Data Mining & Data Warehouse

Dr. Raed Ibraheem Hamed

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

2016 – 2017

Road Map

- Classification: Basic Concepts
- Examples of Cases Where The Data Analysis
- Classification Examples
- Classification Techniques
- Rule-Based Classification
- Rule-Based Classifier with Decision Tree
- Model Evaluation and Selection
- Illustrating Classification Task
- Classification—A Two-Step Process

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.

Examples of Cases Where The Data Analysis Task is Classification

Following are the examples of cases where the data analysis task is Classification −

- A bank loan officer wants to analyze the data in order to know which customer (loan applicant) are risky or which are safe.

- A marketing manager at a company needs to analyze a customer with a given profile, who will buy a new computer.

In both of the above examples, a **model or classifier** is constructed to predict the **classes labels**. These labels are **risky or safe** for loan application data and **yes or no** for marketing data.

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

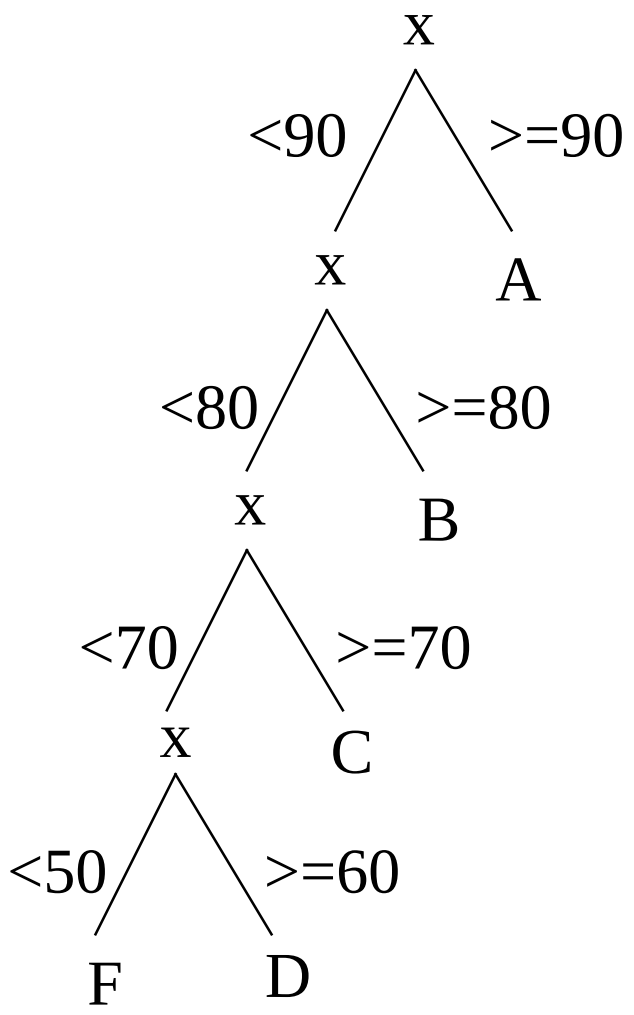
Classification Techniques

1. Decision Tree based Methods
2. Rule-based Methods

Classification Ex: Grading

- If x >= 90 then grade =A.
- If 80<=x<90 then grade =B.
- If 70<=x<80 then grade =C.
- If 60<=x<70 then grade =D.
- If x<50 then grade =F.

Rule-based



Decision Tree based

Rule-Based Classifier

- Classify records by using a collection of "**if…then…**" rules
- Rule: (*Condition*) $\rightarrow$ *y* (consequent)
  - where
    - *Condition* is a conjunctions of attributes
    - *y* is the class label
  - *LHS*: rule condition
  - *RHS*: rule consequent
  - Examples of classification rules:

(Blood Type=Warm) $\wedge$ (Lay Eggs=Yes) $\rightarrow$ Birds

Rule-based Classifier (Example)

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
human	warm	yes	no	no	mammals
python	cold	no	no	no	reptiles
salmon	cold	no	no	yes	fishes
whale	warm	yes	no	yes	mammals
frog	cold	no	no	sometimes	amphibians
komodo	cold	no	no	no	reptiles
bat	warm	yes	yes	no	mammals
pigeon	warm	no	yes	no	birds
cat	warm	yes	no	no	mammals
leopard shark	cold	yes	no	yes	fishes
turtle	cold	no	no	sometimes	reptiles
penguin	warm	no	no	sometimes	birds
porcupine	warm	yes	no	no	mammals
eel	cold	no	no	yes	fishes
salamander	cold	no	no	sometimes	amphibians
gila monster	cold	no	no	no	reptiles
platypus	warm	no	no	no	mammals
owl	warm	no	yes	no	birds
dolphin	warm	yes	no	yes	mammals
eagle	warm	no	yes	no	birds

