- Let S be a collection of positive and negative examples
  - p+: the percentage of positive examples in S
  - p-: the percentage of negative examples in S
- Entropy measures the impurity of S:

$$Entropy(S) = -p_+ \log_2(p_+) - p_- \log_2(p_-)$$

in the general case
(k-classification problem)

$$Entropy(S) = \sum_{i=1}^{k} - p_i \log_2(p_i)$$

  - Entropy= 0, when all members belong to the same class
  - Entropy= 1, when there is an equal number of positive and negative examples

Source of the text: Ntoutsi, 2021

- What is the entropy in the following cases?
  - S: [9+,5-]
  - S: [7+,7-]
  - S: [14+,0-]

$$Entropy(S) = -p_+ \log_2(p_+) - p_- \log_2(p_-)$$

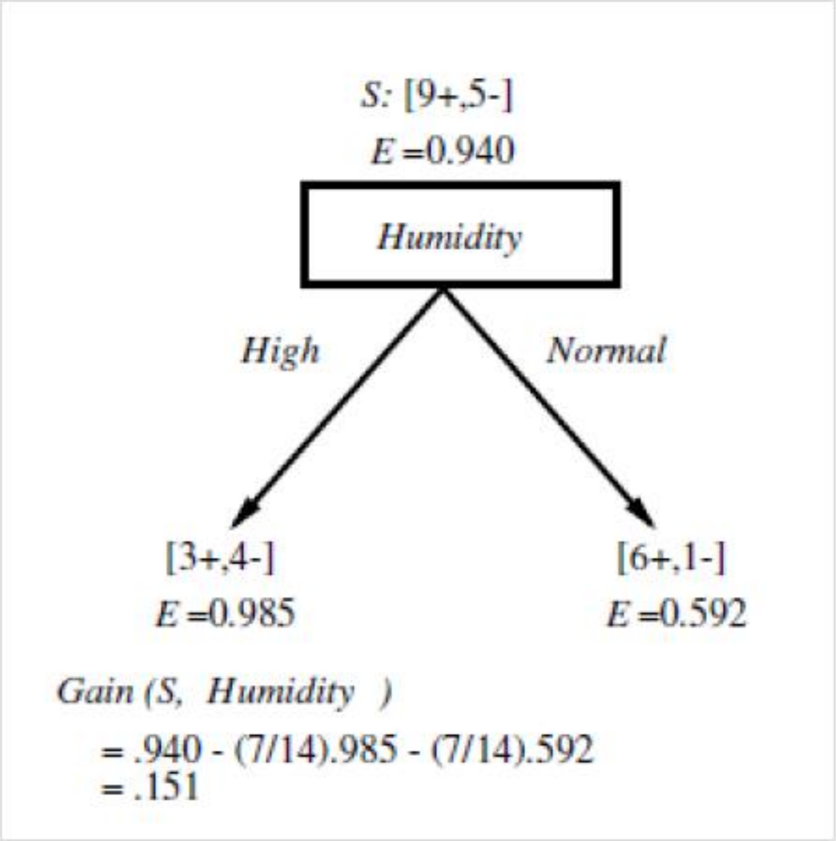
$$Entropy(S) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.940$$

