

# Machine Learning: Introduction to Classification

by Elena Battini Sönmez  
İstanbul Bilgi University

# Classification:

Classification is a supervised learning

- It uses the **training set** to create a model
- It uses the **test set** to test the accuracy of the model
- If **accuracy** is low, it regenerates the model, it consider different features, ...

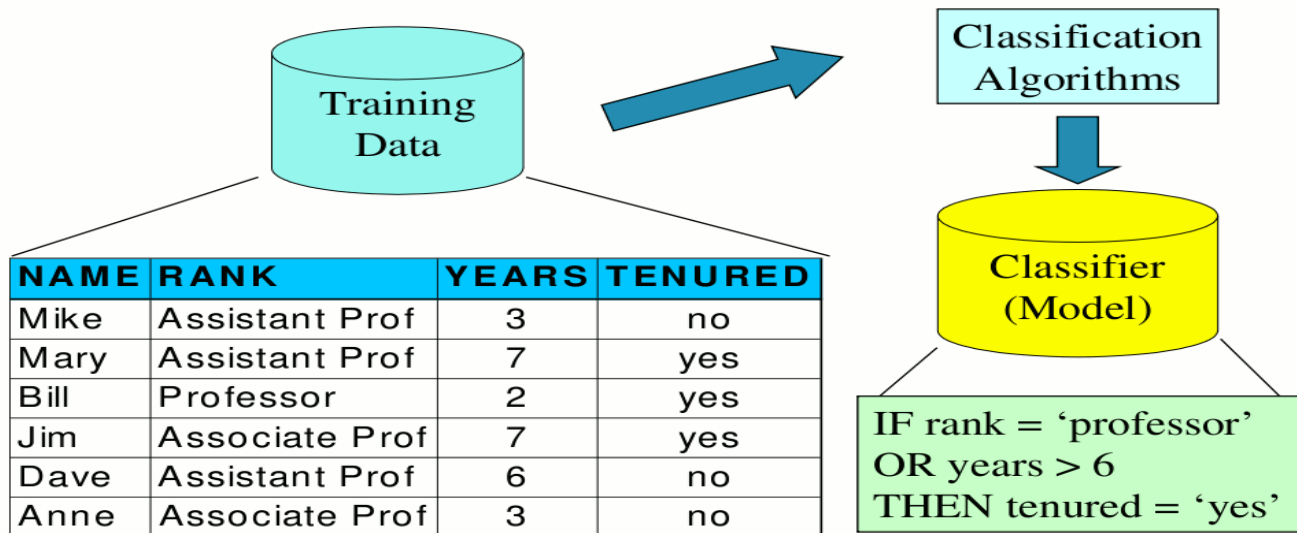
# Training set example:

F1: color	F2: shape	F3: ?	F4: Class Label
red	round	6	apple
red	round	4	cherry
green	round	6	apple
yellow	long	?	banana

- N-class model = {apple, cherry, banana}
- Attributes are features. Which attribute is F3?
- Test sample=(red, round, 6, ?)

# Model Construction:

Taken from: Jiawei Han and Micheline Kamber “Data Mining and Concepts”



The model can be represented as classification rules, decision trees or mathematical formula

# Observations:

- Example of 2-class model, class label={yes,no}
- Positive vs negative samples
- Looking at the model, assign a label to the following (test) samples:

NAME	RANK	YEARS	TENURED
Tom	Ass. Prof	3	?
Mary	Professor	2	?

# Performance measurement :

## Accuracy:

$$\frac{\text{number of correct classified sample}}{\text{total number samples}}$$

## Confusion matrix:

- Square matrix of dimension  $c \times c$ , where  $c$  = total number of class
- Initial values are all zeros:  $CM(i,j)=0$ , for all  $i,j=1,\dots,c$
- Rows label the true class
- Columns label the predicted class
- When a test sample of class “i” is assigned to class “j”,  
 $CM(i,j)=CM(i,j)+1$

# Accuracy (1/2):

Goal: we want to measure the successful rate of our classification system

**Approach1:** Use all labeled samples for training, build the model and test it using the same data

*Very optimistic result! Not working 😞*

**Approach2:** Take the entire data set, cut it in half and use half for training and half for testing

*Potential error in estimate the real classification rate, because the 2 subsets may be very different.*

# Accuracy (2/2):

## Approach3: 10-fold cross validation

- Take all samples and divide them in 10 sub-sets
- At every trial, use 9 subsets for training and 1 for testing
- Each run will result in a particular classification rate
- Get the average of the 10 runs





# Confusion matrix (1/2):

## 2-class model:

- True positive (TP)= hit
- True negative (TN)= correct rejection
- False negative= mis
- False positive= false alarm

$$\text{Accuracy (or performance)} = \frac{(TP+TN)}{(P+N)}$$

Predicted  
class

True class

	Positive	Negative
Yes	True positive	False negative
No	False positive	True negative

Total:

P

N

# Confusion matrix (2/2):

## N-class model

Ex={apple, cherry, banana}  
class\_no=3

True class

A classifier with 100% accuracy produces a CM with all out of diagonal elements equal to “0”

Ex: If-when the test set stores 2 samples belonging to the class “apples”, 3 ... “cherry, and 2 samples ... “banana”

Predicted class

	apple	cherry	banana
apple	2	0	0
cherry	0	3	0
banana	0	0	2