

# Inf-KDDM: Knowledge Discovery and Data Mining

Winter Term 2020/21

## Lecture 5: Classification

Lectures: Prof. Dr. Matthias Renz

Exercises: Steffen Strohm

---

## Outline

---

- Classification basics
- Decision tree classifiers
- Overfitting
- Lazy vs Eager Learners
- k-Nearest Neighbors (or learning from your neighbors)
- Evaluation of classifiers

# The classification problem

- Given:

- a dataset of instances  $D=\{t_1, t_2, \dots, t_n\}$  and
- a set of classes  $C=\{c_1, \dots, c_k\}$

classification is the task of learning a *target function*/ mapping  $f:D \rightarrow C$  that assigns each  $t_i$  to a  $c_j$ .

- The mapping or target function is known informally as a *classification model*.

ID	Age	Car type	Risk
1	23	Family	high
2	17	Sport	high
3	43	Sport	high
4	68	Family	low
5	32	Truck	low

*Predictor attributes:* Age, Car type

*Class attribute:* risk={high, low}

---

# The classification problem

---

## Classification vs Prediction

- Classification
  - predicts categorical (discrete, unordered) class labels
  - Constructs a model (classifier) based on a training set
  - Uses this model to predict the class label for **new unknown-class** instances
- Prediction
  - is similar, but may be viewed as having infinite number of classes (cf. Regression)

## A simple classifier

---

ID	Age	Car type	Risk
1	23	Family	high
2	17	Sport	high
3	43	Sport	high
4	68	Family	low
5	32	Truck	low

A simple classifier:

- if Age > 50 then Risk= low;
- if Age  $\leq$  50 and Car type =Truck then Risk=low;
- if Age  $\leq$  50 and Car type  $\neq$  Truck then Risk = high.

# Applications

---

- Credit approval
  - Classify bank loan applications as e.g. safe or risky.
- Fraud detection
  - e.g., in credit cards
- Churn prediction
  - E.g., in telecommunication companies
- Target marketing
  - Is the customer a potential buyer for a new computer?
- Medical diagnosis
- Character recognition
- ...

# Classification techniques

---

- Typical classification approach:
  - Create specific model by evaluating training data (or using domain experts' knowledge).
    - Assess the quality of the model
  - Apply model developed to new data.
- Classes must be predefined!!!
- Many techniques
  - Decision trees
  - Naïve Bayes
  - kNN
  - Neural Networks
  - Support Vector Machines
  - ....

# Classification technique (detailed)

- **Model construction:** describing a set of predetermined classes
  - The set of tuples used for model construction is the **training set**
  - Each tuple/sample is assumed to belong to a predefined class, as determined by the **class label**
  - The model is represented as classification rules, decision trees, or mathematical formula
  
- **Model evaluation:** estimate accuracy of the model
  - The set of tuples used for model evaluation is the **test set**
  - The class label of each tuple/sample in the test set is known in advance.
  - The known label 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, otherwise over-fitting will occur
  
- **Model usage:** for classifying future or unknown objects
  - If the accuracy is acceptable, use the model to classify data tuples whose class labels are not known.

predefined class values  
↓

Class attribute: tenured={yes, no}

Training set

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

known class label attribute ↗

Test set

NAME	RANK	YEARS	TENURED	PREDICTED
Maria	Assistant Prof	3	no	no
John	Associate Prof	7	yes	no
Franz	Professor	3	yes	yes