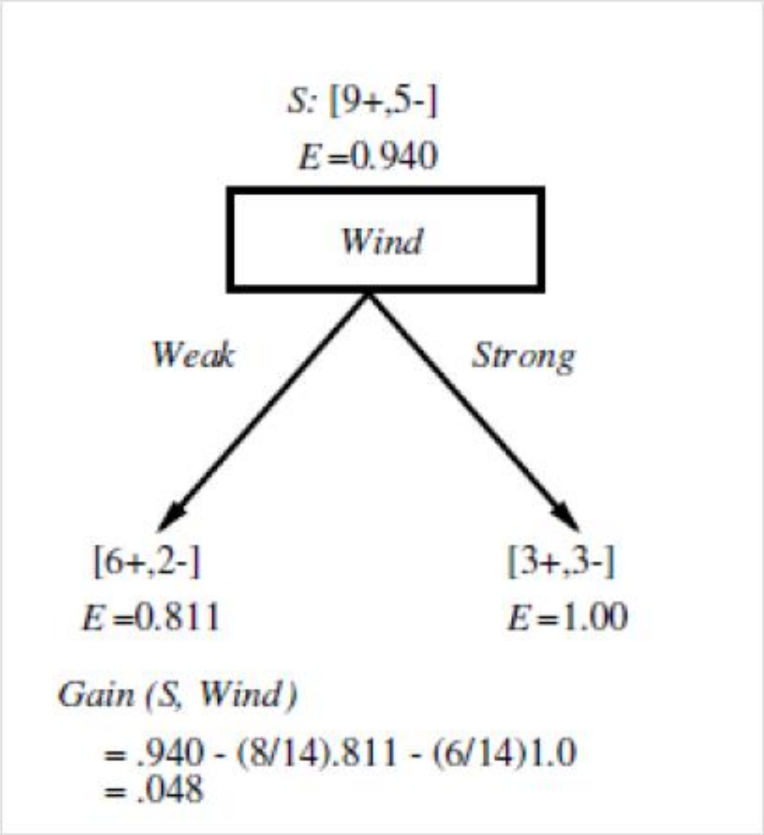
$$Entropy(S) = -\frac{7}{14}\log_2(\frac{7}{14}) - \frac{7}{14}\log_2(\frac{7}{14}) = 1$$

$$Entropy(S) = -\frac{14}{14}\log_2(\frac{14}{14}) - \frac{0}{14}\log_2(\frac{0}{14}) = 0$$

Source of the text: Ntoutsi, 2021

– Two options for splitting: “Humidity” and “Wind”?

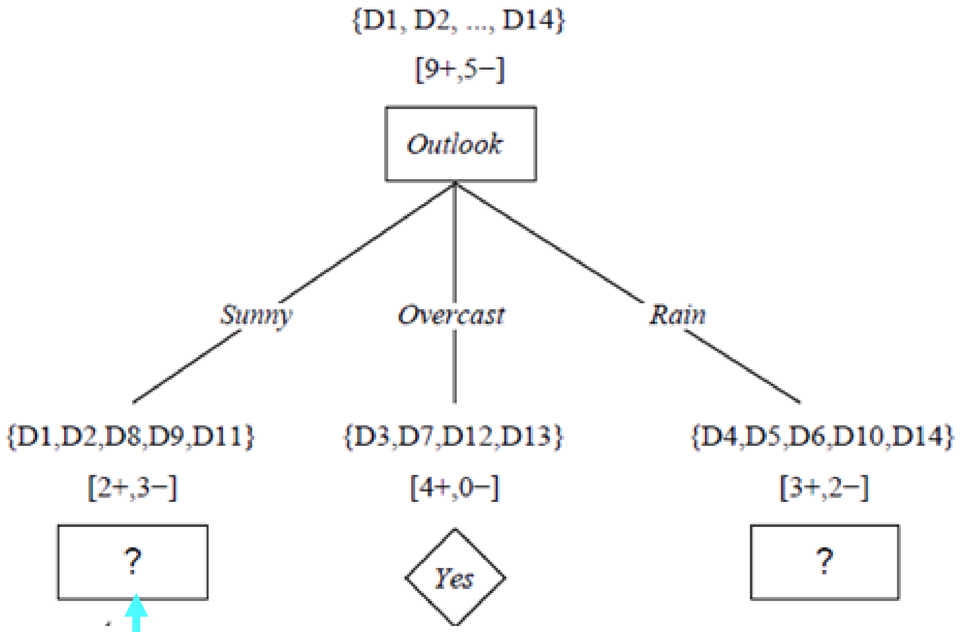




Source of the text: Ntoutsi, 2021

INFORMATION GAIN EXAMPLE

Repeat recursively



Which attribute should we choose for splitting here?

