



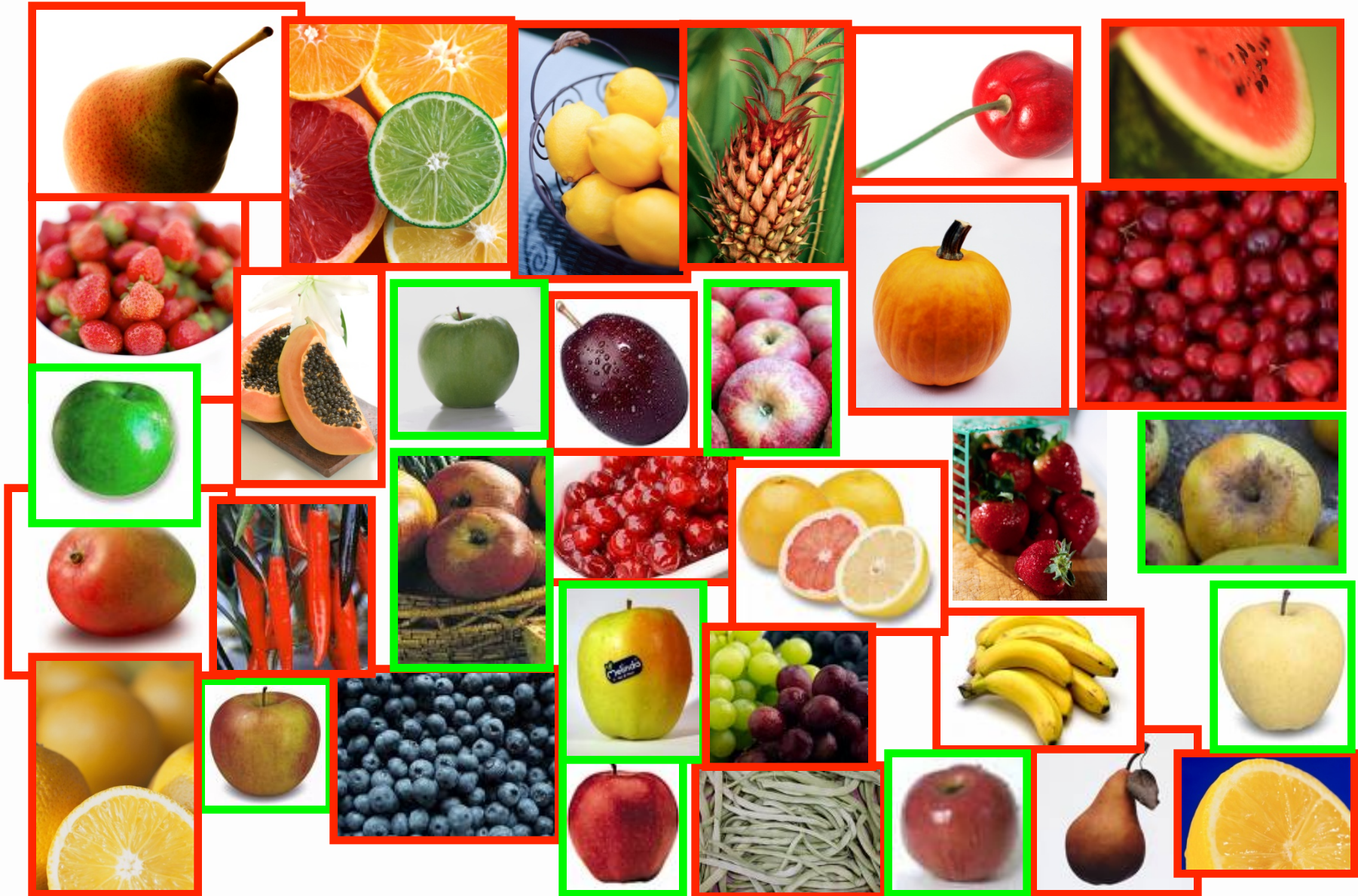
Classification: Introduction

Data Mining and Text Mining (UIC 583 @ Politecnico di Milano)

- Classification and prediction
- Model construction and model usage
- Machine learning view of classification

What is Classification?

What is an apple?



Are these apples?



Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
Young	Myope	No	Reduced	None
Young	Myope	No	Normal	Soft
Young	Myope	Yes	Reduced	None
Young	Myope	Yes	Normal	Hard
Young	Hypermetrope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Young	Hypermetrope	Yes	Reduced	None
Young	Hypermetrope	Yes	Normal	hard
Pre-presbyopic	Myope	No	Reduced	None
Pre-presbyopic	Myope	No	Normal	Soft
Pre-presbyopic	Myope	Yes	Reduced	None
Pre-presbyopic	Myope	Yes	Normal	Hard
Pre-presbyopic	Hypermetrope	No	Reduced	None
Pre-presbyopic	Hypermetrope	No	Normal	Soft
Pre-presbyopic	Hypermetrope	Yes	Reduced	None
Pre-presbyopic	Hypermetrope	Yes	Normal	None
Presbyopic	Myope	No	Reduced	None
Presbyopic	Myope	No	Normal	None
Presbyopic	Myope	Yes	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
Presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Hypermetrope	No	Normal	Soft
Presbyopic	Hypermetrope	Yes	Reduced	None
Presbyopic	Hypermetrope	Yes	Normal	None

A model extracted from the contact lenses data

```
If tear production rate = reduced then recommendation = none
If age = young and astigmatic = no
  and tear production rate = normal then recommendation = soft
If age = pre-presbyopic and astigmatic = no
  and tear production rate = normal then recommendation = soft
If age = presbyopic and spectacle prescription = myope
  and astigmatic = no then recommendation = none
If spectacle prescription = hypermetrope and astigmatic = no
  and tear production rate = normal then recommendation = soft
If spectacle prescription = myope and astigmatic = yes
  and tear production rate = normal then recommendation = hard
If age young and astigmatic = yes
  and tear production rate = normal then recommendation = hard
If age = pre-presbyopic
  and spectacle prescription = hypermetrope
  and astigmatic = yes then recommendation = none
If age = presbyopic and spectacle prescription = hypermetrope
  and astigmatic = yes then recommendation = none
```

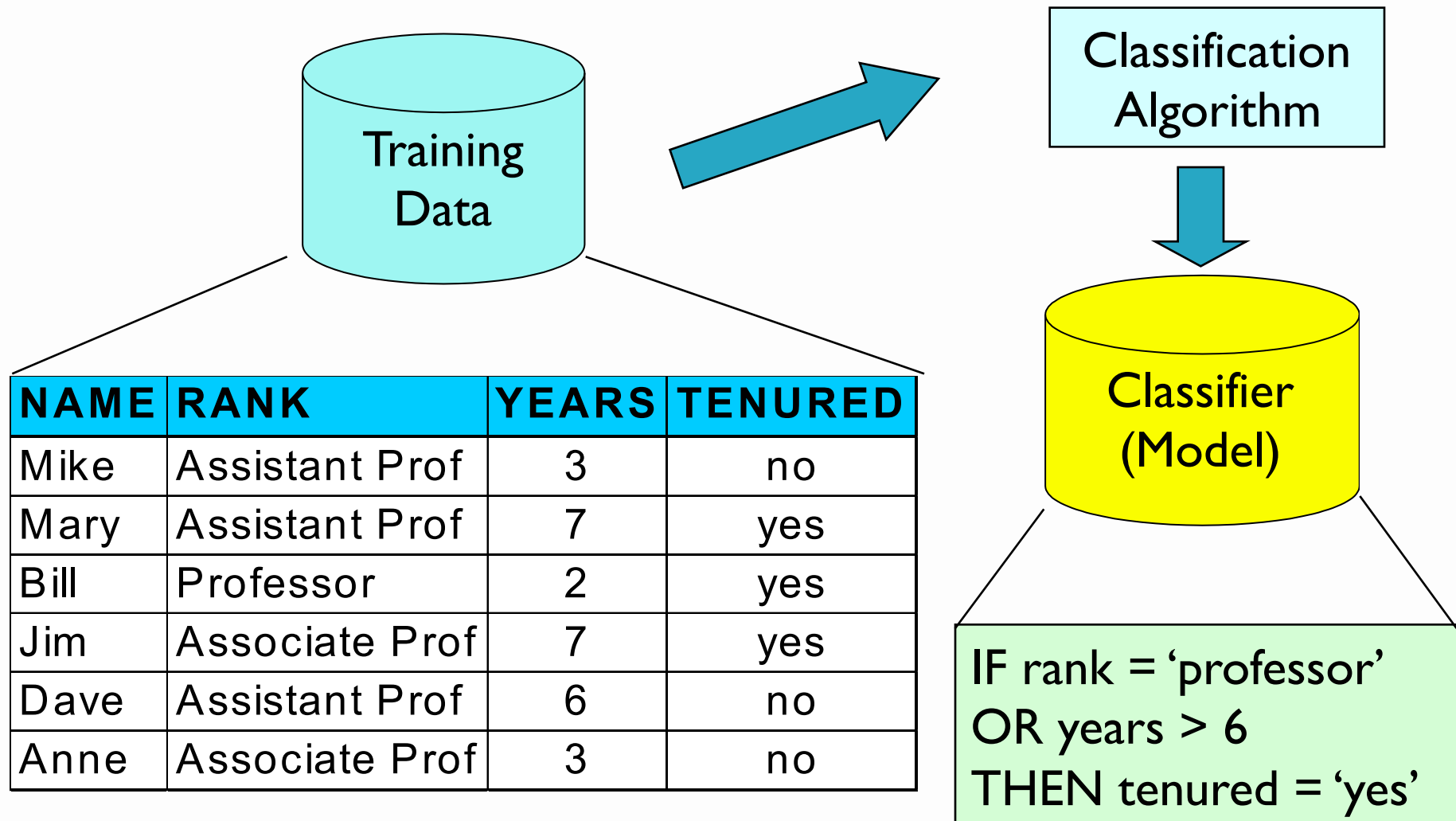
	Cycle time (ns)	Main memory (Kb)		Cache (Kb)	Channels		Performance
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
...							
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