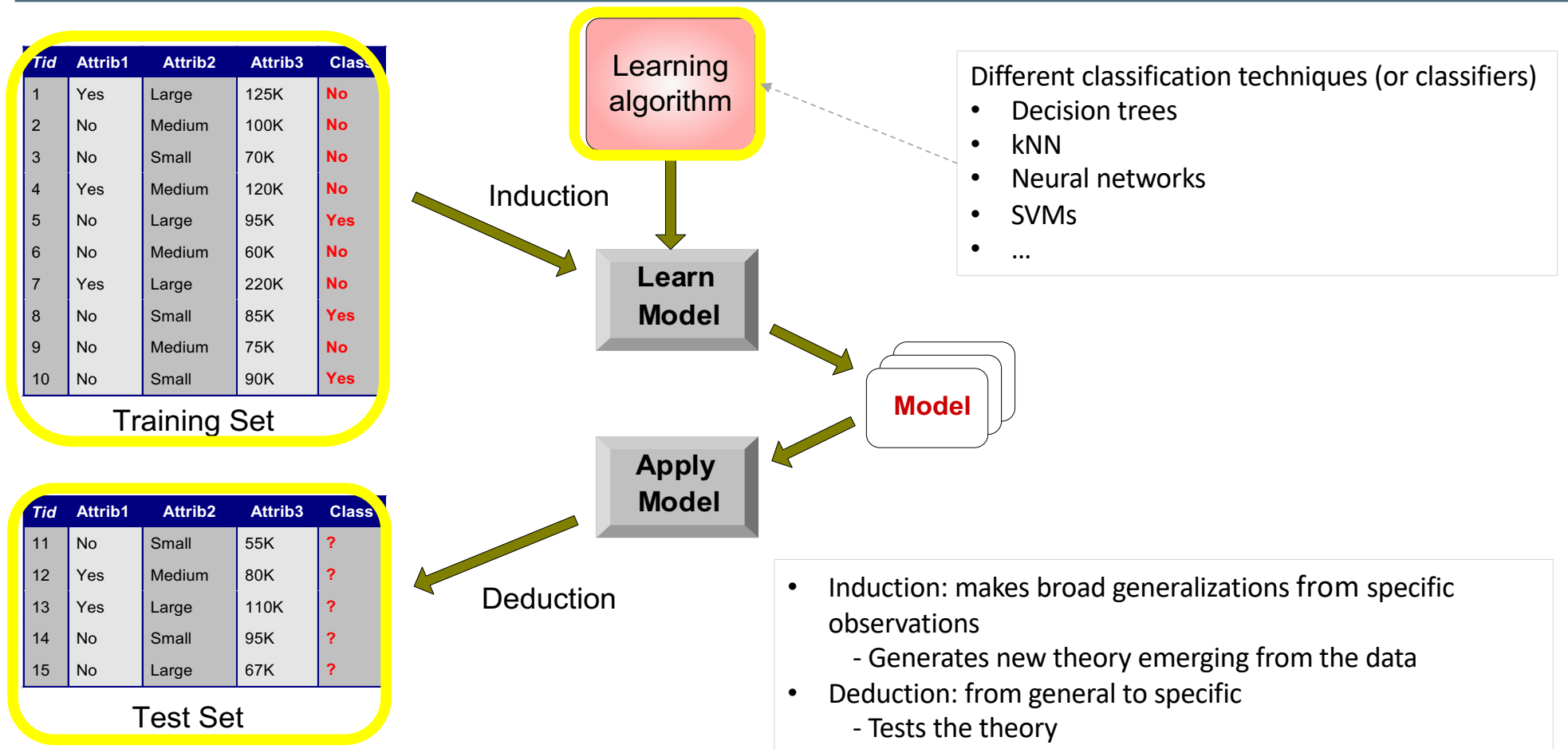
known class label attribute ↗  
predicted class value by the model ↗

NAME	RANK	YEARS	TENURED	PREDICTED
Jeff	Professor	4	?	yes
Patrick	Associate Prof	8	?	yes
Maria	Associate Prof	2	?	no

