uc3m | Universidad Carlos III de Madrid

OPENCOURSEWARE
APRENDIZAJE AUTOMÁTICO PARA EL ANÁLISIS DE DATOS
GRADO EN ESTADÍSTICA Y EMPRESA
Ricardo Aler

INTRODUCCIÓN AL APRENDIZAJE AUTOMÁTICO

1. TIPOS DE TAREAS, MODELOS Y ALGORITMOS

TAREAS

- ¿Qué se puede hacer en aprendizaje automático? **Tareas**:
  - Supervisado: clasificación, predicción (regresión), …
  - No supervisado: asociación, agrupamiento (clustering), …
  - Semi-supervisado
  - Por refuerzo

TAREAS

- Tareas:
  - **Aprendizaje supervisado**:
    - **Clasificación:**
      - **Discreta**
      - Ranking, scoring, probability estimation
    - Predicción numérica (regresión)
  - Aprendizaje semi-supervisado
  - Aprendizaje no supervisado:
    - Agrupamiento o clustering
    - Asociación
  - Aprendizaje por refuerzo

UN EJEMPLO DE PROBLEMA SUPERVISADO DE CLASIFICACIÓN DISCRETA:

- Concesión de créditos bancarios
  - Un banco por Internet desea obtener reglas para predecir qué personas de las que solicitan un crédito no van a devolverlo.
  - La entidad bancaria cuenta con una gran base de datos correspondientes a los créditos concedidos (o no) a otros clientes con anterioridad.
  - Instancias (de la base de datos del banco):
    - Atributos de entrada: años del crédito, cuantía del crédito, tiene cuentas morosas, tiene casa propia
    - Clase: si/no
  - Modelo que se podría aprender:
    - **SI** (cuentas-morosas > 0) **ENTONCES** Devuelve-crédito = no
    - **SI** (cuentas-morosas = 0) **Y** ((salario > 2500) **O** (años > 10)) **ENTONCES** devuelve-crédito = si

UN EJEMPLO DE PROBLEMA SUPERVISADO DE CLASIFICACIÓN DISCRETA:

Instancia de test

Años	Euros	Salario	Casa propia	Cuentas morosas	Crédito
10	50000	3000	Si	0	??

T = Conjunto de instancias de entrenamiento (o ejemplos, datos, patrones, …)

Años	Euros	Salario	Casa propia	Cuentas morosas	Crédito
15	60000	2200	Si	2	No
2	30000	3500	Si	0	Si
9	9000	1700	Si	1	No
15	18000	1900	No	0	Si
10	24000	2100	No	0	No
	…	…	…	…	…

Debido a esta columna, la tarea es supervisada

Algoritmo

Modelo

IF CM >0 **THEN** NO

IF CM =0 **Y** S>2500 **THEN** SI

…

Crédito = Si

x: atributos (o características, predictores, variables independientes, variables de entrada, …)

y: clase (o etiqueta, atributo de salida, variable dependiente, variable respuesta, …)

APRENDIZAJE SUPERVISADO. CLASIFICACIÓN CON CLASES DISCRETAS

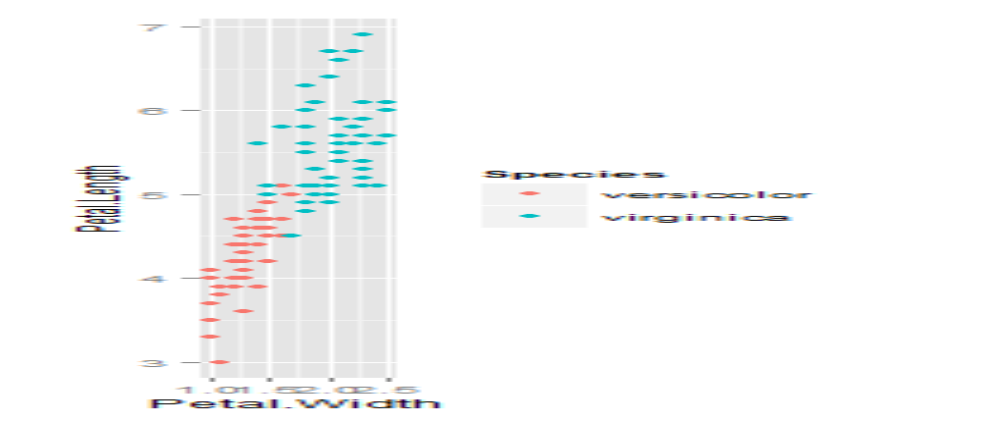
- Tipos de atributos:
  - Nominales / categóricos: verde, rojo, amarillo
  - Ordinales: frío, templado, caliente
  - Reales / enteros: 1.3, 7.9, 10.798, …
- $Y = \{c_1, c_2, \ldots, c_k\}$ son las clases.
  - Si k=2, problema de clasificación binaria: cáncer / no-cáncer
  - Si K>2, problema de clasificación multi-clase: peligroso / normal / inofensivo

