DOUGLAS COLLEGE

FUNDAMENTALS OF MACHINE LEARNING IN DATA SCIENCE

CSIS 3290

CLASSIFICATIONS

DECISION TREE

FATEMEH AHMADI

Decision Tree

- Decision tree induction is the learning of decision trees from class-labeled training datasets.
- A decision tree is a flowchart-like tree structure, where each **internal node (non-leaf node)** denotes a test on an attribute, each branch represents an outcome of the test, and each **leaf node** (or **terminal node**) holds a **class label**.
- Some decision tree algorithms produce only binary trees (where each internal node branches to exactly two other nodes), whereas others can produce nonbinary trees.

Decision Tree

- Because the class label of each training tuple is provided, this step is also known as **supervised learning** (i.e., the learning of the **classifier** is “supervised” in that it is told to which class each training tuple belongs).
- It contrasts with **unsupervised learning** (or clustering), in which the class label of each training tuple is not known, and the number or set of classes to be learned may not be known in advance.

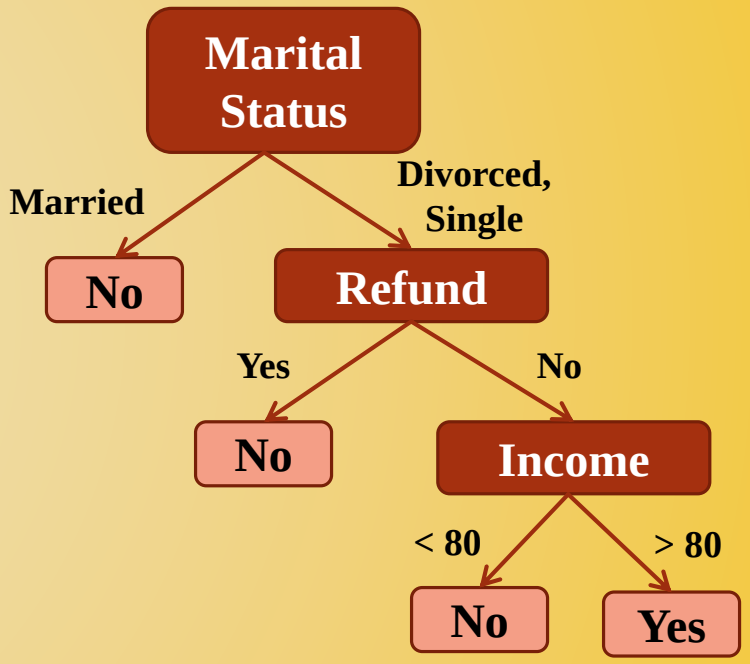
Decision Tree

Famous algorithms to build Decision Tree:

- ✓ **Hunt → E.B Hunt**
- ✓ **ID3 (Iterative Dichotomize)**
- ✓ **C4.5 – C5.0**
- ✓ **CART (Classification and Regression Trees)**

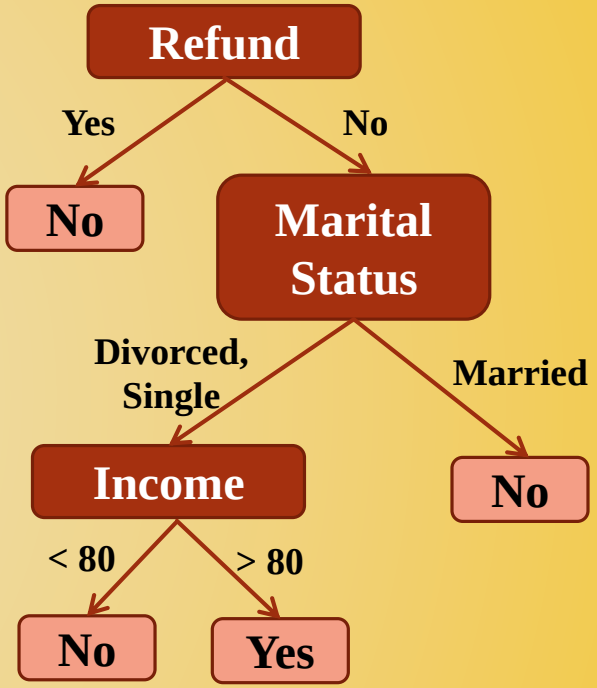
Decision Tree - Modeling

T_ID	Refund	Marital Status	Income	Cheat
1	Yes	Single	125	No
2	No	Married	100	No
3	No	Single	70	No
4	Yes	Married	120	No
5	No	Divorced	95	Yes
6	No	Married	60	No
7	Yes	Divorced	220	No
8	No	Single	85	Yes
9	No	Married	75	No
10	No	Single	90	No