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$$S_{sunny} = \{D1, D2, D8, D9, D11\}$$

$$Gain\,(S_{sunny}, Humidity) = .970 - (3/5)\,0.0 - (2/5)\,0.0 = .970$$

$$Gain\,(S_{sunny}, Temperature) = .970 - (2/5)\,0.0 - (2/5)\,1.0 - (1/5)\,0.0 = .570$$

$$Gain\,(S_{sunny}, Wind) = .970 - (2/5)\,1.0 - (3/5)\,.918 = .019$$

- Information gain is biased towards attributes with a large number of distinct values

$$G(S, A) = \text{Entropy}\left(S\right) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} \text{Entropy}\left(S_v\right)$$

  - Consider unique identifiers like ID or credit card
  - Such attributes have a high information gain, because they uniquely identify each instance, but we do not want to include them in the decision tree
    - E.g., deciding how to treat a customer based on their credit card number is unlikely to generalize to customers we haven't seen before.
  - Measures have been proposed that “correct” this issue:
    - Gini impurity (Gini index)

GINI IMPURITY

- Used in CART (Breiman et al., 1984)
- Measure of impurity or divergence within a dataset
  - The probability of a randomly chosen observation to be misclassified
- Let a dataset S containing examples from k classes. Let p_i be the probability of class i in S. The Gini Index of S is given by: $Gini(S) = 1 - \sum_{i=1}^{k} p_i^2$
- Gini index considers a binary split for each attribute A. Let Sis split based on A into two subsets S_1 and S_2.

$$Gini(S,A) = \frac{|S_1|}{|S|} Gini(S_1) + \frac{|S_2|}{|S|} Gini(S_2)$$

- We want to evaluate the reduction in the impurity of S based on A

$$\Delta Gini(S,A) = Gini(S) - Gini(S,A)$$

- The attribute A that provides the smallest Gini(S,A)(or the largest reduction in impurity) is chosen to split the node

SUM OF SQUARED ERRORS (SSE)

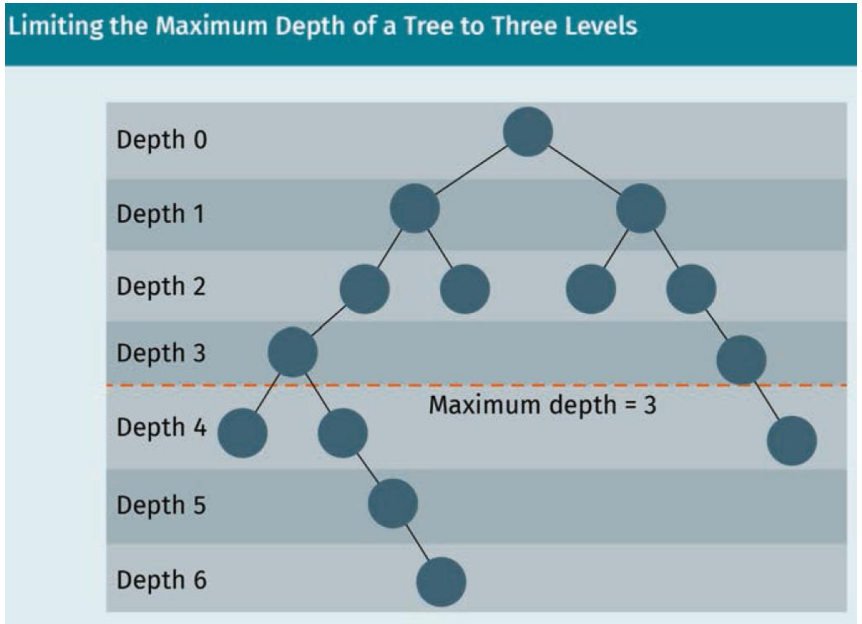
- Used in regression trees
- If we want to split a dataset S into two subsets S_1 and S_2,
  - Y: actual value, $\bar{y}$: mean value of the left/right side of the possible split