DEFINICIÓN: FEATURE SPACE (ESPACIO DE INSTANCIAS)

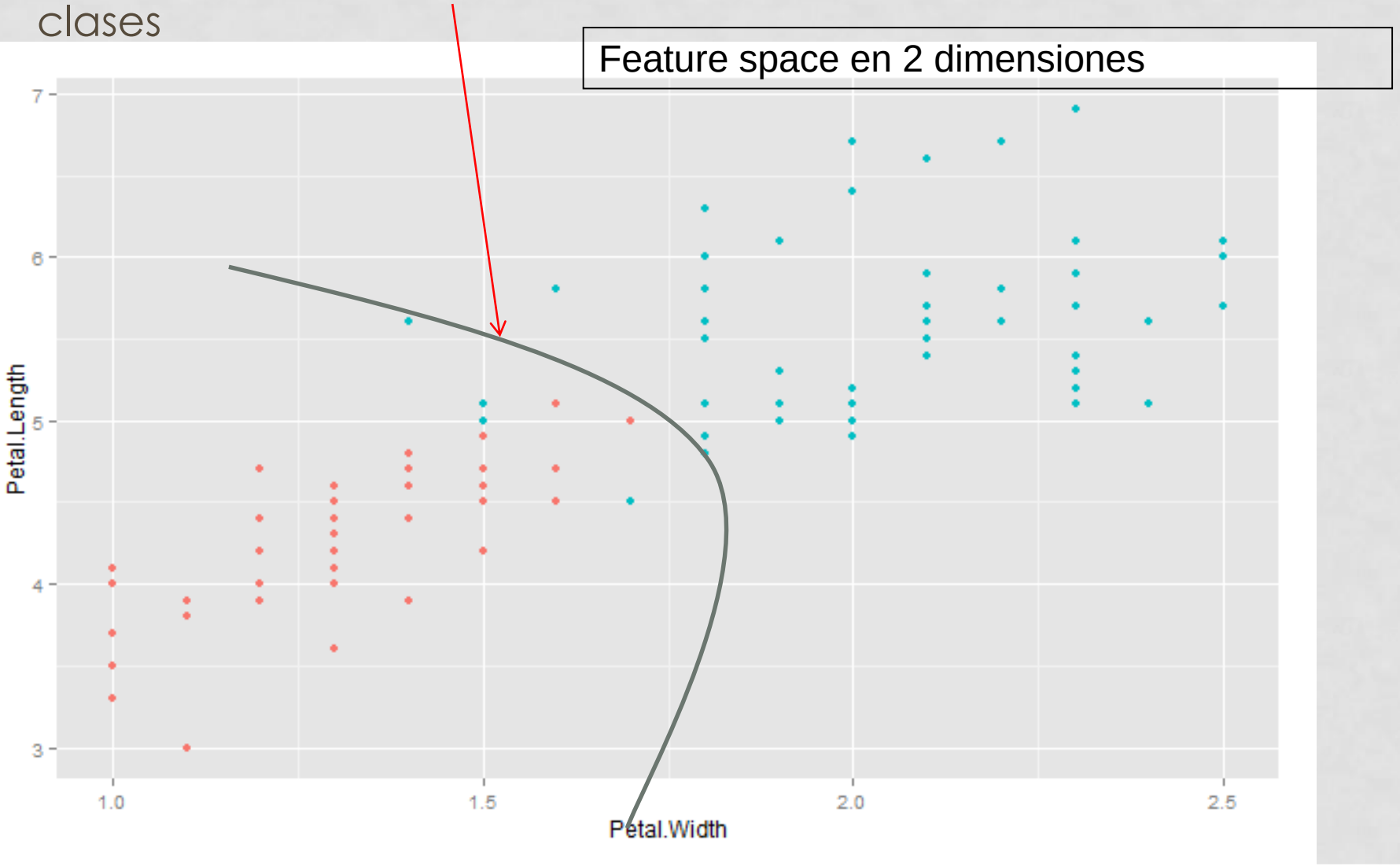
- Las instancias posibles "habitan" un espacio d-dimensional (donde *d* es el número de atributos de entrada)
  - Esta instancia tiene 5 atributos de entrada y 1 de salida