Decision Tree - Prediction

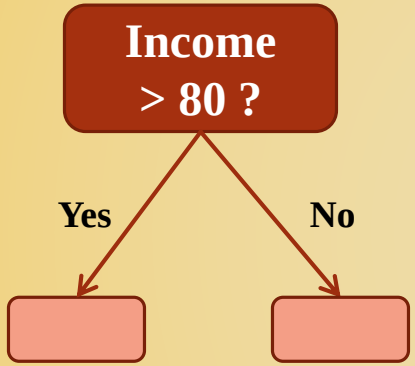
T_ID	Refund	Marital Status	Income	Cheat
1	Yes	Single	125	No
2	No	Married	100	No
3	No	Single	70	No
4	Yes	Married	120	No
5	No	Divorced	95	Yes
6	No	Married	60	No
7	Yes	Divorced	220	No
8	No	Single	85	Yes
9	No	Married	75	No
10	No	Single	90	No



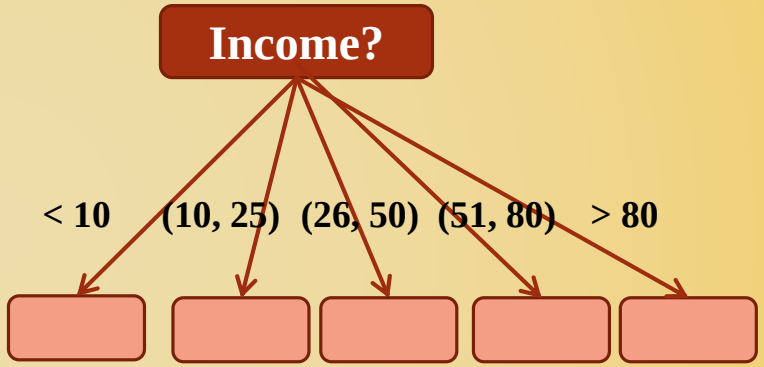
New Data
Cheat: Yes OR No

Refund	Marital Status	Income	Cheat
No	Married	130	?

Decision Tree



Binary Split

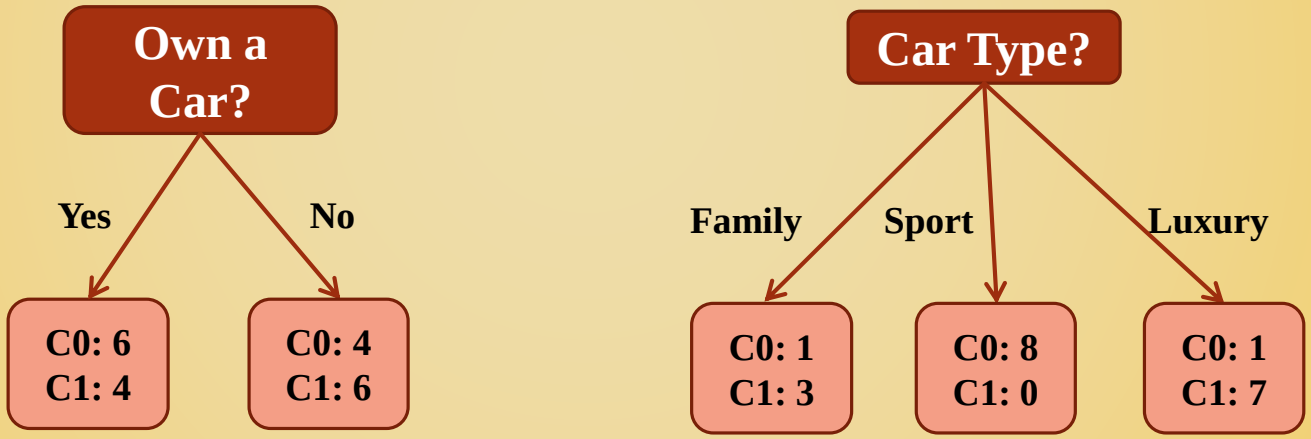


Multi-Way Split

Decision Tree: Node Purity

For two attributes of Own a Car and Car Type

Here 10 records are belong to C0 and 10 record belong to C1



The question is: Which split is the best in which the most information is achieved from the records