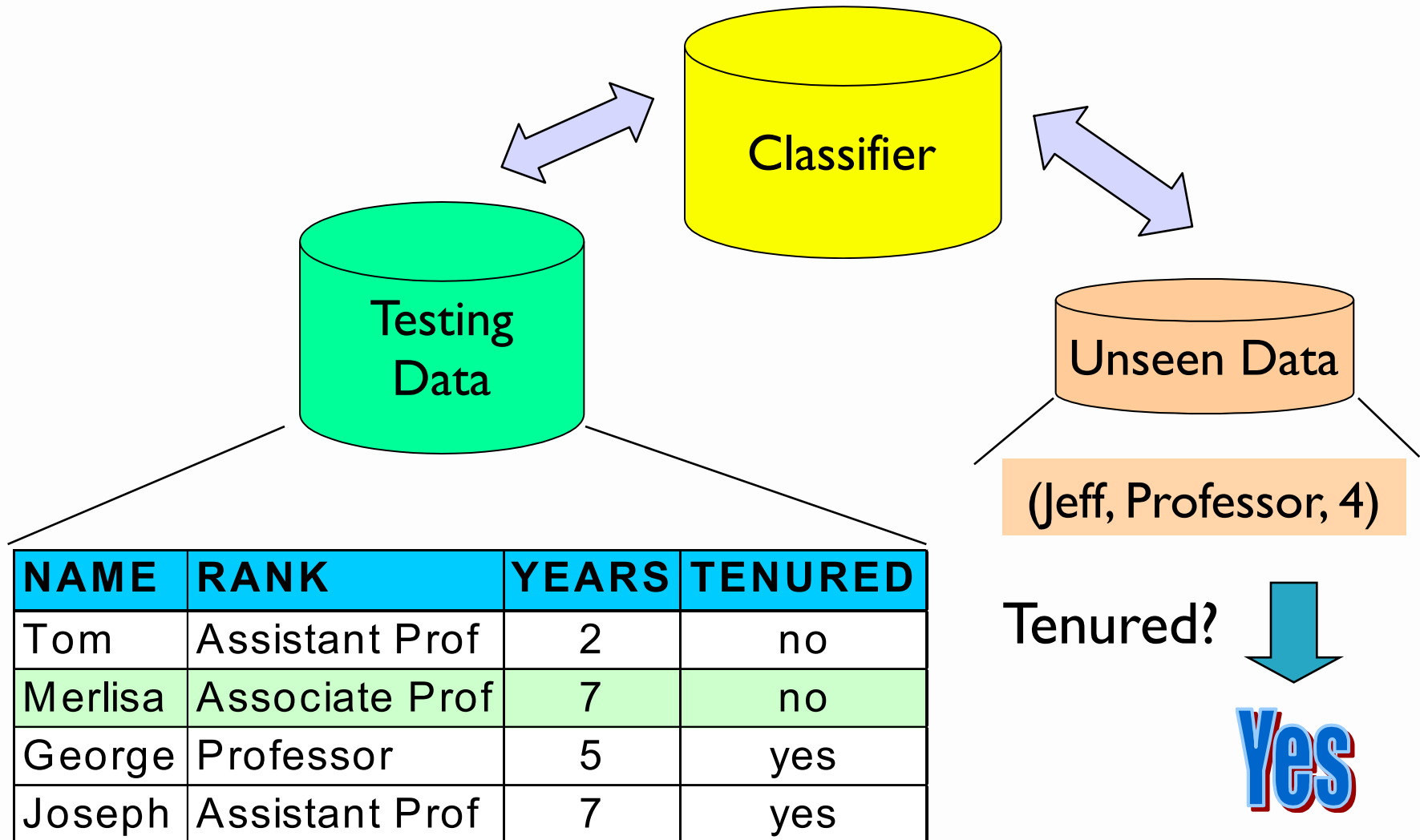
$$\text{PRP} = -55.9 + 0.0489 \text{ MYCT} + 0.0153 \text{ MMIN} + 0.0056 \text{ MMAX} \\ + 0.6410 \text{ CACH} - 0.2700 \text{ CHMIN} + 1.480 \text{ CHMAX}$$