Años	Euros	Salario	Casa propia	Cuentas morosas	Crédito
10	50000	3000	Si	0	Si

  - En 2 dimensiones (2 atributos), cada instancia es un punto en el feature space



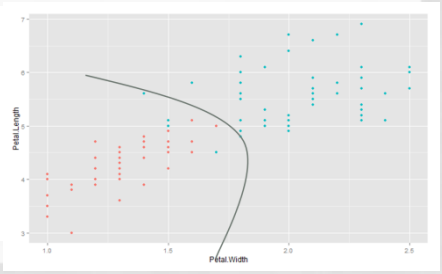
- Ejemplo: clasificar plantas en dos clases ("versicolor“ / roja vs. "virginica“ / azul)
- 2 atributos = (Petal.Width, Petal.Length) = 2 dimensiones
- Clasificación = encontrar una función **g: X → Y** frontera entre las clases



TAREAS

- Tareas:
  - **Aprendizaje supervisado**:
    - **Clasificación:**
      - Discrete
      - **Ranking, scoring, probability estimation**
    - Predicción (regresión)
  - Aprendizaje semi-supervisado
  - Aprendizaje no supervisado:
    - Agrupamiento o clustering
    - Asociación
  - Aprendizaje por refuerzo

SCORING-CLASSIFIERS / ESTIMACIÓN DE PROBABILIDADES



- *Scoring: en caso de que* la función g nos devuelva una medida de pertenencia de una instancia a una clase:
  - $g : X \rightarrow \boldsymbol{R}$ (en caso de clasificación binaria)
  - Por ejemplo, *g(x)* = distancia de x a la frontera de separación
- Estimación de probabilidades / densidades:
  - Un "score" no es una probabilidad (no cumple las leyes de las probabilidades)
  - g(x) = p(y | x)
    - p(y = si | x)
    - p(y = no | x)

TAREAS

- Tareas:
  - **Aprendizaje supervisado**:
    - Clasificación**:**
      - Discrete
      - Ranking, scoring, probability estimation
    - **Predicción numérica (regresión): la "clase" es contínua**
  - Aprendizaje semi-supervisado
  - Aprendizaje no supervisado:
    - Agrupamiento o clustering
    - Asociación
  - Aprendizaje por refuerzo

PREDICCIÓN (REGRESIÓN)

- Si la clase a predecir es continua el problema es de **predicción** (o **regresión**)
- Los modelos que se aprenden son funciones matemáticas *y=g(x)*
  - Lineal: y = ax+b
  - No lineal: $y = a*x^2+bx+c$ / y = log(sin(x))