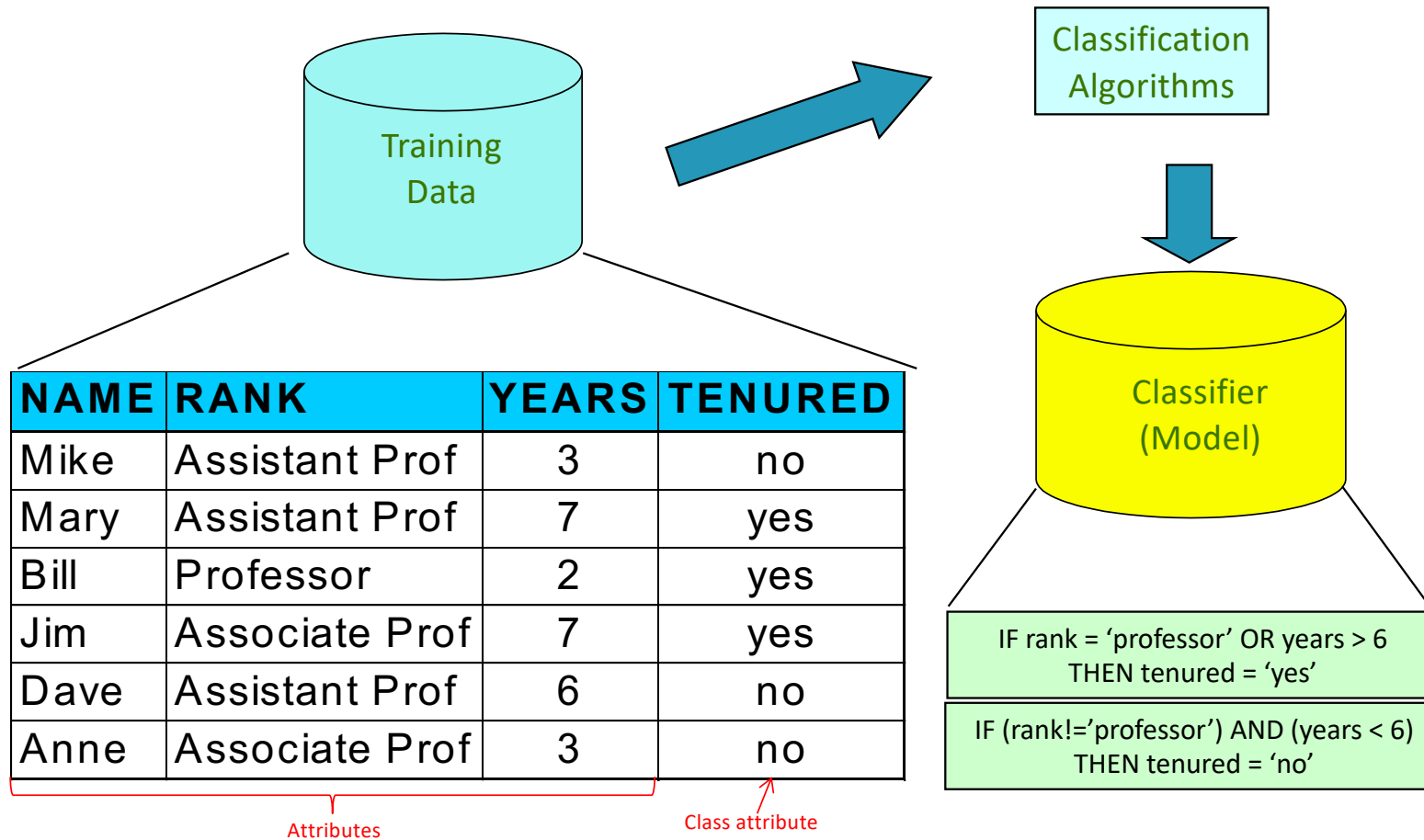
unknown class label attribute ↗  
predicted class value by the model ↗