- **Classification**
 - Predicts categorical class labels (discrete or nominal)
 - Classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data
- **Prediction**
 - Models continuous-valued functions, i.e., predicts unknown or missing values
- **Applications**
 - Credit approval
 - Target marketing
 - Medical diagnosis
 - Fraud detection

Model Building and Model Usage

- Classification is a two-step Process
- **Model construction**
 - Given a set of data representing examples of a target concept, build a model to “explain” the concept
- **Model usage**
 - The classification model is used for classifying future or unknown cases
 - Estimate accuracy of the model



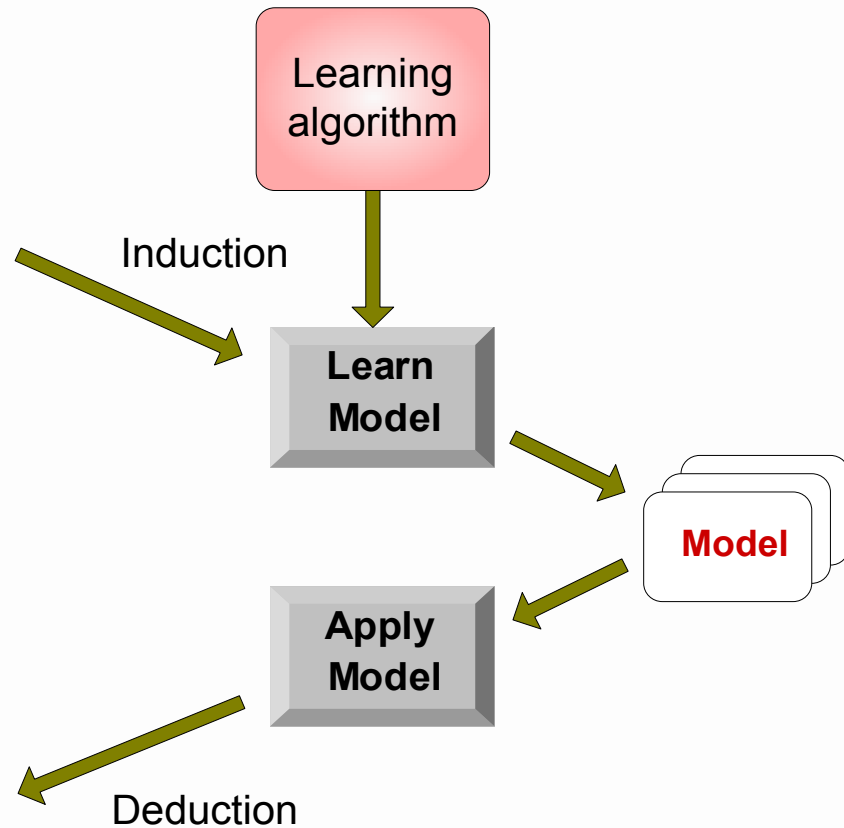


Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

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



- **Accuracy**
 - classifier accuracy: predicting class label
 - predictor accuracy: guessing value of predicted attributes
- **Speed**
 - time to construct the model (training time)
 - time to use the model (classification/prediction time)
- **Robustness**: handling noise and missing values
- **Scalability**: efficiency in disk-resident databases
- **Interpretability**: understanding and insight provided
- **Other measures**, e.g., goodness of rules, such as decision tree size or compactness of classification rules

Example

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

- For each attribute A , For each value VA of the attribute, make a rule as follows:
- count how often each class appears find the most frequent class C_f create a rule when $A=VA$; class attribute value = C_f
- End For-Each
- Calculate the error rate of all rules End For-Each
- Chose the rule with the smallest error rate

The Weather Dataset: Building the Model

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

- Write one rule like “if $A=v_1$ then X , else if $A=v_2$ then Y , ...” to predict whether the player is going to play or not
- A is an attribute; v_i are attribute values; X, Y are class labels

- Write one rule like “if $A=v1$ then X , else if $A=v2$ then Y , ...” to predict whether the player is going to play or not

The Weather Dataset: Testing the Model

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Rainy	Mild	High	False	Yes
Sunny	Mild	High	False	No
Rainy	Mild	Normal	False	Yes

DMTM2011-Example01-OneRule.xml - RapidMiner@pc124-25.elet.polimi.it

