Data Science Fundamentals: CE 364

Course Lecturer
Vincent M. Nofong, Ph.D.

Computer Science and Engineering Department
University of Mines and Technology, Tarkwa - Ghana

June 26, 2024

Introduction

Outline

- Who I am
- Course Information and Outline of CE 364
- Expected Learning Outcomes
- Rules
- Chapter One: Introduction

Introduction

About me

- Name: **Vincent M. Nofong, PhD**
- Email: **vnofong@umat.edu.gh**
- Personal Website: https://vincentnofong.com/
- Uni website: `https://www.umat.edu.gh/staffinfo/staffDetailed.php?contactID=385`
- Office hours (Working days): **09:00 am - 16:00 pm GMT**
- Research interest: **data mining, trend prediction, classification, bioinformatics, artificial intelligence, machine learning**

Introduction

Course Information (CE 364)

- Credit hours: **2**
- Attendance: **10%**
- Continuous Assessment: **30%**
  - Quizzes - two or three
  - Programming assignment - one
    - Marking will be one-on-one code explanation and modification
- End of Semester: **60%**

Introduction

Course Outline (CE 364)

1. Introduction and Data Exploration
2. Classification
3. Regression Methods
4. Association Analysis
5. Clustering
6. Text Mining
7. Time Series Forecasting
8. Anomaly Detection
9. Model Evaluation and Feature Selection
10. Data Visualization and Data Ethics

Introduction

Expected Learning Outcomes (CE 364)

Students should understand and be able to:

1. Explain the fundamental concepts in data science.
2. Apply data science techniques for knowledge discovery from data.
3. Present and communicate knowledge discovered from data effectively.
4. Utilize various data science tools and software for data analysis.
5. Implement machine learning algorithms for predictive modeling.
6. Conduct data preprocessing and cleaning to prepare datasets for analysis.
7. Interpret and evaluate the results of data science experiments.
8. Develop visualizations to represent data insights and findings.
9. Integrate data science methods into real-world problem-solving scenarios.
10. Understand ethical considerations and best practices in data science.

Introduction

Reference Materials

1. Kroese, Dirk P., Zdravko Botev, and Thomas Taimre. Data science and machine learning: mathematical and statistical methods. Chapman and Hall/CRC, 2023.
2. Kotu, Vijay, and Bala Deshpande. Data science: concepts and practice. Morgan Kaufmann, 2018.
3. Grus, Joel. Data science from scratch: first principles with python. O'Reilly Media, 2019.
4. Aggarwal, Charu C. Data mining: the textbook. Vol. 1. New York: Springer, 2015.
5. Zaki, Mohammed J., and Wagner Meira. Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press, 2014.

Introduction

Rules

1. Feel free to ask questions in class, unless they are too "personal".
2. Students should not be late for lectures or practicals.
3. Students should attend all lectures and practicals.
4. **In case you are unable to attend lectures or will be late, send me an email - at least 30 minutes before lectures.**
5. Students should do and submit all assignments before the given deadline.
6. Unless otherwise permitted, students should not use their mobile phones in class - note usage of Laptops/Desktops is permitted.

Classification

What is Classification?

Classification

What is Classification?

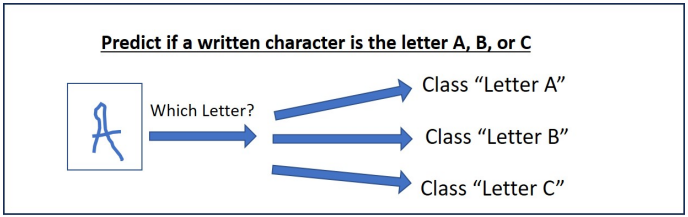
- Classification is the process of categorizing objects or instances into predefined groups or classes based on their attributes and characteristics
- It is referred to as supervised learning because an example data set is used to learn the structure of the groups.
- It involves using algorithms (techniques) to identify which category an item belongs to among several possible options.

Classification

Examples: (A) Movie Preference Prediction

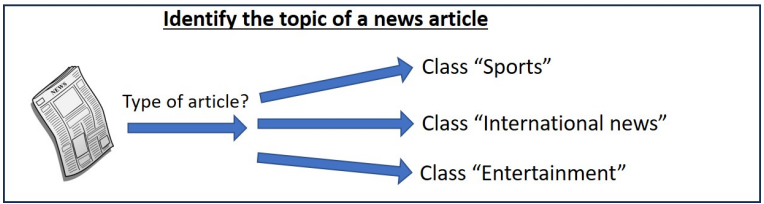
1. **Compare Preferences:**
   - Analyze the person's likes and preferences.
   - Compare these with the preferences of people who like the movie and those who do not.
2. **Make a Prediction:**
   - Determine which group the person is more similar to.
   - Based on this comparison, predict whether the person will like the movie or not.