R1: (Give Birth = no) $\wedge$ (Can Fly = yes) $\rightarrow$ Birds

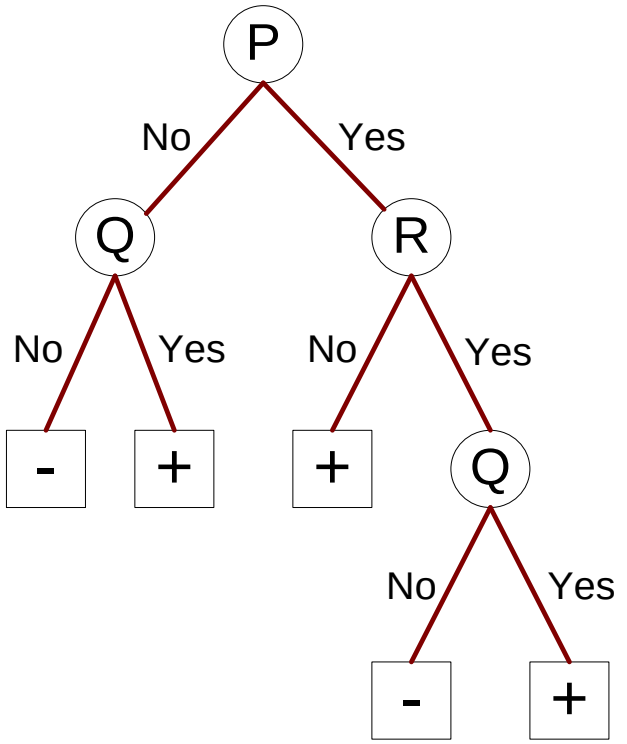
R2: (Give Birth = no) $\wedge$ (Live in Water = yes) $\rightarrow$ Fishes

R3: (Give Birth = yes) $\wedge$ (Blood Type = warm) $\rightarrow$ Mammals

R4: (Give Birth = no) $\wedge$ (Can Fly = no) $\rightarrow$ Reptiles

R5: (Live in Water = sometimes) $\rightarrow$ Amphibians

Rule-Based Classifier with Decision Tree



Model: Decision Tree

Rule Set

r1: (P=No,Q=No) ➡ -
r2: (P=No,Q=Yes) ➡ +
r3: (P=Yes,R=No) ➡ +
r4: (P=Yes,R=Yes,Q=No) ➡ -
r5: (P=Yes,R=Yes,Q=Yes) ➡ +

Classification: Definition

- Given a collection of records (**training set**)
  - Each record contains a set of **attributes**, one of the attributes is the **class**.
- Find a **model** for class attribute as a function of the values of other attributes.
- Goal: **previously unseen** records should be assigned a class as accurately as possible.
- A **test set** is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with **training set** used to build the model and **test set** used to validate it.

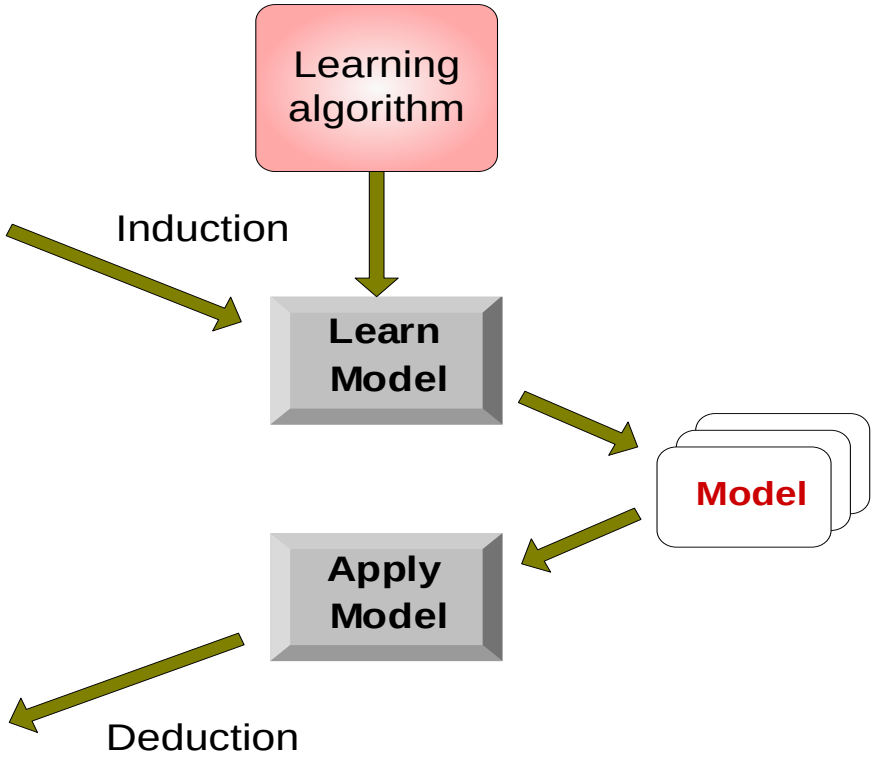
Illustrating Classification Task

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



How Does Classification Works?