File Edit Process Tools View Help

Overview Process XML

Process

Parameters

Read ARFF (3) (Read ARFF)

data file: .nominal.arff

decimal character: .

grouped digits

2 hidden expert parameters

Help Comment

Read ARFF

Synopsis
This operator can read arff files.

Description
This operator can read ARFF files know...

Problems Log

8 potential problems

Message	Fixes	Location
The attribute 'play' is missing in the input example set.	Change value of parameter "...	Set Role.example se...
Input example set must have special attribute 'label'.	Select an attribute whose rol...	OneRule (My Testset...
The attribute 'play' is missing in the input example set.	Change value of parameter "...	Set Role (2).exampl...

- if outlook = sunny then no (3 / 2)
if outlook = overcast then yes (0 / 4)
if outlook = rainy then yes (2 / 3)

correct: 10 out of 14 training examples

- if outlook = sunny then yes (1 / 2)
if outlook = overcast then yes (0 / 4)
if outlook = rainy then no (2 / 1)

correct: 8 out of 10 training examples

The Machine Learning Perspective

- Classification algorithms are methods of supervised Learning
- The experience E consists of a set of examples of a target concept that have been prepared by a **supervisor**
- The task T consists of finding an hypothesis that accurately explains the target concept
- The performance P depends on how accurately the hypothesis h explains the examples in E

- Let us define the problem domain as the set of instance X (for instance, X contains different different fruits)
- We define a concept over X as a function c which maps elements of X into a range D or $c: X \rightarrow D$
- The range D represents the type of concept analyzed
- For instance, $c: X \rightarrow \{\text{isApple}, \text{notAnApple}\}$

- Experience E is a set of $\langle x, d \rangle$ pairs, with $x \in X$ and $d \in D$.
- The task T consists of finding an hypothesis h to explain E :
- $\forall x \in X \ h(x) = c(x)$
- The set H of all the possible hypotheses h that can be used to explain c it is called the hypothesis space
- The goodness of an hypothesis h can be evaluated as the percentage of examples that are correctly explained by h

$$P(h) = |\{x \mid x \in X \text{ e } h(x) = c(x)\}| / |X|$$

- Concept Learning
when $D = \{0, 1\}$
- Supervised classification
when D consists of a finite number of labels
- Prediction
when D is a subset of \mathbb{R}^n