Examples: (B) Hand Written Character Prediction



1. **Compare Features:**
   - Analyze the features of the handwritten character.
   - Compare these with the features of several known handwritten characters.
2. **Make a Prediction:**
   - Determine which character group the handwritten character is more similar to.
   - Based on this comparison, predict the character written.

Examples: (C) News Article Topic Categorization



1. **Compare Features:**
   - Analyze the features of the news article.
   - Compare these features with those of articles in known categories (e.g., sports, entertainment, politics).
2. **Make a Prediction:**
   - Determine which topic category the news article is more similar to.
   - Based on this comparison, predict the topic of the news article.

Classification

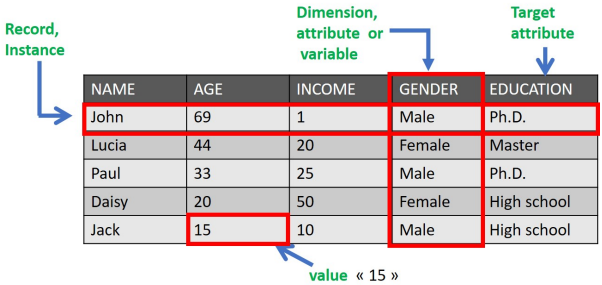
Phases of Classification

Most classification algorithms typically have two phases:

1. **Training phase**
   - In this phase, a training model is constructed from the training instances.
2. **Testing phase**
   - In this phase, the training model is used to determine the class label (or group identifier) of one or more unseen test instances.

Classification

Typical Data Used in Classification



NAME	AGE	INCOME	GENDER	EDUCATION
John	69	1	Male	Ph.D.
Lucia	44	20	Female	Master
Paul	33	25	Male	Ph.D.
Daisy	20	50	Female	High school
Jack	15	10	Male	High school

- Data is usually stored in a table.
- **Instance/Record:** A row in the table.
- **Dimension/Attribute/Variable:** A column in the table.
- **Value:** The data in a cell.
- **Target Attribute:** The attribute to be predicted.

Classification

Goal of Classification

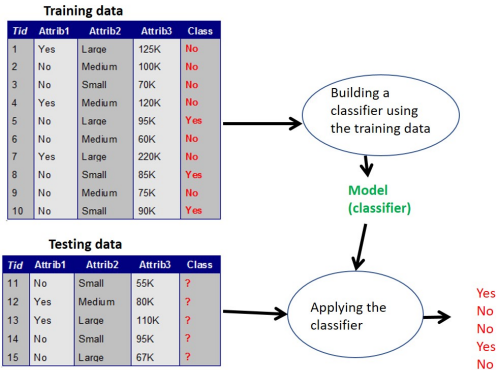
- Predict the value of the target attribute for new data.
- Example: Given the training data in the table below, what is the highest educational level of Victoria?

	NAME	AGE	INCOME	GENDER	EDUCATION
Training data	John	69	1	Male	Ph.D.
	Lucia	44	20	Female	Master
	Paul	33	25	Male	Ph.D.
	Daisy	20	50	Female	High school
	Jack	15	10	Male	High school
	Victoria	35	50	Female	?????????

Classification

Building/Using a Classifier

- A classifier is a model firstly built from the training data.
- The model is then used to predict the values of the target attribute based on the values of other attributes as shown below.



Classification

What is a Good Classifier?

A good classifier:

- Can perform predictions for new records.
- Can perform accurate predictions.

Performance Measures:

- Accuracy: $\frac{\text{Number of correct predictions}}{\text{Number of records}}$
- Precision: $\frac{\text{Number of correct predictions}}{\text{Number of predictions}}$
- Other Measures:
  - Recall
  - F1 Score
  - ROC-AUC

Classification

Classification vs Regression

- **Classification:** Target attribute is a discrete value (e.g., fraud or not fraud).
- **Regression:** Target attribute is a continuous value (e.g., weight of a person).

Suitability:

- Classification works well for predicting binary or nominal attributes.
- It may not work as well for ordinal (e.g. small, medium, larger) or hierarchical (e.g. human, mammal, animal,) attributes.

