

Métodos Matemáticos en Ciencias de la Computación 2007–2008

Abdelmalik Moujahid, Iñaki Inza y Pedro Larrañaga

Departamento de Ciencias de la Computación e Inteligencia Artificial

Universidad del País Vasco

<http://www.sc.ehu.es/isg/>

Temario

Introducción a la Asignatura

PARTE I: HEURÍSTICOS ESTOCÁSTICOS DE BÚSQUEDA

1. Heurísticos en Optimización Combinatorial
2. Algoritmos Genéticos
3. Algoritmos de Estimación de Distribuciones

Temario

PARTE II: RECONOCIMIENTO DE PATRONES

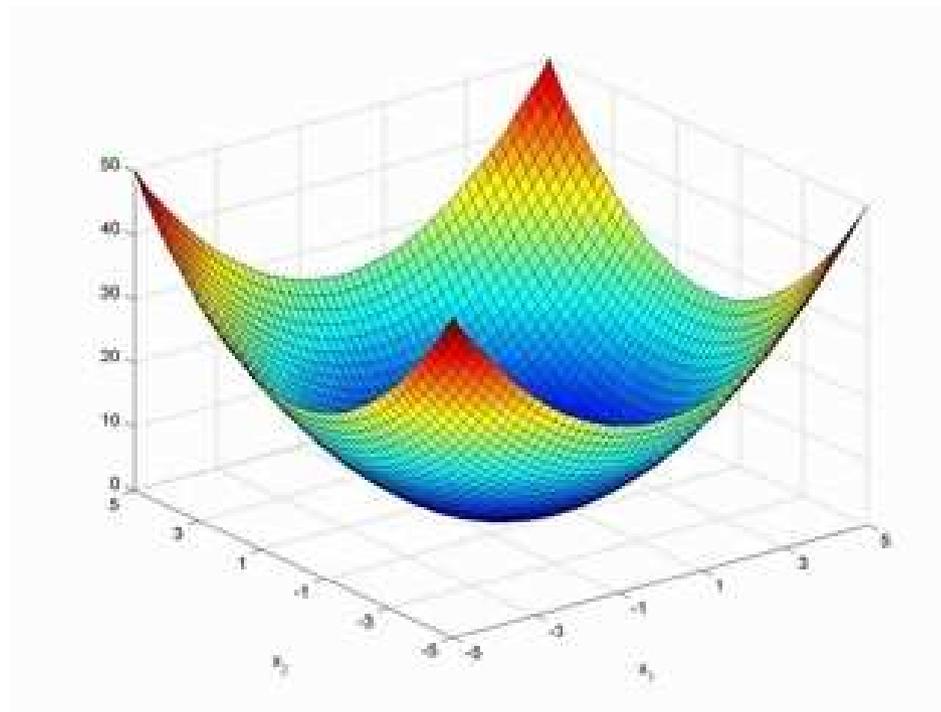
4. Aprendizaje Automático
5. Clasificadores k -NN
6. Evaluación de Modelos de Clasificación Supervisada
7. Teoría de la información
8. Árboles de Clasificación
9. Inducción de Reglas
10. Selección de Variables
11. Clustering
12. Clasificadores Bayesianos
13. Regresión Logística
14. Redes Neuronales
15. Combinación de Clasificadores

Métodos Matemáticos en Ciencias de la Computación

Inteligencia Computacional (optimización, modelización)

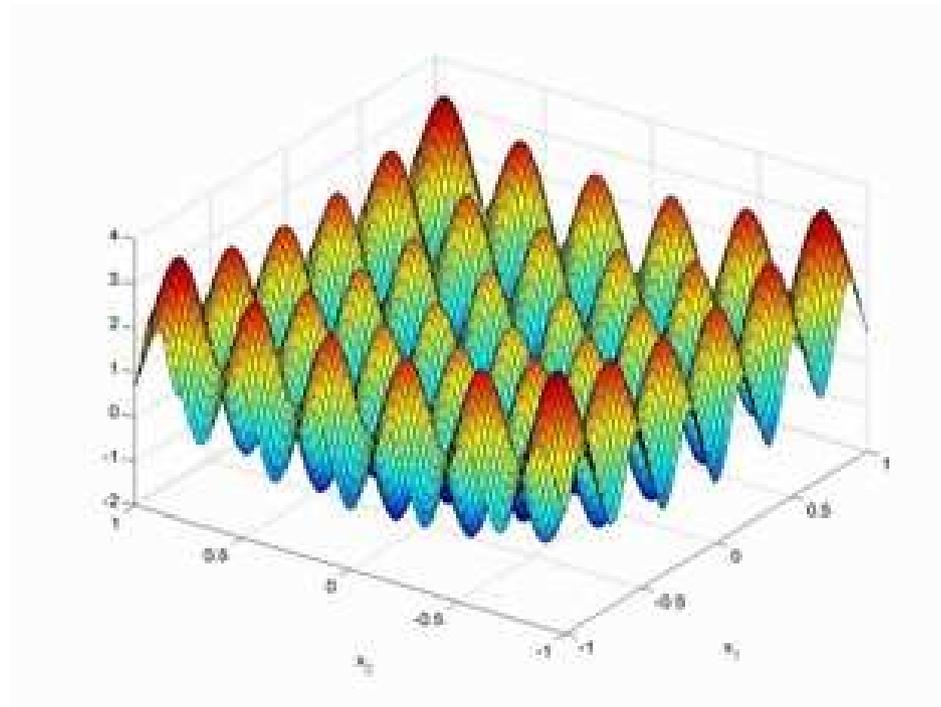
- Optimización (exacta - heurística)
 - Exacta: Investigación operativa
 - Heurística: Determinista versus estocástica
- Modelización (minería de datos)
 - De datos a conocimiento (descubrir relaciones, estructurar un dominio, ...)
 - Clasificación supervisada (*supervised pattern recognition*)
 - Clasificación no supervisada (*unsupervised pattern recognition*)