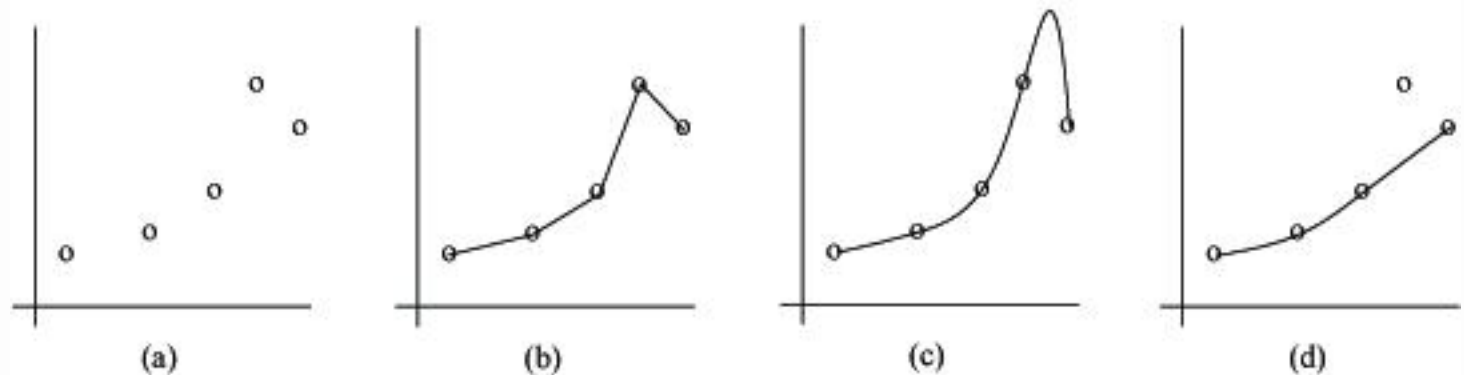
The Machine Learning Perspective on Classification

- Supervised learning algorithms, given the examples in E , search the hypotheses space H for the hypothesis h that best explains the examples in E
- Learning is viewed as a search in the hypotheses space

- The type of hypothesis required influences the search algorithm
- The more complex the representation the more complex the search algorithm
- Many algorithms assume that it is possible to define a partial ordering over the hypothesis space
- The hypothesis space can be searched using either a general to specific or a specific-to-general strategy

- **General to Specific**
 - Start with the most general hypothesis and then go on through specialization steps
- **Specific to General**
 - Start with the set of the most specific hypothesis and then go on through generalization steps

- Set of assumptions that together with the training data deductively justify the classification assigned by the learner to future instances
- There can be a number of hypotheses consistent with training data
- Each learning algorithm has an inductive bias that imposes a preference



- Syntactic Bias
 - Depends on the language used to represent hypotheses
- Semantic Bias
 - Depends on the heuristics used to filter hypotheses
- Preference Bias
 - Depends on the ability to rank and compare hypotheses
- Restriction Bias
 - Depends on the ability to restrict the search space

Why Looking for h ?

- Any hypothesis (h) found to approximate the target function (c) over a sufficiently large set of training examples will also approximate the target function (c) well over other unobserved examples.
- Training and Testing
 - Training: the hypothesis h is developed to explain the examples in E_{Train}
 - Testing: the hypothesis h is evaluated (verified) with respect to the previously unseen examples in E_{Test}
- The underlying hypothesis is that if h explains E_{Train} then it can also be used to explain other unseen examples in E_{Test} (not previously used to develop h)

- **Generalization**
 - When h explains “well” both E_{Train} and E_{Test} we say that h is general and that the method used to develop h has adequately generalized
- **Overfitting**
 - When h explains E_{Train} but not E_{Test} we say that the method used to develop h has overfitted
 - We have overfitting when the hypothesis h explains E_{Train} too accurately so that h is not general enough to be applied outside E_{Train}

What are the general issues for classification in Machine Learning?

- Type of training experience
 - Direct or indirect?
 - Supervised or not?
- Type of target function and performance
- Type of search algorithm
- Type of representation of the solution
- Type of Inductive bias

Summary

- Classification is a two-step process involving the building, the testing, and the usage of the classification model
- In Machine Learning, classification is viewed as an instance of supervised learning
- The focus is on the search process aimed at finding the classifier (the hypothesis) that best explains the data
- Major issues for Machine Learning include:
 - The type of input experience
 - The search algorithm
 - The inductive biases