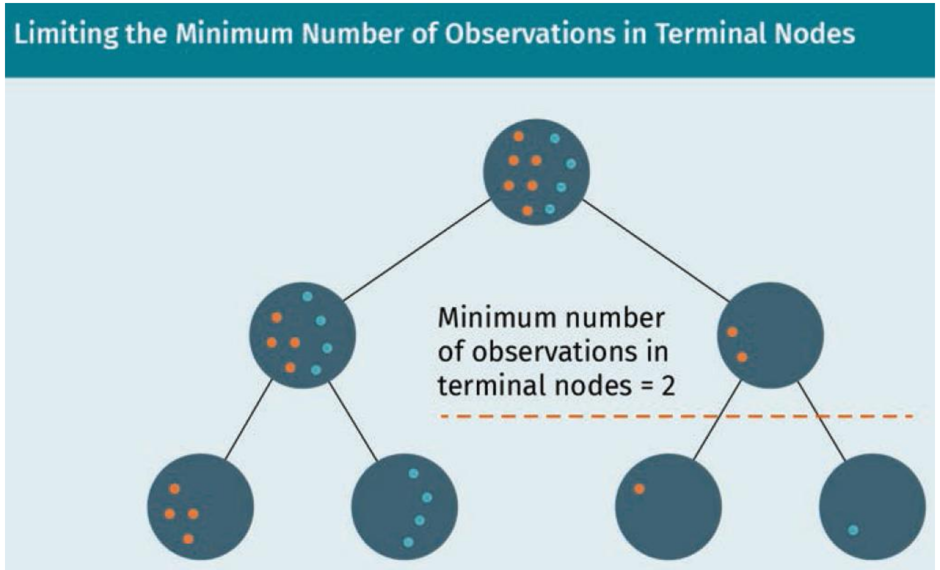
$$SSE = \sum_{i \in S_1} (y_i - \overline{y}_1)^2 + \sum_{i \in S_2} (y_i - \overline{y}_2)^2$$

- Finding the minimization of the SSE

STOP CRITERIA

- Choosing large trees with many leaves → risk of reducing prediction performance
- The goal is to build an ideal sized tree (Breiman et al., 1984)
- Stop criteria restrict the growth of the tree to avoid the risk of overfitting
  - Restrict the tree depth to a certain level
  - Restrict the minimum number of observations allowed in any terminal node