TAREAS

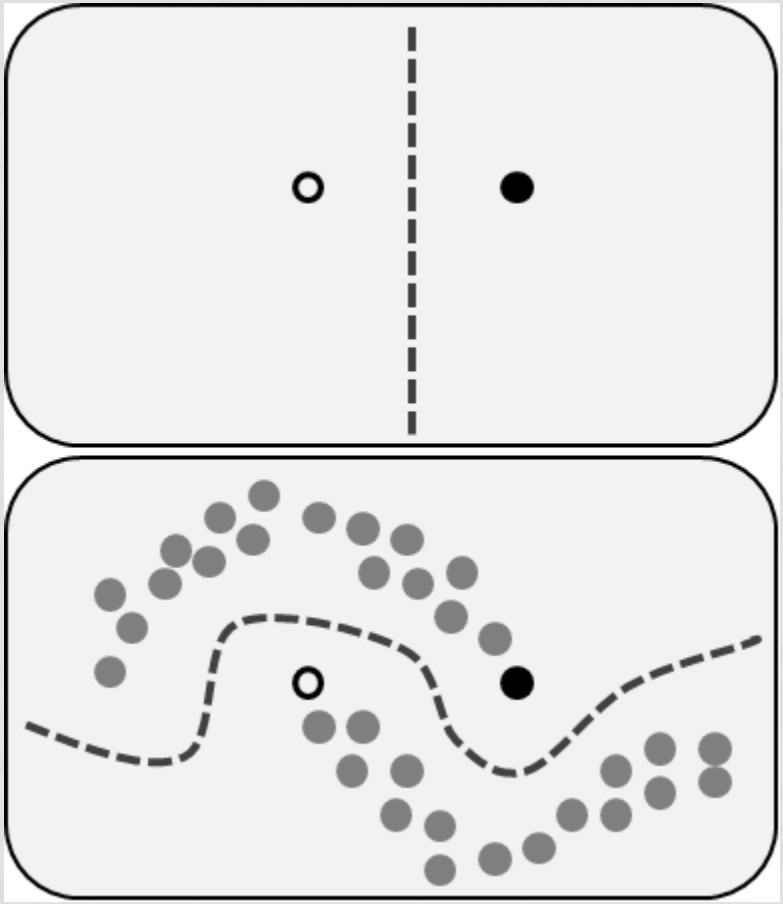
- Tareas:
  - **Aprendizaje supervisado**:
    - Clasificación**:**
      - Discrete
      - Ranking, scoring, probability estimation
    - Predicción numérica (regresión)
  - Aprendizaje semi-supervisado
  - Aprendizaje no supervisado:
    - Agrupamiento o clustering
    - Asociación
  - Aprendizaje por refuerzo

TAREAS

- Tareas:
  - Aprendizaje supervisado:
    - Clasificación
    - Predicción (regresión)
  - **Aprendizaje semi-supervisado**
  - Aprendizaje no supervisado:
    - Agrupamiento o clustering
    - Asociación
  - Aprendizaje por refuerzo

APRENDIZAJE SEMISUPERVISADO

- Se dispone de algunas instancias de entrenamiento etiquetadas y de otras muchas sin etiquetar (sólo los atributos de entrada)
- Justificación: etiquetar datos puede resultar costoso (ej: hacer una biopsia para saber si una persona tiene cáncer)



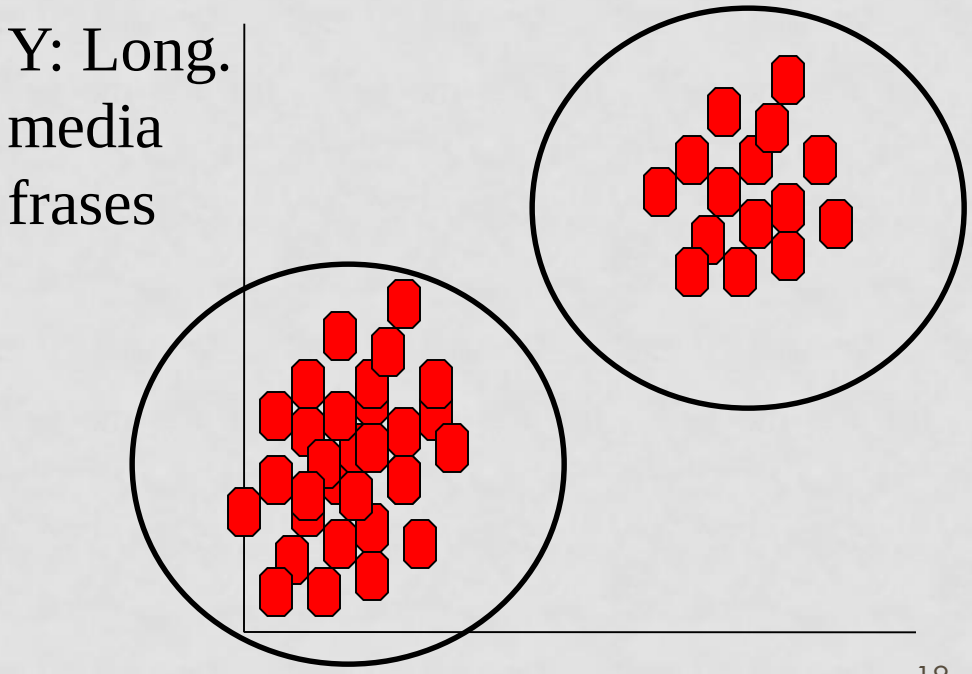
Source: https://en.wikipedia.org/wiki/Semi-supervised_learning

