



Classification: Introduction

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

- ❑ What is classification?
- ❑ What data?
- ❑ The classification process
- ❑ A Machine Learning perspective
- ❑ Training, testing, generalization and overfitting

What is classification?

Are these apples?



- ❑ **Unsupervised learning** (clustering)
 - ▶ The class labels of training data is unknown
 - ▶ Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data
- ❑ **Supervised learning** (classification)
 - ▶ Supervision: The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations
 - ▶ New data is classified based on the training set

What data?

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
Young	Myope	No	Reduced	None
Young	Myope	No	Normal	Soft
Young	Myope ^a	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
```

- 209 different computer configurations

	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

- A model to predict the performance

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

Classification process

□ 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

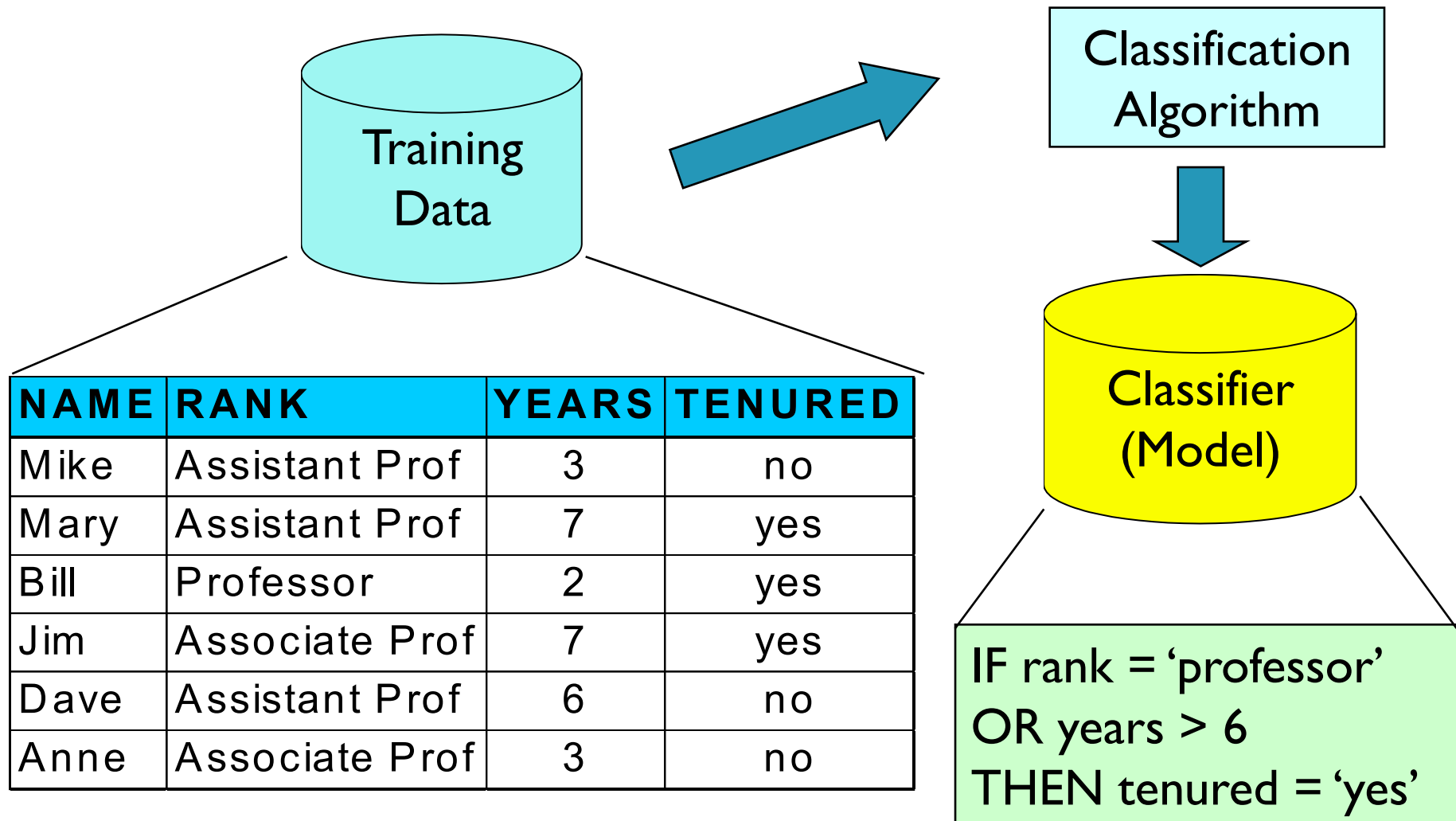
□ Applications

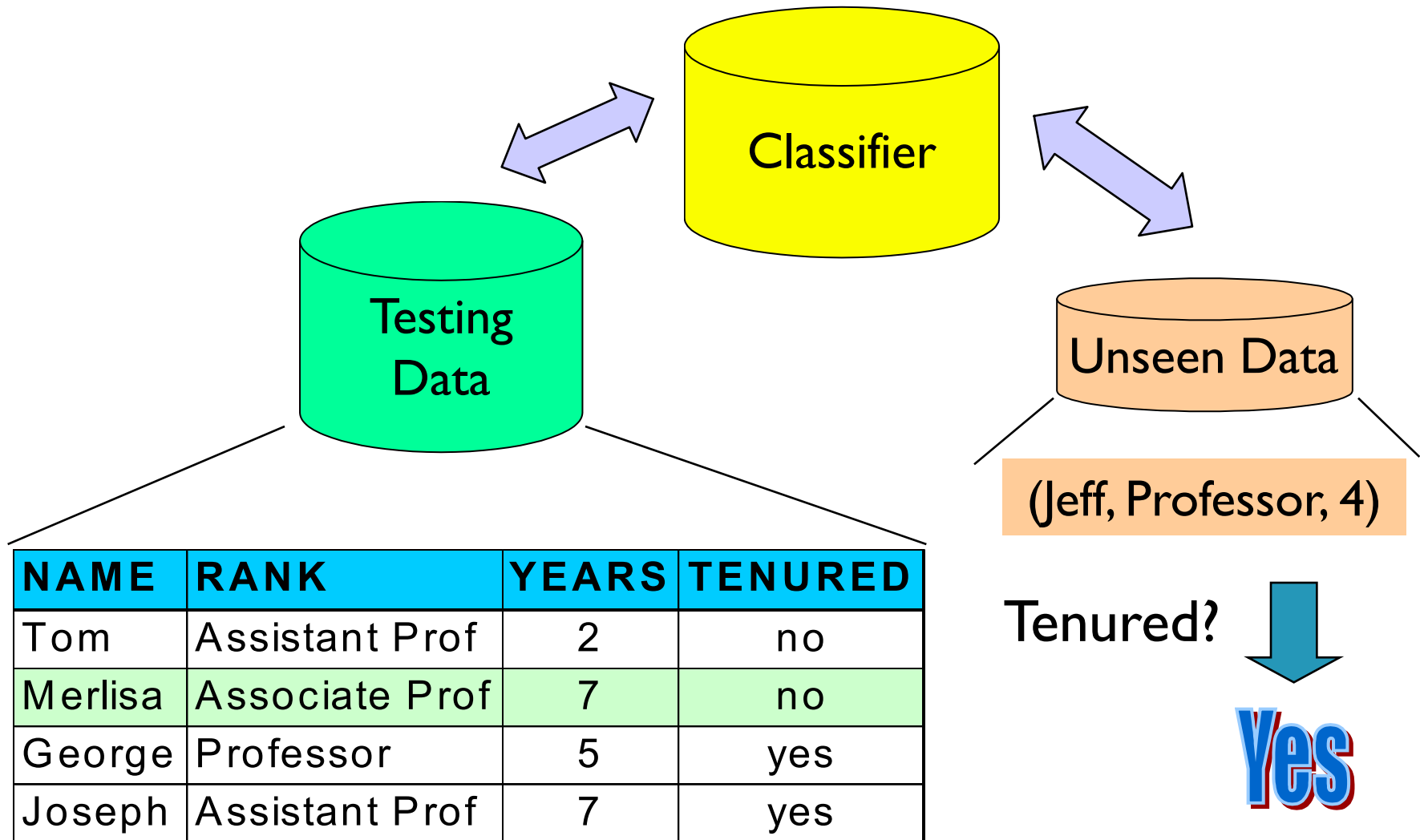
- ▶ Credit approval
- ▶ Target marketing
- ▶ Medical diagnosis
- ▶ Fraud detection

- ❑ 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





❑ 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