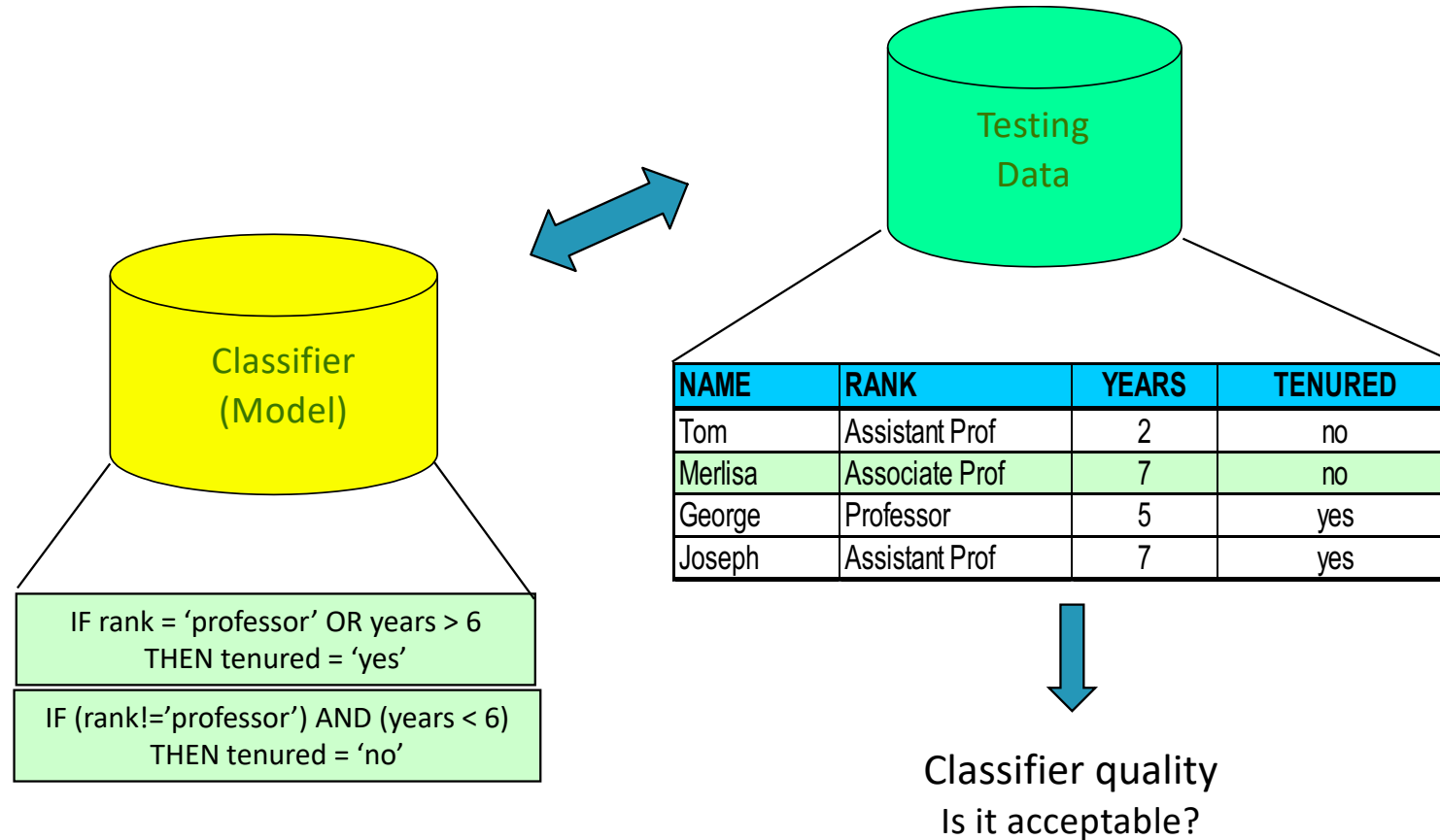
# General approach for building a classification model