TAREAS

- Tareas:
  - Aprendizaje inductivo (a partir de ejemplos)
    - Aprendizaje supervisado:
      - Clasificación
      - Predicción (regresión)
    - Aprendizaje semi-supervisado
    - **Aprendizaje no supervisado**:
      - **Agrupamiento o clustering**
      - Asociación
    - Aprendizaje por refuerzo

IDEA GENERAL DE AGRUPACIÓN

- Detectar agrupaciones naturales en los datos
- Agrupación (o "clustering") = aprendizaje no supervisado: se parte de una tabla, como en clasificación, pero sin la clase



Ejemplo: clustering de libros. 2 grupos:

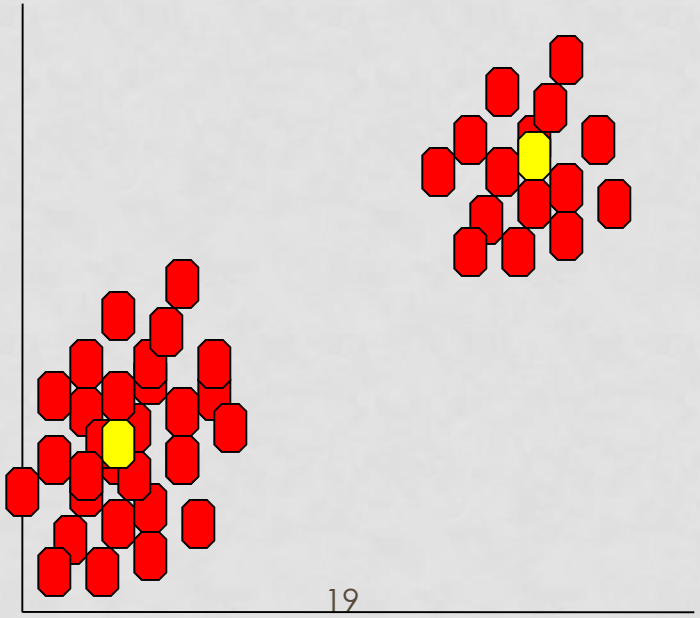
* Palabras y frases largas (¿filosofía?)

* Palabras y frases cortas (¿novela?)

REPRESENTACIÓN DE CLUSTERS

- Por sus centroides (ej: algoritmo k-medias)

K-MEDIAS / K-MEANS: http://www.youtube.com/watch?v=74rv4snLl70



$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

AGRUPACIÓN DE EMPLEADOS ("**CLUSTERING**")

- El departamento de RRHH de una empresa desea categorizar a sus empleados en distintos grupos con el objetivo de entender mejor su comportamiento y tratarlos de manera adecuada

EJEMPLO DE CLUSTERING

Id	Sueldo	Casado	Coche	Hijos	Alq/Prop	Sindicado	Bajas	Antigüedad	Sexo
1	1000	Si	No	0	Alq	No	7	15	H
2	2000	No	Si	1	Alq	Si	3	3	M
3	1500	Si	Si	2	Prop	Si	5	10	H
4	3000	Si	Si	1	Alq	No	15	7	M
5	1000	Si	Si	0	Prop	Si	1	6	H
…	…	…	…	…		…	…	…	…