Classification

Interpretable and Complex Classifiers:

- Some classifiers indicate the criteria used to distinguish between classes (e.g., decision trees, some associative classifiers).
- Other classifiers, such as neural networks, may be difficult for humans to interpret despite their effectiveness.

Types Classifiers:

- Decision Trees (CART, ID3, C4.5)
- Neural Networks / Deep Learning
- Support Vector Machines (SVM)
- Naïve Bayes Classifier
- Associative Classifiers, etc.

We will discuss a few

Applications of Classification

Classification

Applications of Classification

There are several applications of classification:

- Customer Target Marketing
- Medical Disease Management
- Document Categorization and Filtering
- Multimedia Data Analysis
- Detecting Malignant Human Cells
- Identifying Legitimate or Fraudulent Credit Card Transactions
- Determining the Topic of a News Article (e.g., sports, entertainment, weather)
- Predicting Political Views, Age, and Gender on Social Networks

Types of Classifiers: Decision Trees

- A classification methodology where the classification process is modeled using a set of hierarchical decisions on feature variables, arranged in a tree-like structure.
- The decision at a node, known as the *split criterion*, is typically a condition on one or more feature variables in the training data.
- This criterion divides the training data into two or more parts.

Classification

Types of Classifiers: Decision Trees

Goal of Splitting:

- The aim is to identify a split criterion that reduces the "mixing" of class variables in each branch of the tree as much as possible.

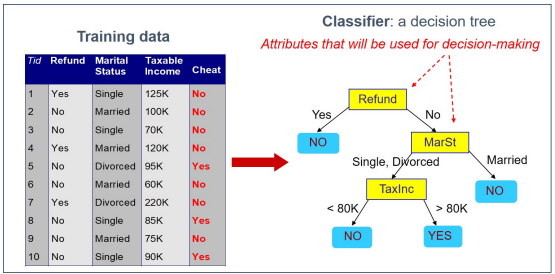
Types of Splits:

- **Univariate Splits:** Use only one attribute as the split criterion.
- **Multivariate Splits:** Use more than one attribute in the split criterion.

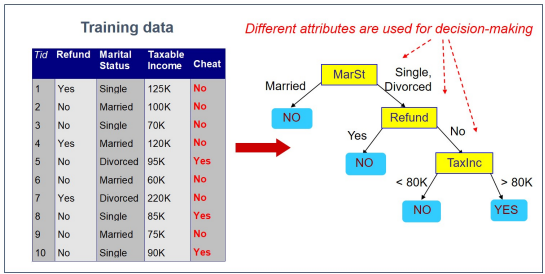
For a given dataset, several decision trees with different attributes may be created.

Types of Classifiers: Decision Trees

- For a given dataset, several decision trees with different attributes may be created.
- Which tree is better?



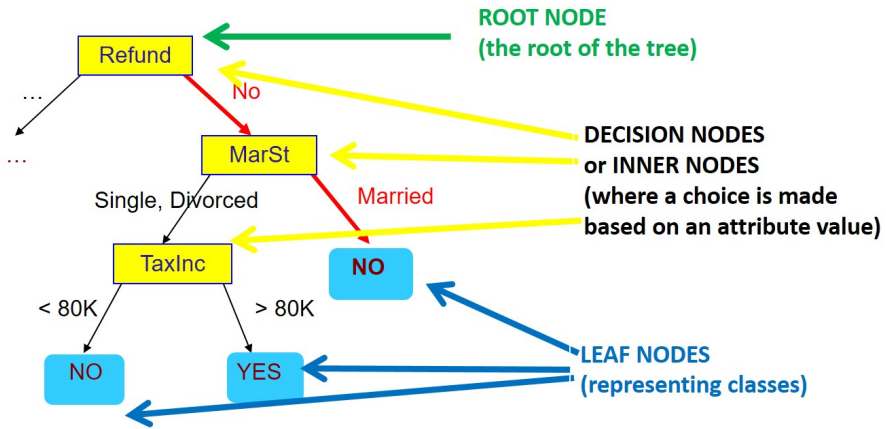
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