With the help of the bank loan application that we have discussed above, let us understand the working of classification. The Data Classification process includes two steps:-

- Building the **Classifier or Model**
- Using **Classifier or Model** for Classification

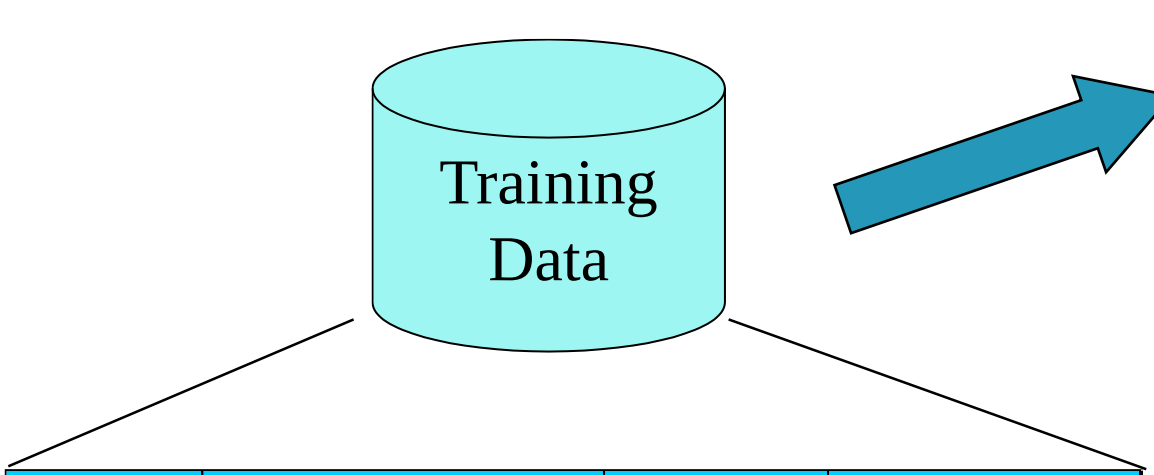
Classification—A Two-Step Process

- **Classifier or Model Construction**: describing a set of predetermined classes:-
  - Each object is assumed to belong to a predefined **class**, as determined by the class label attribute
  - The set of objects used for model construction is **training set**
  - The model is represented as **classification rules or decision trees,**

Classification—A Two-Step Process

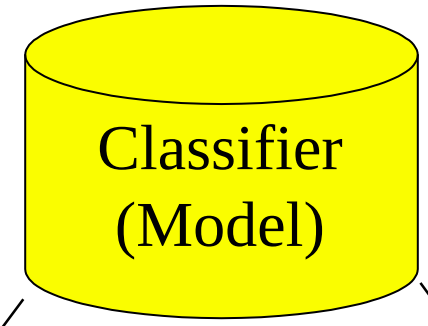
- **Use Classifier or Model for Classification** : for classifying future or unknown objects
  - Estimate accuracy of the model
    - The known class of test sample is compared with the classified result from the model
    - Accuracy rate is the percentage of test set samples that are correctly classified by the model
    - Test set is independent of training set
  - If the accuracy is acceptable, use the model to classify new data

Process (1): Model Construction



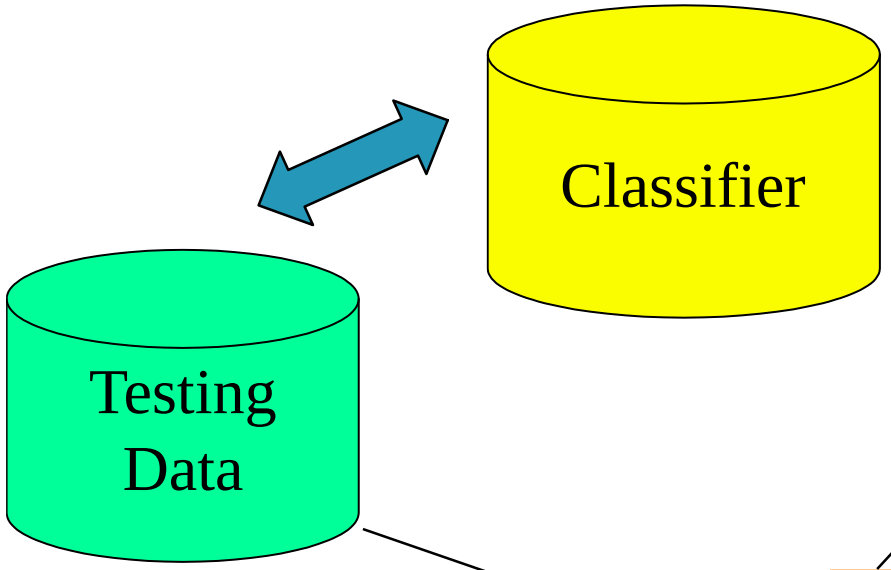
NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Classification
Algorithms



IF rank = ‘professor’
OR years > 6
THEN tenured = ‘yes’

Process (2): Using the Model for Classification



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	?
Merlisa	Associate Prof	7	?
George	Professor	5	?
Joseph	Assistant Prof	7	?

Tenured?

Yes

Thank you