MODELO

	GRUPO 1	GRUPO 2	GRUPO 3
Sueldo	1535	1428	1233
Casado (No/Si)	77%/22%	98%/2%	0%/100%
Coche	82%/18%	1%/99%	5%/95%
Hijos	0.05	0.3	2.3
Alq/Prop	99%/1%	75%/25%	17%/83%
Sindicado	80%/20%	0%/100%	67%/33%
Bajas	8.3	2.3	5.1
Antigüedad	8.7	8	8.1
Sexo (H/M)	61%/39%	25%/75%	83%/17%

- Grupo 1: sin hijos y con vivienda de alquiler. Poco sindicados. Muchas bajas
- Grupo 2: sin hijos y con coche. Muy sindicados. Pocas bajas. Normalmente son mujeres y viven en alquiler
- Grupo 3: con hijos, casados y con coche. Mayoritariamente hombres propietarios. Poco sindicados.

TAREAS

- Tareas:
  - Aprendizaje supervisado:
    - Clasificación
    - Predicción (regresión)
  - Aprendizaje semi-supervisado
  - **Aprendizaje no supervisado**:
    - Agrupamiento o clustering
    - **Asociación**
  - Aprendizaje por refuerzo

ANÁLISIS DE LA CESTA DE LA COMPRA (**ASOCIACIÓN**)

- Un supermercado quiere obtener información sobre el comportamiento de compra de sus clientes.
  - Si compra este producto también compra este otro producto
- Se piensa que de esta manera se puede mejorar el servicio, colocando ciertos productos juntos, etc.

