

# Chapter VI: Classification

## Knowledge Discovery in Databases

Luciano Melodia M.A.

Evolutionary Data Management, Friedrich-Alexander University Erlangen-Nürnberg

Summer semester 2021



## Chapter VI: Classification

### **Classification: basic concepts.**

Decision-tree induction.

Bayes classification methods.

Rule-based classification.

Model evaluation and selection.

Techniques to improve classification accuracy: ensemble methods.

Summary.

# Supervised vs. unsupervised learning

## 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 a **model** created from the training data.

## Unsupervised learning (clustering).

The class labels of training data are unknown.

Or rather, there are no training data.

Given a set of measurements, observations, etc., the goal is to find classes or clusters in the data.

See next chapter.

## Prediction problems: classification vs. numerical prediction

### Classification:

Predicts **categorical class labels** (discrete, nominal).

Constructs a model based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data.

### Numerical prediction:

Models **continuous-valued functions**.

I.e. predicts missing or unknown (future) values.

### Typical applications of classification:

Credit/loan approval: Will it be paid back?

Medical diagnosis: Is a tumor cancerous or benign?

Fraud detection: Is a transaction fraudulent or not?

Web-page categorization: Which category is it?

## Classification – a two-step process

### Model construction: describing a set of predetermined classes:

Each tuple/sample is assumed to belong to a predefined class, as determined by the **class-label attribute**.

The set of tuples used for model construction is the **training set**.

The **model** is represented as classification rules, decision trees, or mathematical formulae.

### Model usage, for classifying future or unknown objects:

Estimate **accuracy** of the model:

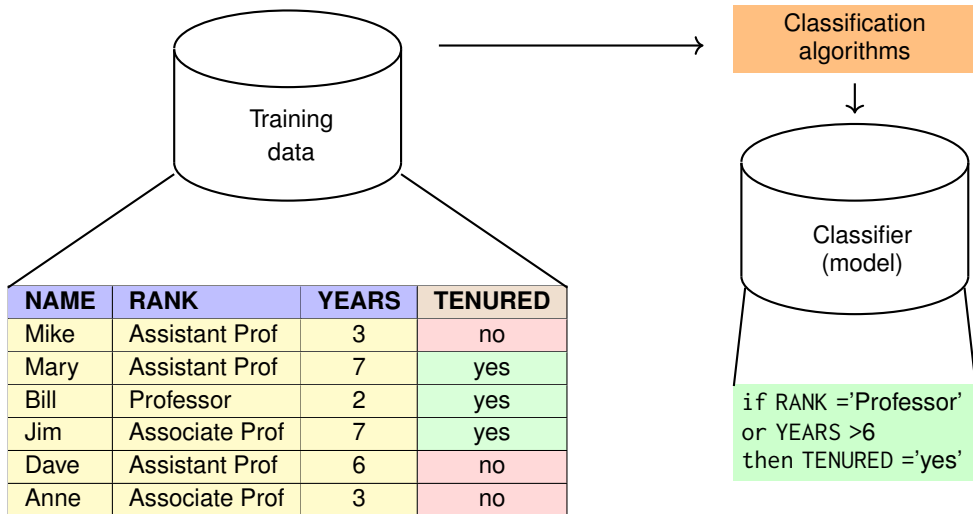
The known label of **test samples** is compared with the result from the model.

**Accuracy rate** is the percentage of test-set samples that are correctly classified by the model.


Test set is independent of training set (otherwise overfitting).

If the accuracy is acceptable, **use the model** to classify data tuples whose class labels are not known.

## Classification – a two-step process



Thank you for your attention.  
**Any questions about the sixth chapter?**

Ask them now, or again, drop me a line:  
 `luciano.melodia@fau.de`.