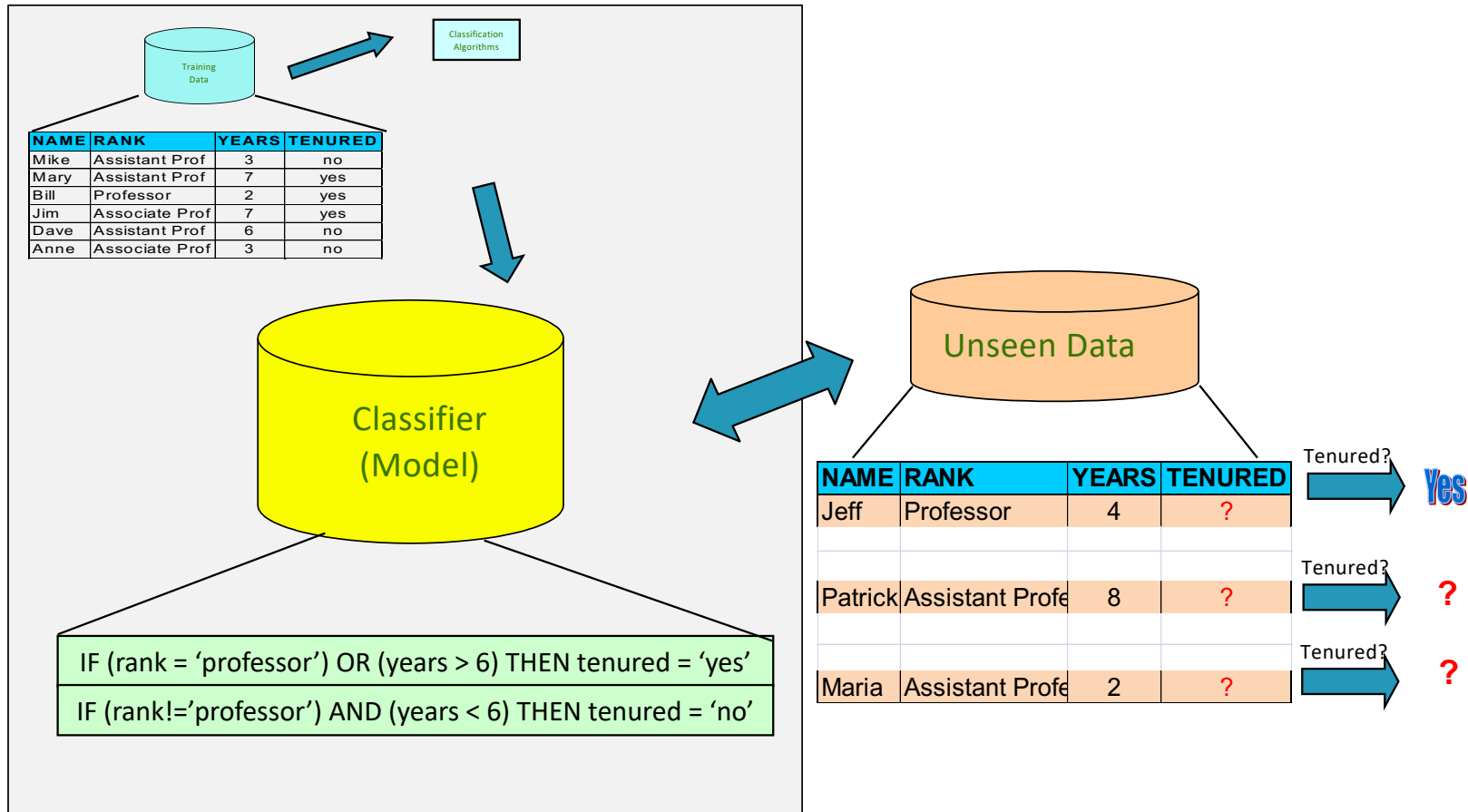
# Model construction



# Model evaluation



# Model usage for prediction



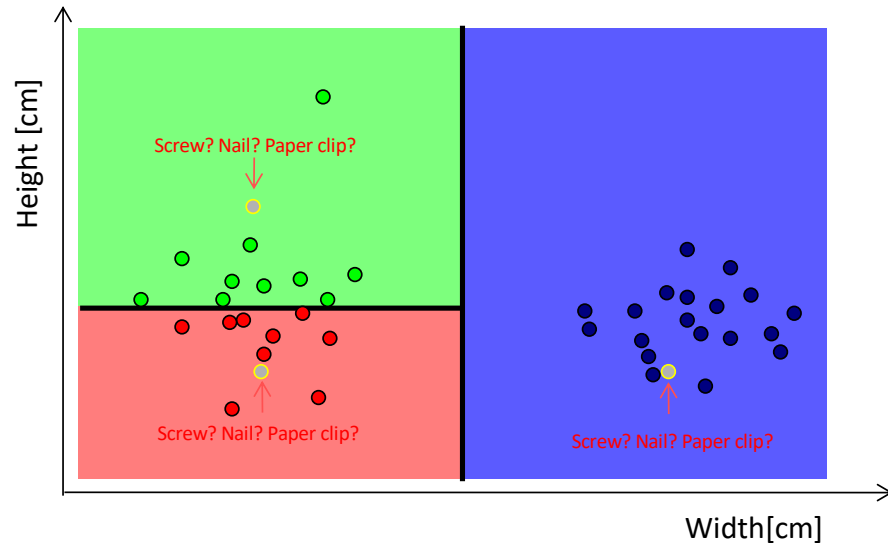
---

## A supervised learning task

---

- Classification is a **supervised** learning task
  - 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
  
