Data Mining

Dr. Raed Ibraheem Hamed

University of Human Development,
College of Science and Technology
Department of Computer Science

2016 – 2017

Road Map

- Classification: Basic Concepts
- Decision Tree Induction
- Using IF-THEN Rules for Classification
- Rule Extraction from a Decision Tree
- Rule Generation
- Illustrating Classification Task
- Example of a Decision Tree
- Apply Model to Test Data

Classification: Basic Concepts

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks.

Classification Examples

1. Teachers classify students' grades as **A, B, C, D, or F.**
2. Identify mushrooms as poisonous or edible.
3. Predict when a river will flood.
4. Credit/loan approval:
5. Medical diagnosis: if a tumor is cancerous or benign
6. Fraud detection: if a transaction is fraudulent

Using IF-THEN Rules for Classification

- Represent the knowledge in the form of IF-THEN rules

  Rule: IF age = youth AND student = yes THEN buys_computer = yes

  - Rule antecedent vs. rule consequent
- Assessment of a rule: coverage and accuracy

Rule: $\underbrace{\text{IF age = youth AND student = yes}}_{\text{antecedent}}$ THEN $\underbrace{\text{buys\_computer = yes}}_{\text{consequent}}$

Rule Extraction from a Decision Tree

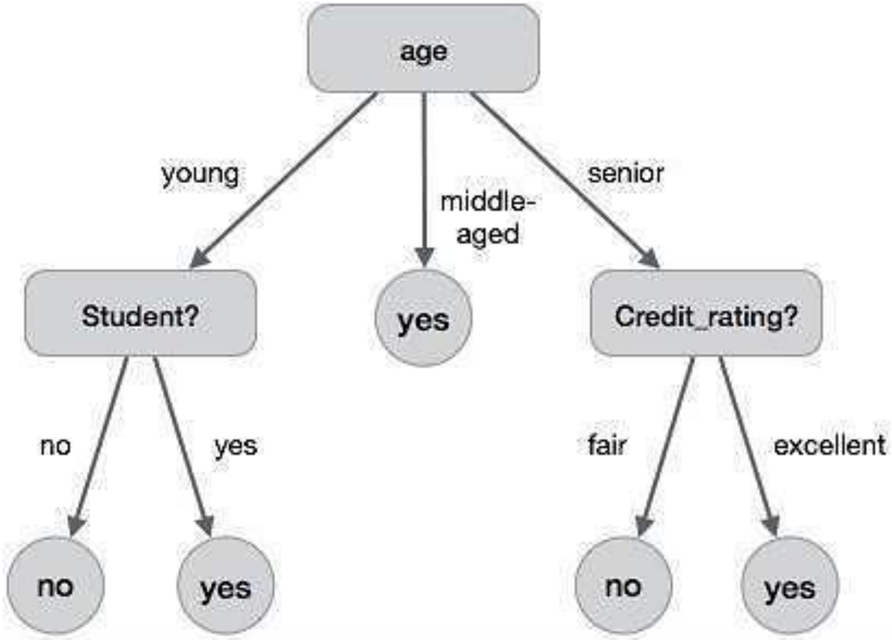
- Rules are easier to understand than large trees
- One rule is created for each path from the root to a leaf
- Each attribute-value pair along a path forms a **conjunction**: the leaf holds the **class** prediction
- Rules are mutually exclusive and exhaustive

The benefits of having a decision tree are as follows :-

1. It does not require any domain knowledge.
2. It is easy to comprehend.
3. The learning and classification steps of a decision tree are simple and fast.