Impurity Calculation: Gini Index

$$Gini(t) = 1 - \sum_{j} [p(j \mid t)]^2$$

Worst case: Gini(t) = 1- $(1/n_c)^2$ - $(1/n_c)^2$ - $(1/n_c)^2$ ….- $(1/n_c)^2$ = 1- $n_c(1/n_c)2$ = 1-$1/n_c$

Best case: Gini(t) = 1- $(1)^2$ – $(0)^2$ - … - $(0)^2$ = 0

C1	0
C2	6

$P(C_1)$ = 0/6 = 0 $P(C_2)$ = 6/6 =1
Gini(t) = 1- $p(C_1)^2$ – $P(C_2)^2$ = 1-0-1 = 0

C1	1
C2	5

$P(C_1)$ = 1/6 $P(C_2)$ = 5/6
Gini(t) = 1- $(1/6)^2$- $(5/6)^2$ = 0.278

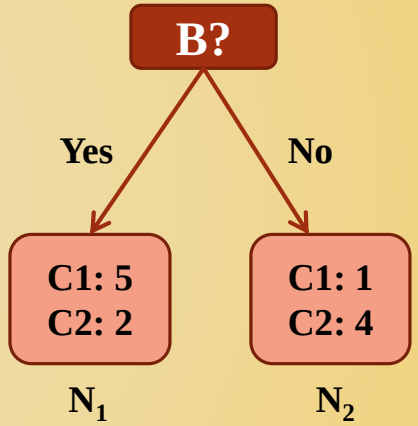
C1	2
C2	4

$P(C_1)$ = 2/6 $P(C_2)$ = 4/6
Gini(t) = 1- $(2/6)^2$- $(4/6)^2$ = 0.444

Impurity Calculation: Gini Index

$$Gini_{split} = \sum_{i=1}^{k} \frac{n_i}{n} Gini(i)$$

	N_1	N_2
C_1	5	1
C_2	2	4



Gini(N_1) = 1- $(5/7)^2$-$(2/7)^2$ = 0.41
Gini(N_2) = 1- $(1/5)^2$-$(4/5)^2$ = 0.32
Gini(B) = 7/12* 0.41 + 5/12 * 0.32 = 0.372

Impurity Calculation: Entropy

$$Entropy(t) = -\sum_{j} P(j|t) log P(j|t)$$

Worst case: Entropy(t) = - $(1/n_c)\log 1/n_c$ - $(1/n_c)\log 1/n_c$ ….- $(1/n_c)\log 1/n_c$ = - n_c $(1/n_c)\log 1/n_c$ = $\log n_c$