- Clustering is an **unsupervised** learning task
  - The class labels of training data is unknown
  - Given a set of measurements, observations, etc., the goal is to group the data into groups of similar data (clusters)

# Supervised learning example



**Classification model**

- Screw 
- Nails 
- Paper clips 

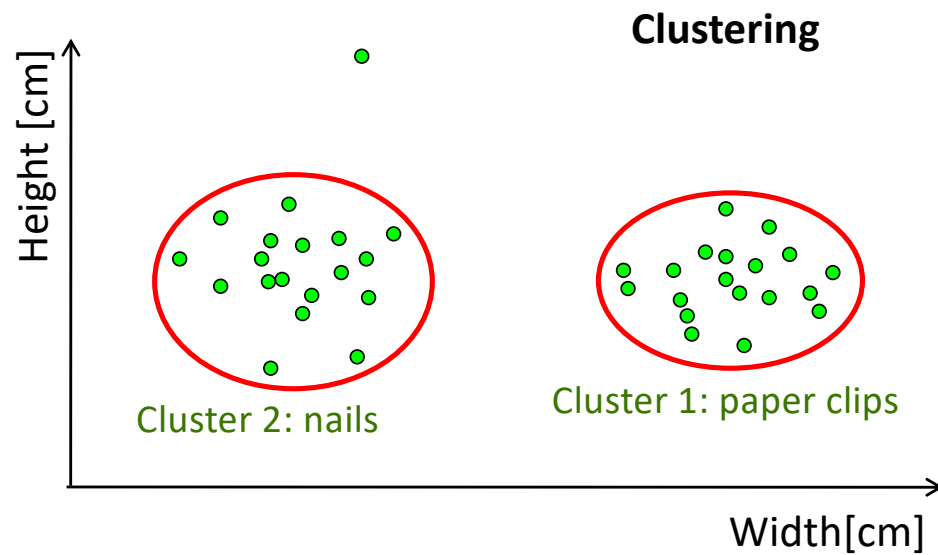
- New object (unknown class)

Question:

What is the class of a new object???

Is it a screw, a nail or a paper clip?

## Unsupervised learning example



Question:

Is there any structure in data (based on their characteristics, i.e., width, height)?

# Classification techniques

- Statistical methods

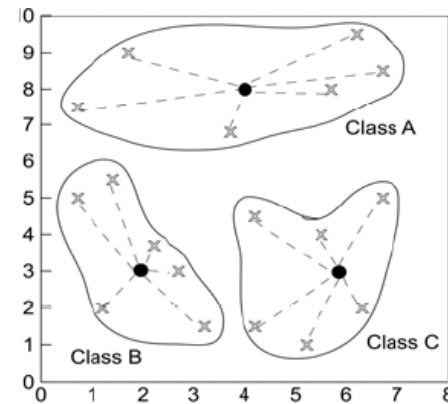
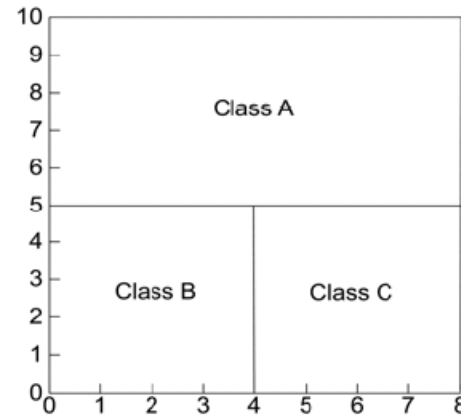
- Bayesian classifiers etc

- Partitioning methods

- Decision trees etc

- Similarity based methods

- K-Nearest Neighbors etc





## Outline

---

- Classification basics
- Decision tree classifiers
- Overfitting
- Lazy vs Eager Learners
- k-Nearest Neighbors (or learning from your neighbors)
- Evaluation of classifiers