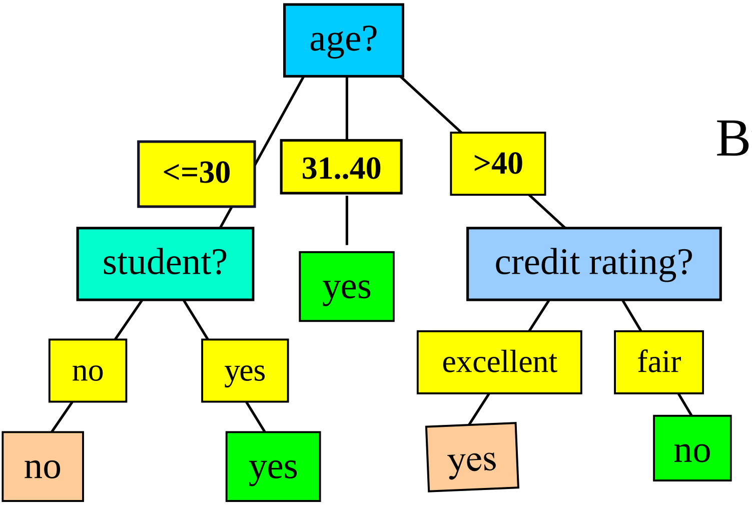
Data Mining - Decision Tree

The following decision tree is for the concept **buy_computer** that indicates whether a customer at a company is likely to **buy a computer** or **not**. Each internal node represents a test on an attribute. Each leaf node represents a class.



Rule Extraction from a Decision Tree

Decision-tree



Buys_computer

- **Example: Rule extraction from our buys_computer decision-tree**

IF *age* = young AND *student* = *no* THEN *buys_computer* = *no*

IF *age* = young AND *student* = *yes* THEN *buys_computer* = *yes*

IF *age* = mid-age THEN *buys_computer* = *yes*

IF *age* = old AND *credit_rating* = *excellent* THEN *buys_computer* = *yes*

IF *age* = old AND *credit_rating* = *fair* THEN *buys_computer* = *no*

Application of Rule-Based Classifier

- A rule ***r*** covers an instance **x** if the attributes of the instance satisfy the condition of the rule

R1: (Give Birth = no) ∧ (Can Fly = yes) → Birds

R2: (Give Birth = no) ∧ (Live in Water = yes) → Fishes

R3: (Give Birth = yes) ∧ (Blood Type = warm) → Mammals

R4: (Give Birth = no) ∧ (Can Fly = no) → Reptiles

R5: (Live in Water = sometimes) → Amphibians

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
hawk	warm	no	yes	no	?
grizzly bear	warm	yes	no	no	?

The rule R1 covers a hawk => Bird

The rule R3 covers the grizzly bear => Mammal

Rule Coverage and Accuracy

- Coverage of a rule:
  - Fraction of records that satisfy the antecedent of a rule
- Accuracy of a rule:
  - Fraction of records that satisfy both the antecedent and consequent of a rule

Tid	Refund	Marital Status	Taxable Income	Class
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**

(Status=Single) → No

Coverage = 40%, Accuracy = 50%

How does Rule-based Classifier Work?

R1: (Give Birth = no) ∧ (Can Fly = yes) → Birds

R2: (Give Birth = no) ∧ (Live in Water = yes) → Fishes

R3: (Give Birth = yes) ∧ (Blood Type = warm) → Mammals

R4: (Give Birth = no) ∧ (Can Fly = no) → Reptiles

R5: (Live in Water = sometimes) → Amphibians

A grizzly bear triggers rule R3, so it is classified as a mammal

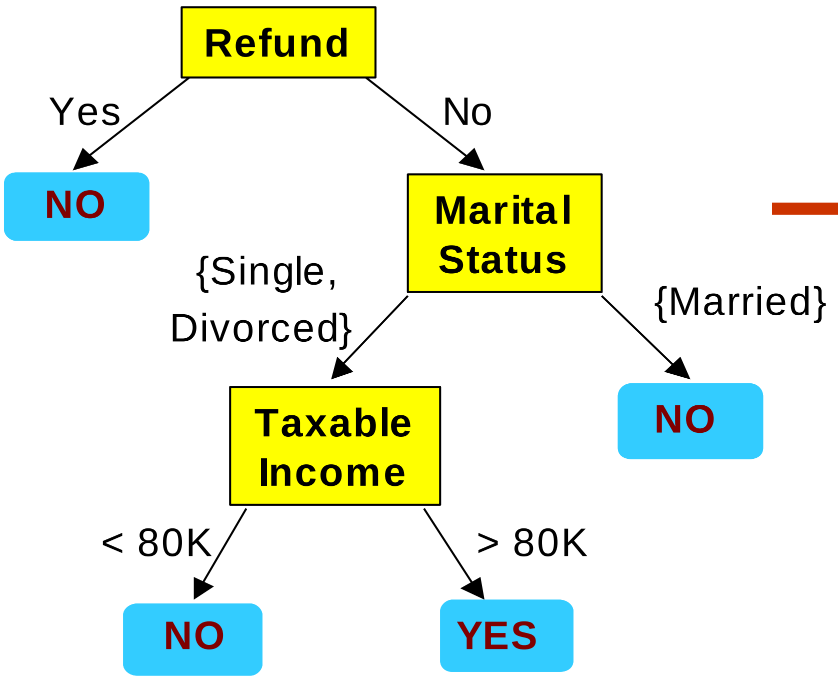
A Salomon triggers rule R3, so it is classified as a fish

