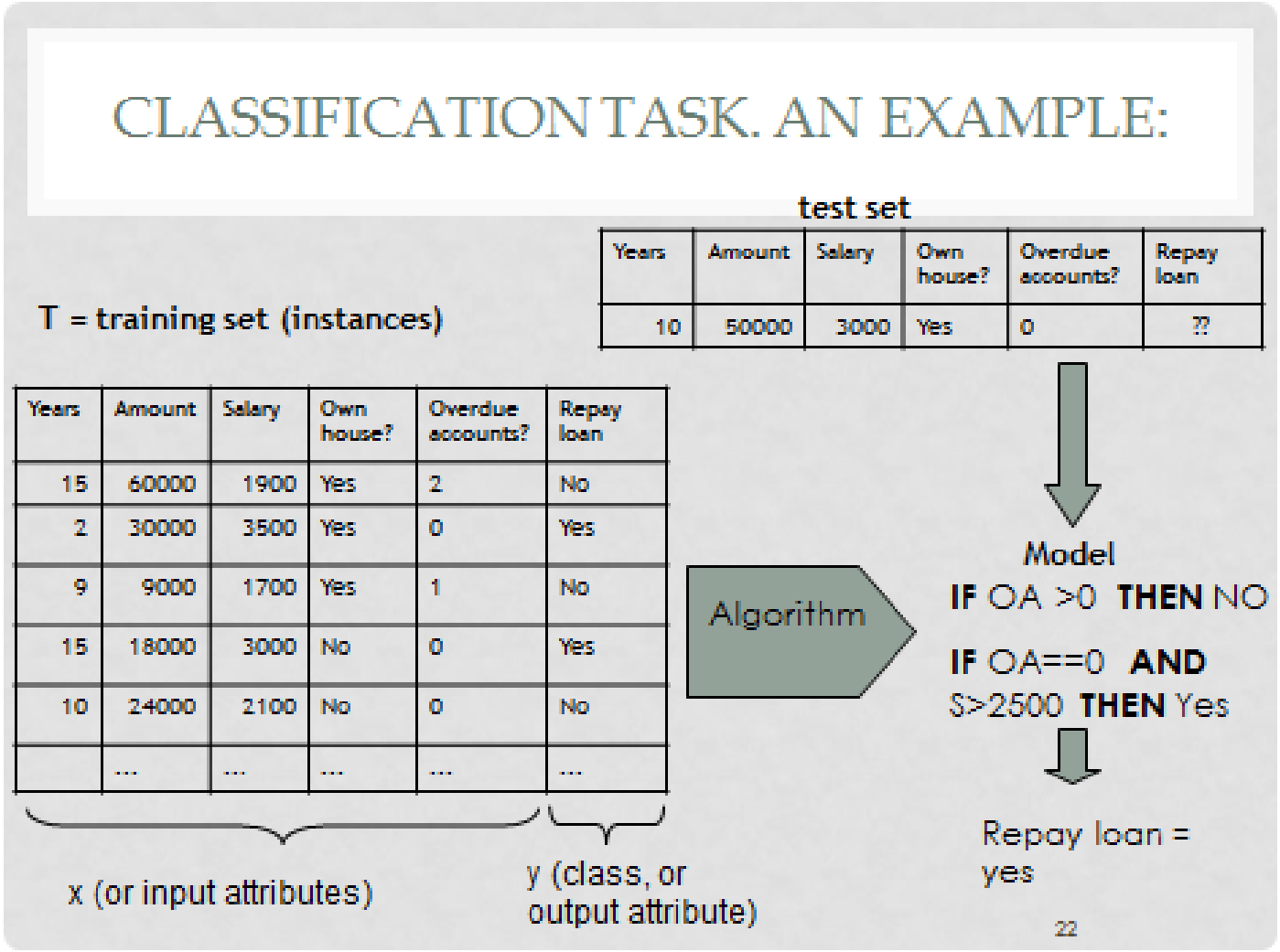
uc3m | Universidad Carlos III de Madrid

OPENCOURSEWARE
ADVANCED PROGRAMMING
STATISTICS FOR DATA SCIENCE

Ricardo Aler

First Assignment: a programming assignment for feature extraction

MACHINE LEARNING SCHEMA



ML dataset

- We know that our starting dataset should look like something like this:

Years	Amount	Salary	Own house?	Overdue accounts?	Repay loan
15	60000	1900	Yes	2	No
2	30000	3500	Yes	0	Yes
9	9000	1700	Yes	1	No
15	18000	3000	No	0	Yes
10	24000	2100	No	0	No
	…	…	…	…	…

Feature extraction

- But in many cases, we are not given data in that format and we have to carry out a process in order to convert it to this table format. This is called feature (or attribute) extraction
- For instance: text data mining

Text data mining

- Let’s suppose we are given different messages (posts) from forums, like twitter, facebook, or similar. And we want to be able to classify them into different categories
- In the old times, such forums were called newsgroups, and there were thousands of them

Text categorization

- Let's suppose we are given many messages from the newsgroup "alt.atheism" and the newsgroup "comp.graphics" (computer graphics)
  - Note: all messages from 20 newsgroups can be found here: http://qwone.com/~jason/20Newsgroups/
- And a news agency is interested in building a model that is able to classify messages (news) into two categories: atheism and computer graphics

Initial dataset for message classification (categorization)

- Typically, each message is characterized by word frequencies

	"God"	"nothingness"	"video"	"card"	"aliens"	…	Class
message1	30%	7%	0%	1%	10%		atheism
message2	1%	0%	40%	50%	5%		Comp.graphics
message3	0%	0%	10%	40%	0%		Comp.graphics
message4	25%	0%	0%	No	30%		atheism
message5	0%	1%	20%	20%	1%		Comp.graphics
message6		…	…	…	…		…
…				…			…

- But we are not given the table, but the individual messages.
- For instance, this is the first message from “alt.atheism”

EVOLUTION DESIGNS

Evolution Designs sell the "Darwin fish". It's a fish symbol, like the ones Christians stick on their cars, but with feet and the word "Darwin" written inside. The deluxe moulded 3D plastic fish is $4.95 postpaid in the US.

Write to: Evolution Designs, 7119 Laurel Canyon #4, North Hollywood, CA 91605

…

- And this is the first message from “comp.graphics”

From: weston@ucssun1.sdsu.edu (weston t)
Subject: graphical representation of vector-valued functions

gnuplot, etc. make it easy to plot real valued functions of 2 variables but I want to plot functions whose values are 2-vectors. I have been doing this by plotting arrays of arrows (complete with arrowheads) but before going further, I thought I would ask whether someone has already done the work. Any pointers?? thanx in advance Tom Weston | USENET: weston@ucssun1.sdsu.edu

Bag of words representation

- In order to compute the table:

	"God"	"nothingness"	"video"	"card"	"aliens"	...	Class
message1	30%	7%	0%	1%	10%		atheism
message2	1%	0%	40%	50%	5%		Comp.graphics
message3	0%	0%	10%	40%	0%		Comp.graphics
message4	25%	0%	0%	No	30%		atheism
message5	0%	1%	20%	20%	1%		Comp.graphics
message6		…	…	…	…		…
…				…			…

- For every message, we have to :
  - Split it into words
  - Count how many times each word appears in the message
  - Divide by the total number of words in the message, in order to compute the frequency

NLTK

- Important: there are libraries specialized in text mining, that automatize the bag-of-words feature extraction process
  - Natural Language Toolkit (NLTK):
    - https://www.nltk.org/
- But in this assignment, we will program it by hand, in order to learn Python programming