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, ...

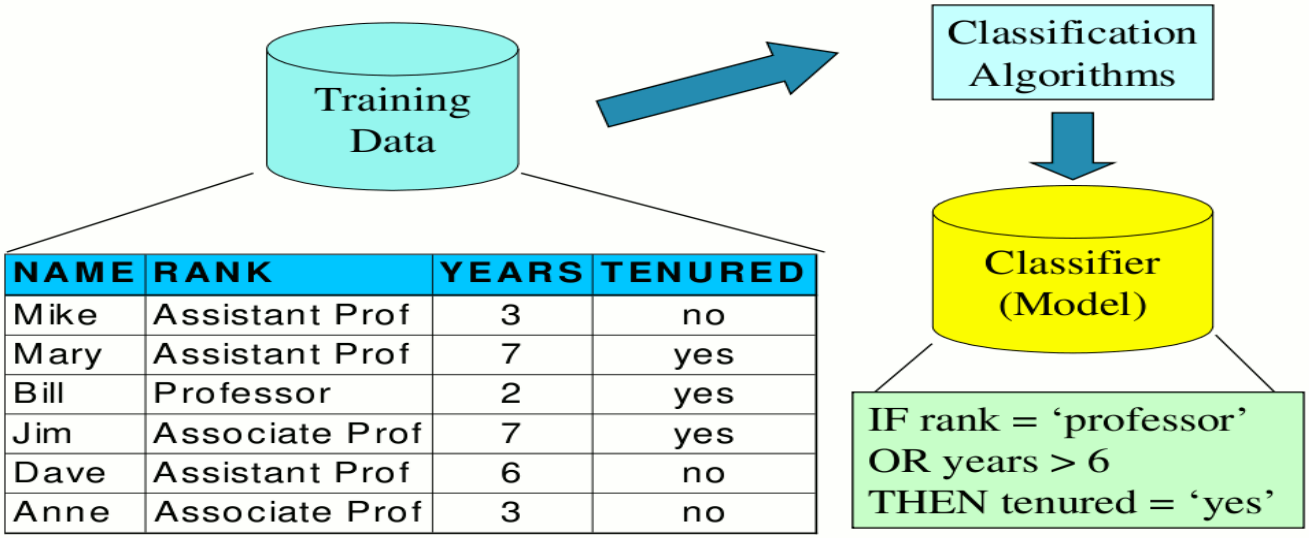
Training set example:

F1: color	F2: shape	F3: ?	F4: Class Label
red	round	6	apple
red	round	4	cherry
green	round	7	apple
yellow	long	?	banana

- N-class model = {apple, cherry, banana}
- Attributes are features. Which attribute is F3?
- Test sample=(red, round, 7, ?)

Model Construction:

Taken from: Jiawei Han and Micheline Kamber "Data Mining and Concepts"



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

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

Observations:

- Example of 2-class model, class label={yes, no}
- Positive versus 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×c, where c= total number of class
- Initial values are all zeros: CM(i, j)=0, for all i, j=1,...,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

How to divide all samples into training and test sets (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☹

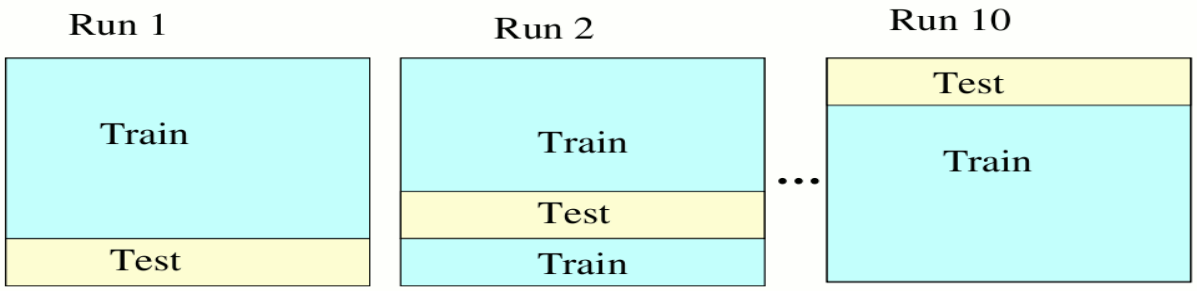
Approach2: Take the entire data set, cut it in half and use half for training and half for testing

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

How to divide all samples into training and test sets (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 (1/3):

2-class model:

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

Accuracy (or performance) = (TP+TN)/(P+N)

	True class: Positive	True class: Negative
Predicted class: Yes	True positive	False positive
Predicted class: No	False negative	True negative
Total:	P	N

Confusion matrix (2/3):

N-class model

Ex={apple, cherry, banana}

class_no=3

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”

True class \ Predicted class	apple	cherry	banana
apple	2	0	0
cherry	0	3	0
banana	0	0	2

Confusion matrix (3/3):

N-class model

Ex={apple, cherry, banana}

class_no=3

Which is the performance of this classifier?

Which are the confused classes?

Predicted class

True class

	apple	cherry	banana
apple	2	4	0
cherry	6	3	0
banana	0	0	2