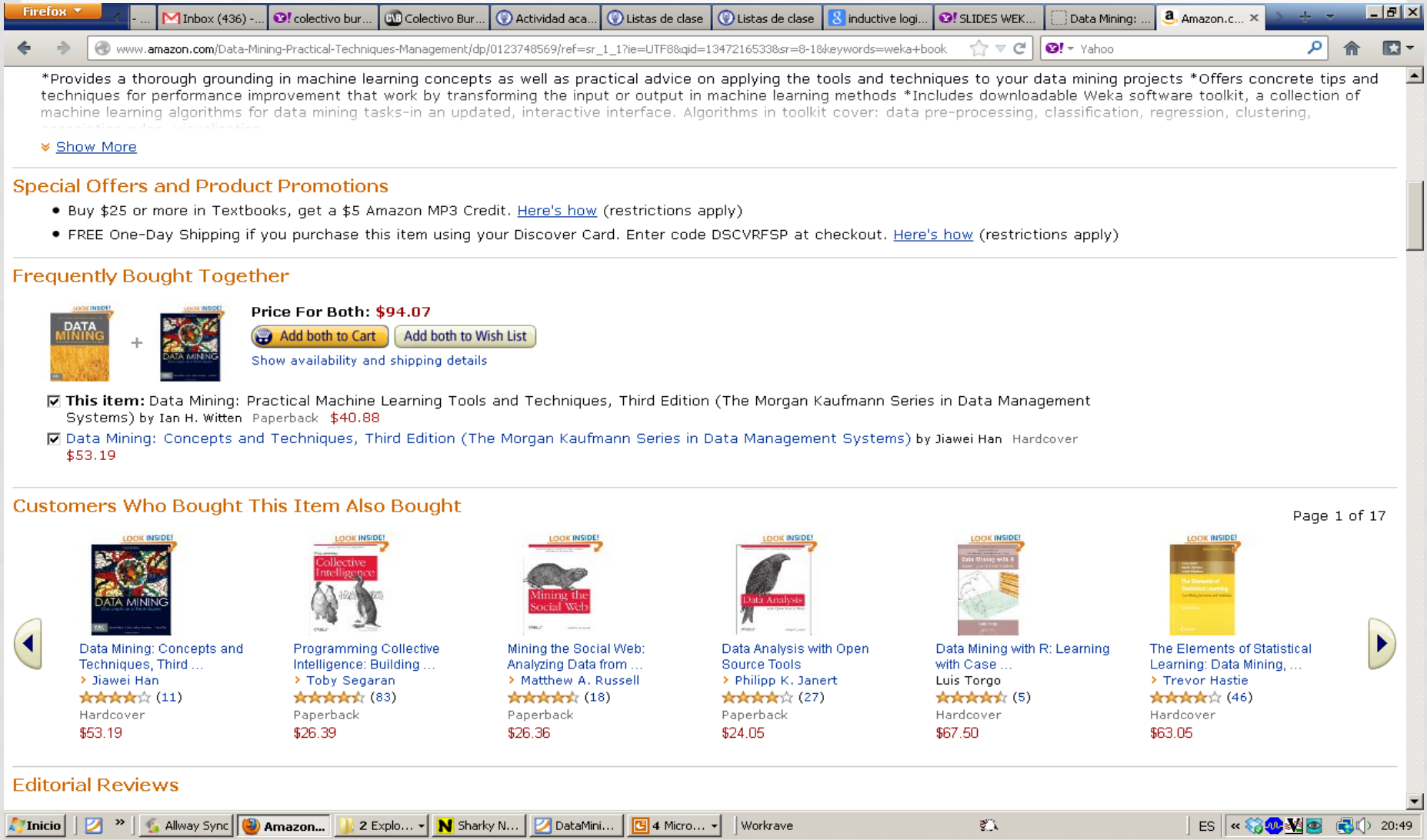
INSTANCIAS DE LAS CESTAS

Id	Huevos	Aceite	Pañales	Vino	Leche	Mantequilla	Salmón	Lechugas	…
1	Si	No	No	Si	No	Si	Si	Si	…
2	No	Si	No	No	Si	No	No	Si	…
3	No	No	Si	No	Si	No	No	No	…
4	No	Si	Si	No	Si	No	No	No	…
5	Si	Si	No	No	No	Si	No	Si	…
6	Si	No	No	Si	Si	Si	Si	No	…
7	No	No	No	No	No	No	No	No	…
8	Si	Si	Si	Si	Si	Si	Si	No	…
…	…	…	…	…	…	…	…	…	…

MODELO

- Reglas **Si** At_1=a y At_2=b y ... **Entonces** At_n=c
  - **Si** pañales=si, **entonces** leche=si
- Las reglas también pueden ser:
  - **Si** At_1=a y At_2=b **Entonces** At_n=c, **At_4=D**
  - **Si** huevos = si **Y** vino = si **Entonces** Mantequilla = si **Y** Salmón = si

ASOCIACIÓN: OTROS EJEMPLOS

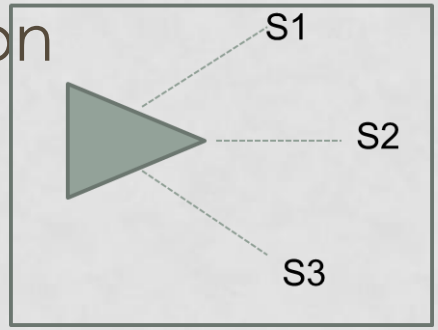


TAREAS

- Tareas:
  - Aprendizaje inductivo (a partir de ejemplos)
    - Aprendizaje supervisado:
      - Clasificación
      - Predicción (regresión)
    - Aprendizaje semi-supervisado
    - Aprendizaje no supervisado:
      - Agrupamiento o clustering
      - Asociación
    - **Aprendizaje por refuerzo**

APRENDIZAJE POR REFUERZO

- El agente tiene que aprender la "política" (policy) o estrategia π para saber que acción tomar en cada situación s
- Sea un robot que:
  - Tiene 3 sensores de distancia S1, S2, S3
  - Acciones: girar hacia la izquierda, hacia la derecha, avanzar
- Objetivo: aprender estrategia (policy) π para saber que acción tomar en cada situación s, para por ejemplo, moverse por una habitación con obstáculos sin chocar
  - π(s1, s2, s3) = acción



APRENDIZAJE POR REFUERZO

- No puede ser un problema supervisado, porque en principio, no podemos decirle al robot que acción realizar en cada posible situación.

S1	S2	S3	π
1.3	0.5	7	?
10	8.7	5	?
0.5	0.5	0.6	?
…	…	…	…

- La política es aprendida dejando que el robot explore un espacio simulado, recibiendo refuerzos positivos (de vez en cuando) al hacer algo bien, y negativos en caso contrario (ej: chocar).