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

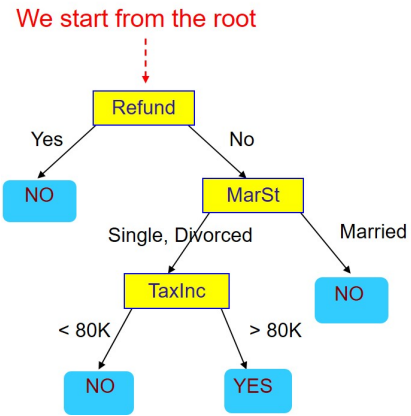
Types of Classifiers: Decision Trees

Vocabulary on Decision Trees



Types of Classifiers: Decision Trees

Using Decision Trees for Prediction (1/6)

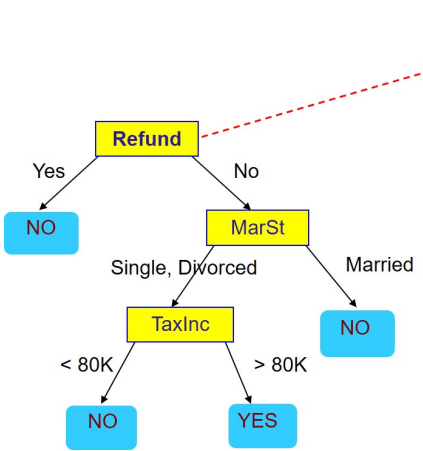


Testing data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

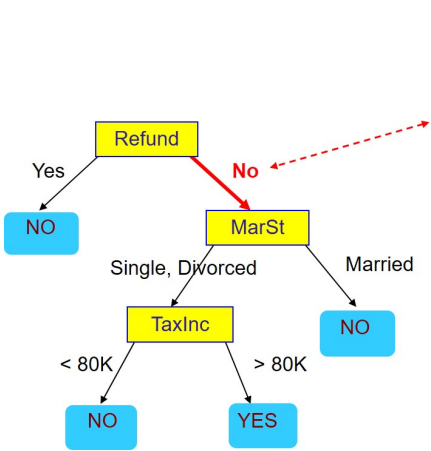
Types of Classifiers: Decision Trees

Using Decision Trees for Prediction (2/6)



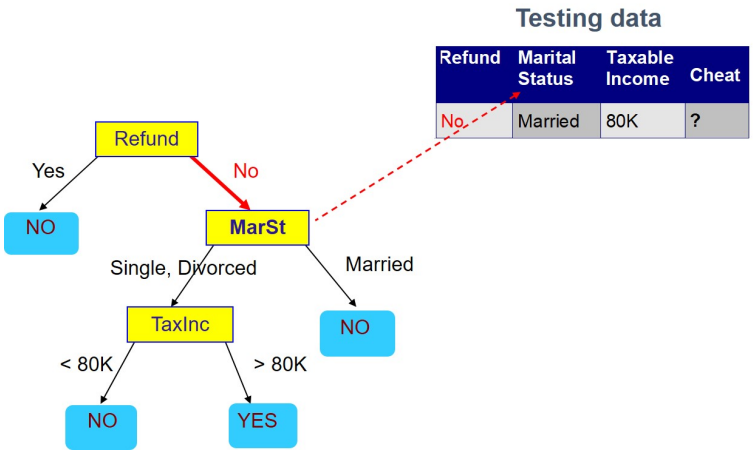
Types of Classifiers: Decision Trees

Using Decision Trees for Prediction (3/6)



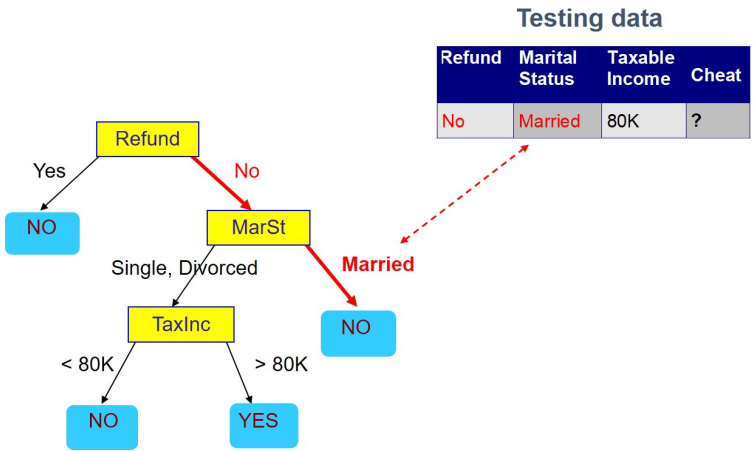
Types of Classifiers: Decision Trees

Using Decision Trees for Prediction (4/6)