STOP CRITERIA

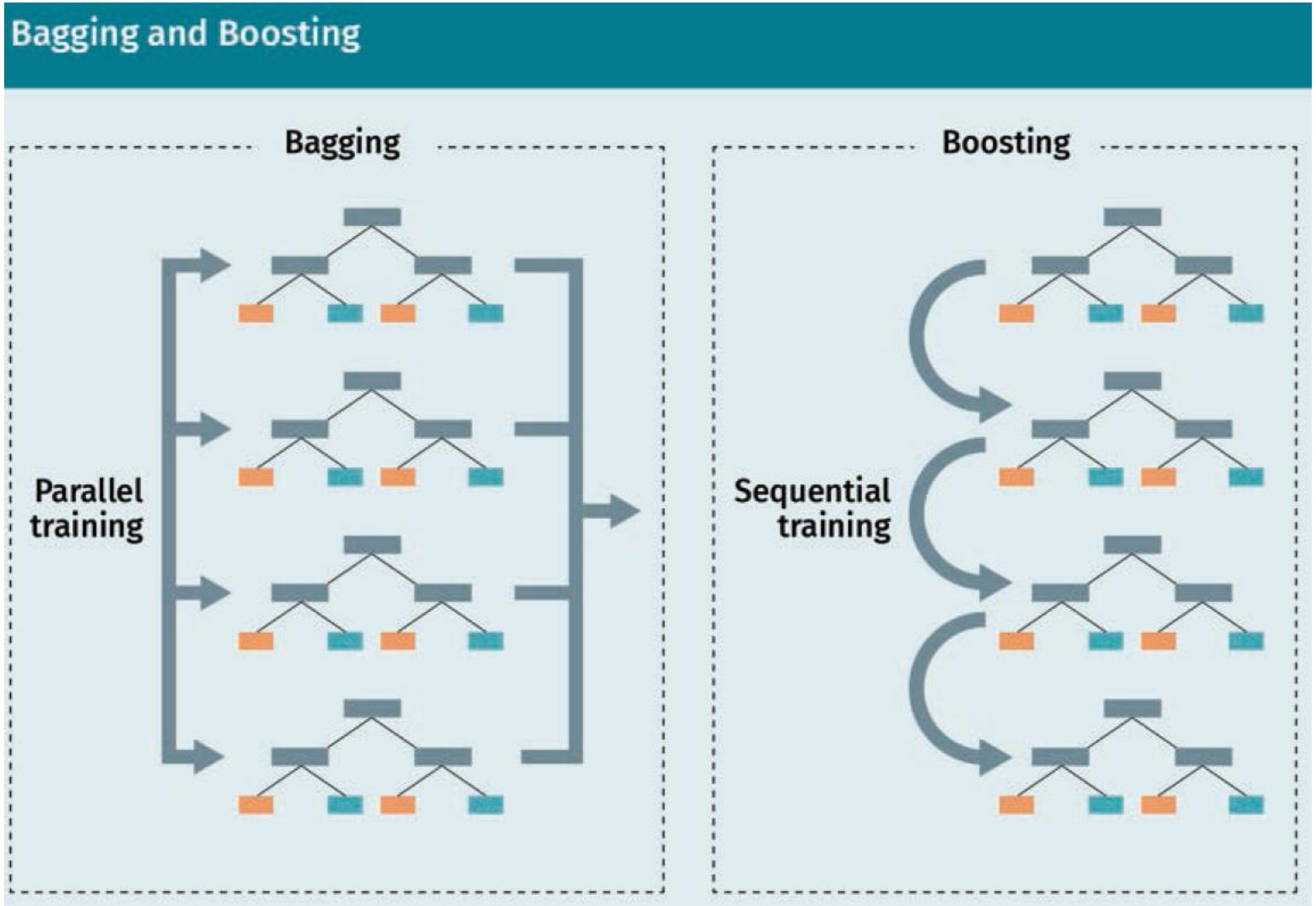
- Tree Pruning: we allow the tree to fully develop and later remove insignificant branches.
  - Starting at the leave nodes and moving toward the root of the tree
  - The branches are pruned according to the lowest level of influence on the prediction error Error(T) of tree T
  - This is done until the desired stop criterion is fulfilled
    - e.g., a defined maximum tree depth or a minimum number of observations per leaf node
  - The branching to be pruned in each pruning step, i.e., the pruning candidate C of tree T, is thus determined as follows:

$$C(T) = Error(T) + \lambda\ L(T)$$

  - Pre-specified cost complexity parameter λ that penalizes the number of terminal nodes L of tree T
  - Smaller values for the cost complexity parameter λ tend to produce larger trees; larger values for λ result in smaller trees.
  - Evaluate several models across a spectrum of λ and use cross-validation to identify the optimal value

- Several trees to be bundled together to form one strong estimator
- Ensemble methods can be divided into two categories:
  - Bagging algorithms: individual decision trees are independently trained in parallel
  - Boosting algorithms: decision trees are trained sequentially, and one tree takes the errors of
  - the previously constructed tree into consideration
- The most well-known representatives of bagging and boosting are:
  - Random forest
  - Gradient boosting

ENSEMBLE METHODS



STUDY GOALS REVIEW

- explain the concept of decision and regression trees.
- define bagging and boosting.
- apply decision tree and regression tree models on your own with the use of Python.

UNIT 5.1

TRANSFER TASK

Credit Score Classification: Case Study

– The **credit score** of a person determines the creditworthiness of the person. It helps financial companies determine if you can repay the loan or credit you are applying for.

– Explain and describe how decision tree or regression tree techniques might be applied.

TRANSFER TASK
PRESENTATION OF THE RESULTS

Please present your results.
The results will be discussed in plenary.