A hawk triggers rule R1, so it is classified as a bird

Characteristics of Rule-Based Classifier

- Mutually exclusive rules
- Every record is covered by at most one rule
- Exhaustive rules
- Similar expressions to those of decision trees

From Decision Trees To Rules



Classification Rules

(Refund=Yes) ==> No

(Refund=No, Marital Status={Single,Divorced}, Taxable Income<80K) ==> No

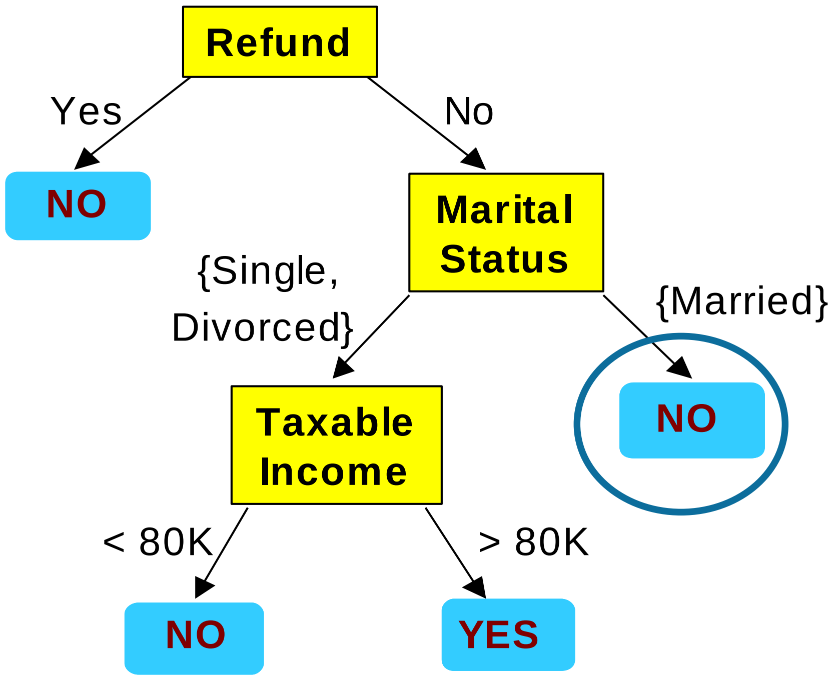
(Refund=No, Marital Status={Single,Divorced}, Taxable Income>80K) ==> Yes

(Refund=No, Marital Status={Married}) ==> No

Rules are mutually exclusive and exhaustive

Rule set contains as much information as the tree

Rules Can Be Simplified



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

Initial Rule: **(Refund=No) $\wedge$ (Status=Married) → No**

Simplified Rule: **(Status=Married) → No**

Rule Ordering Schemes

- Rule-based ordering
  - Individual rules are ranked based on their quality
- Class-based ordering
  - Rules that belong to the same class appear together

Rule-based Ordering

(Refund=Yes) ==> No

(Refund=No, Marital Status={Single,Divorced}, Taxable Income<80K) ==> No

(Refund=No, Marital Status={Single,Divorced}, Taxable Income>80K) ==> Yes

(Refund=No, Marital Status={Married}) ==> No

Class-based Ordering

(Refund=Yes) ==> No

(Refund=No, Marital Status={Single,Divorced}, Taxable Income<80K) ==> No

(Refund=No, Marital Status={Married}) ==> No

(Refund=No, Marital Status={Single,Divorced}, Taxable Income>80K) ==> Yes

Building Classification Rules

- Direct Method:
  - Extract rules directly from data
- Indirect Method:
  - Extract rules from other classification models (e.g. decision trees, neural networks, etc).

Thank you