Machine Learning

- ❑ Classification algorithms are methods of **supervised Learning**

- ❑ In 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 fruits
- ❑ We define a concept over X as a function c which maps elements of X into a range D

$$c: X \rightarrow D$$

- ❑ The range D represents the type of concept that is going to be analyzed
- ❑ For instance, $c: X \rightarrow \{\text{apple, not_an_apple}\}$

- ❑ 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 :

$$\forall x \in X \quad h(x) = c(x)$$

- ❑ 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

- ❑ 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

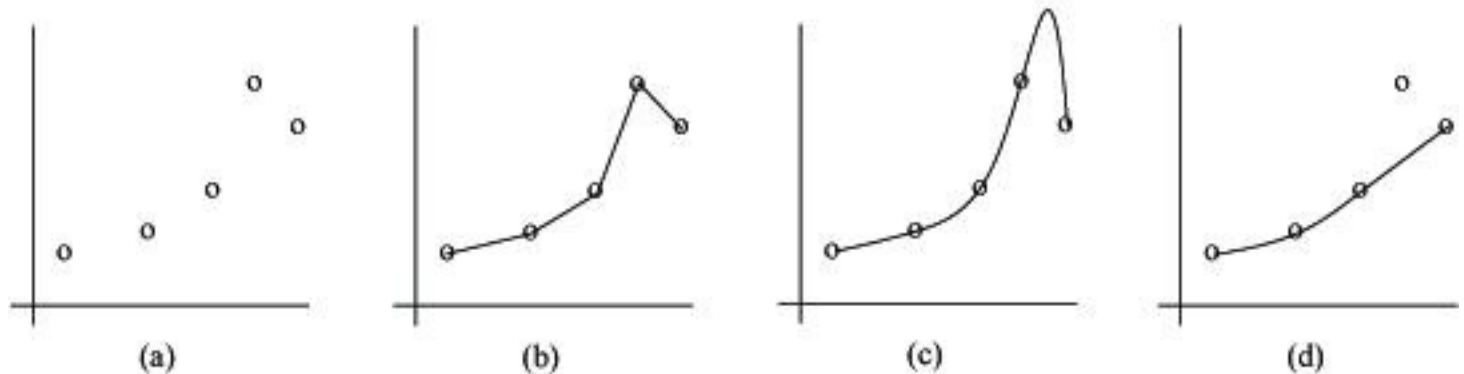
□ **General to Specific**

- ▶ 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 on the space of all possible hypotheses



- ❑ 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

- **Inductive Learning Hypothesis**: 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

- ▶ 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}

□ Generalization and Overfitting

- ▶ 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
- ▶ 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
- ❑ Major issues for Data Mining include:
 - ▶ The type of input data
 - ▶ The representation used for the model
 - ▶ The generalization performance on unseen data
- ❑ 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