Best Case: Entropy(t) = - (1)log 1 – (0) log 0 - … - (0) log 0 = 0

C1	0
C2	6

$P(C_1)$ = 0/6 = 0 $P(C_2)$ = 6/6 =1
Entropy = -0 log 0 – 1log1 = -0-0 = 0

C1	1
C2	5

$P(C_1)$ = 1/6 $P(C_2)$ = 5/6
Entropy = - (1/6)log(1/6)-(5/6)log(5/6) =0.65

C1	2
C2	4

$P(C_1)$ = 2/6 $P(C_2)$ = 4/6
Entropy = - (2/6)log(2/6)-(4/6)log(4/6) =0.92

Impurity Calculation: Classification Error

$$Error(t) = 1 - max\ P(i \mid t)$$

Worst Case: Error(t) = 1- max($1/n_c$, $1/n_c$, …, $1/n_c$)= 1- $1/n_c$

Best Case: Error(t) = 1- max(1, 0, …, 0) = 0

C1	0
C2	6

$P(C_1)$ = 0/6 = 0 $P(C_2)$ = 6/6 =1
Error = 1- max (0, 1) = 1-1 = 0

C1	1
C2	5

$P(C_1)$ = 1/6 $P(C_2)$ = 5/6
Error = 1- max (1/6, 5/6) = 1-5/6 = 1/6

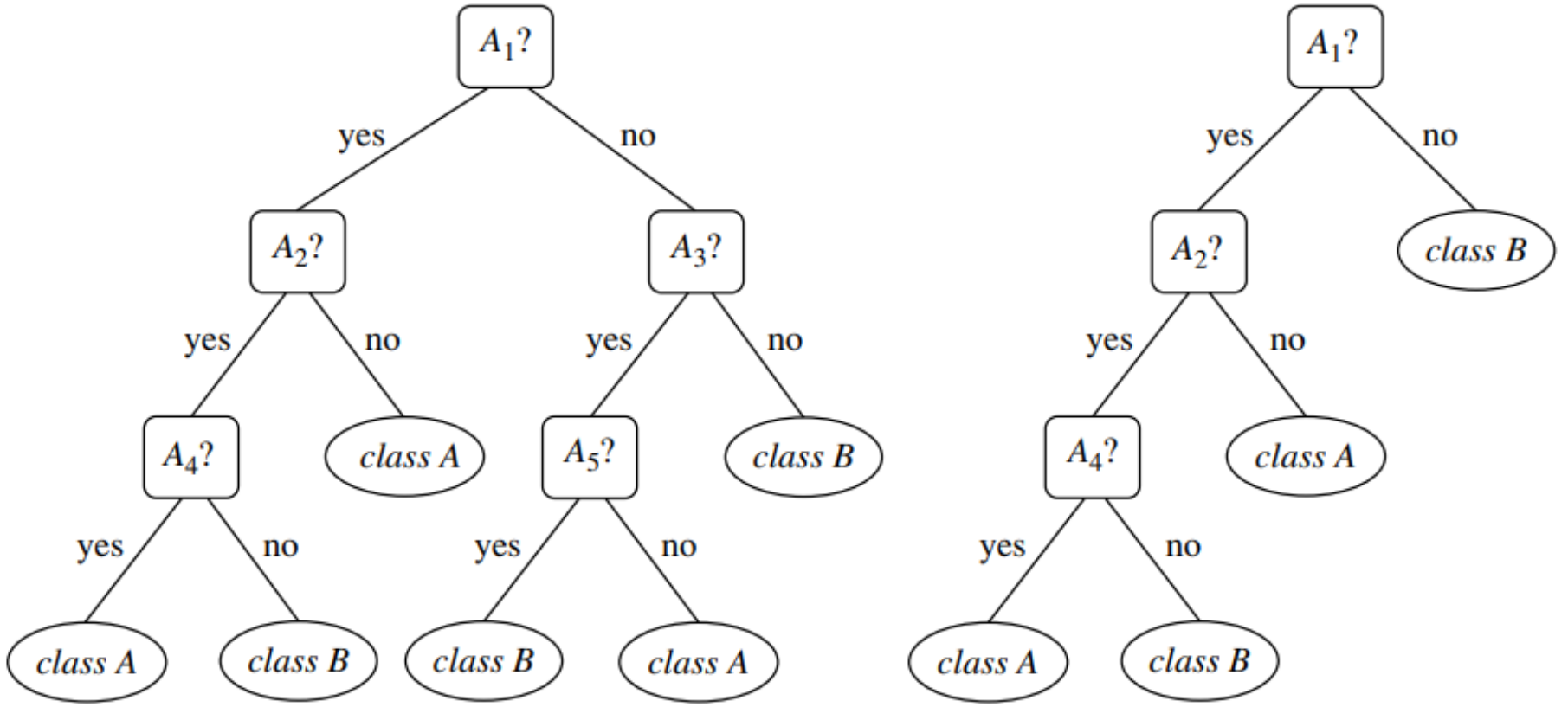
C1	2
C2	4

$P(C_1)$ = 2/6 $P(C_2)$ = 4/6
Error = 1- max (2/6, 4/6) = 1-4/6 = 1/3

Tree Pruning

- When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods address this problem of overfitting the data. Such methods typically use statistical measures to remove the least-reliable branches.
- An unpruned tree and a pruned version of it are shown in Figure 8.6. Pruned trees tend to be smaller and less complex and, thus, easier to comprehend. They are usually faster and better at correctly classifying independent test data (i.e., of previously unseen tuples) than unpruned trees. “How does tree pruning work?”
- There are two common approaches to tree pruning: **pre-pruning** and **post-pruning.**

Tree Pruning



An unpruned decision tree and a pruned version of it.

Advantages of Decision Trees

- One of the main benefits of decision trees is that they are easy to understand and interpret.
- You can visualize the structure and logic of the tree, and trace how each decision is made. This makes them transparent and explainable, unlike some other complex models such as neural networks.
- Another advantage of decision trees is that they can handle both numerical and categorical data, and do not require much preprocessing or scaling.
- They can also deal with missing values and outliers by creating separate branches or ignoring them.
- Furthermore, decision trees can capture non-linear relationships and interactions among the features, and perform well on both large and small datasets.

Reference

Data Mining, Concepts and Techniques, Jiawei Han, Micheline Kamber, Jian Pei. MK. Chapter 8.