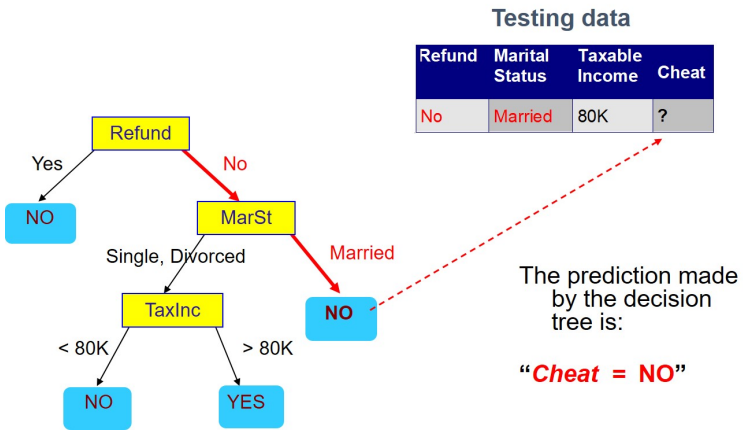
Types of Classifiers: Decision Trees

Using Decision Trees for Prediction (5/6)



Types of Classifiers: Decision Trees

Using Decision Trees for Prediction (6/6)



Types of Classifiers: Decision Trees

How Decision Trees are Built - Hunt's Algorithm (1/8)

- Let D_t be the set of records reaching a node t.
- Initially, D_t includes all records in the database.
- Procedure:
- If all records in D_t belong to the same class y_t, then node t becomes a leaf with the label y_t.
- If D_t is empty, then node t becomes a leaf node with the default class y_d.
- If records in D_t belong to multiple classes, node t becomes a decision node, and an attribute is used to split the records.

Types of Classifiers: Decision Trees

How Decision Trees are Built - Hunt's Algorithm (2/8)

Training data

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

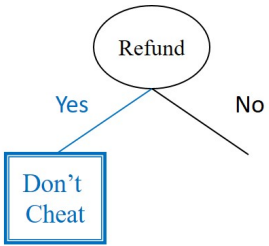
The target attribute is « **cheat** »

The records do not belong to the same class (we have **Yes** and **No** for the **« Cheat »** attribute).

We can choose the « *refund* » atribute to try to separate the records.

Types of Classifiers: Decision Trees

How Decision Trees are Built - Hunt's Algorithm (3/8)



If « refund = Yes » then all records belong to the same class (« Cheat = No »)

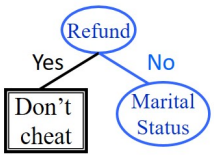
Hence, we create a **leaf node** « don't cheat ».

Training data

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

Types of Classifiers: Decision Trees

How Decision Trees are Built - Hunt's Algorithm (4/8)



If « refund = No » then all records do not belong to the same class.

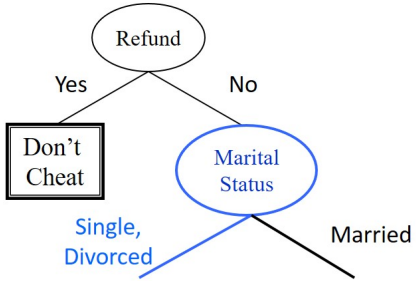
Hence, we must create a decision node. We can choose the attribute « marital status ».

Training data

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

Types of Classifiers: Decision Trees

How Decision Trees are Built - Hunt's Algorithm (5/8)



If « refund = No » and « Marital status = single or divorced » not all records are of the same class.

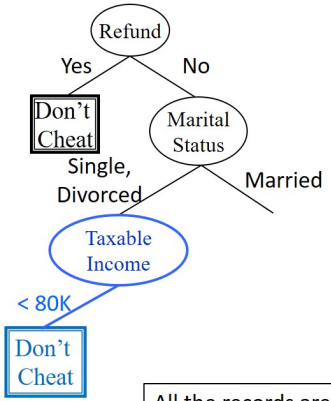
We can create a node « income » to try to separate the records

Training data

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

Types of Classifiers: Decision Trees

How Decision Trees are Built - Hunt's Algorithm (6/8)



Training data

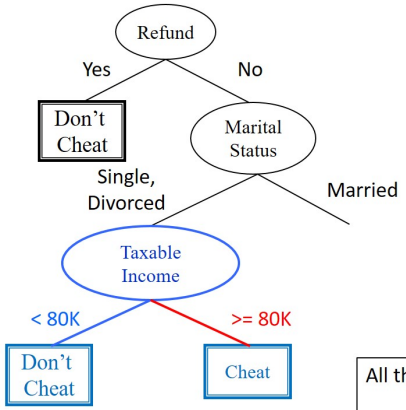
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

All the records are of the same class

Thus we create a leaf node

Types of Classifiers: Decision Trees

How Decision Trees are Built - Hunt's Algorithm (7/8)



Training data

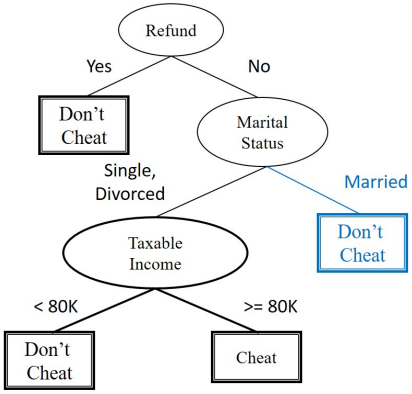
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

All the records are of the same class

Thus we create leaf nodes

Types of Classifiers: Decision Trees

How Decision Trees are Built - Hunt's Algorithm (8/8)



Training data

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

All records are of the same class

We create a leaf node

End of Decision Tree Building

Types of Classifiers: Decision Trees

How to choose the attributes for building a decision tree?

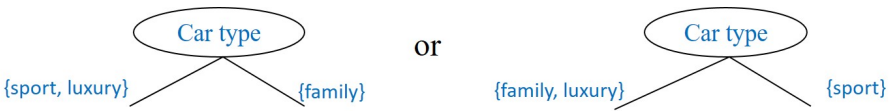
- The "Greedy" Approach:
  - Decision trees are built by always choosing the attribute that best separates the data using a single attribute.
  - The goal is to obtain the best possible tree, though it is not guaranteed.
- Challenges:
  - It is sometimes possible to separate records using many different attributes.
  - Deciding when to stop growing the tree. Should we use a small tree or a very large tree?
  - Determining the criterion to use for separating records - depends on attributes

Types of Classifiers: Decision Trees

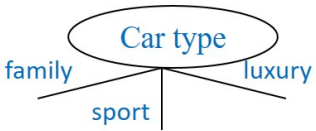
How to choose the attributes for building a decision tree?

For nominal attributes:

- **Binary split:** only two branches, we must find the best way to separate the records



- **Multiple splits:** a branch for each value.

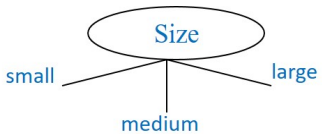


Types of Classifiers: Decision Trees

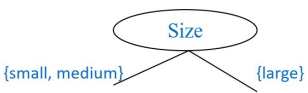
How to choose the attributes for building a decision tree?

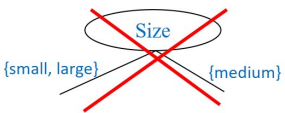
For ordinal attributes:

- **Multiple split:** a branch for each value



- **Binary split:** two branches. The order between values must be respected



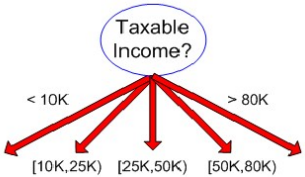


Types of Classifiers: Decision Trees

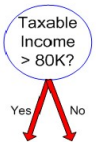
How to choose the attributes for building a decision tree?

For continuous attributes:

- Separate the continuous attribute values into several distinct ranges.



- Make a binary decision based on the continuous attribute.



Types of Classifiers: Decision Trees

Why Use Decision Trees?

- **Easy to Understand:** Small trees are easy for humans to interpret.
- **Fast Building Process:** Building a decision tree is very fast.
  - Complexity: $O(n \times d \times log(d)$ where n is the number of attributes, and d is the number of records.
- **Efficient Classification:** Classifying new instances is extremely fast.
  - Complexity: $O(w)$ where w is the tree depth.

Types of Classifiers: Decision Trees

Why Use Decision Trees? (2/2)

- **Comparable Accuracy:** Accuracy is similar to other classifiers for simple data.
- **Noise Tolerance:** Decision trees can be quite tolerant of noise in the data.
- **Overfitting Avoidance:** With certain techniques, decision trees can avoid the problem of overfitting.