PARTE I: HEURÍSTICOS ESTOCÁSTICOS DE BÚSQUEDA



Optimización sencilla

PARTE I: HEURÍSTICOS ESTOCÁSTICOS DE BÚSQUEDA



Optimización difícil

PARTE II: RECONOCIMIENTO DE PATRONES (SUPERVISADO)

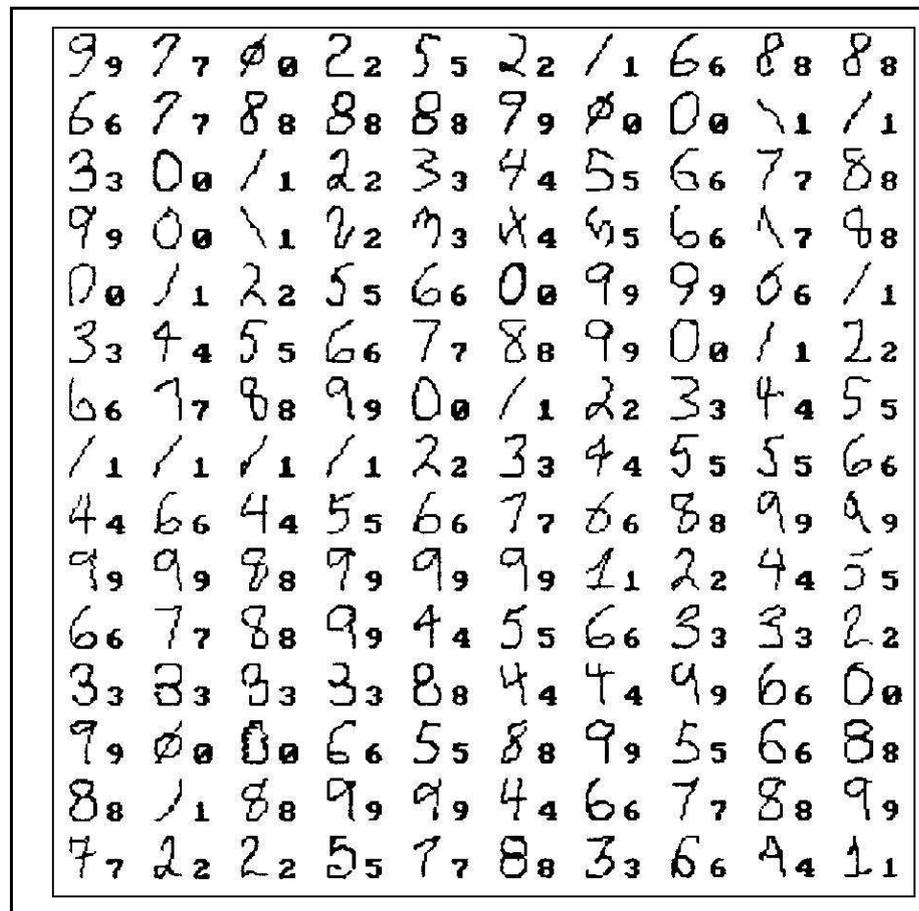
- Sistemas de ayuda al diagnóstico y al pronóstico
- Concesión de créditos
- Detección de correo basura
- Mercado bursátil
- Predicción de resultados deportivos

PARTE II: RECONOCIMIENTO DE PATRONES (SUPERVISADO)

age	employe	education	edur	marital	...	job	relation	race	gender	hour	country	wealth
39	State_gov	Bachelors	13	Never_mar	...	Adm_cleric	Not_in_fan	White	Male	40	United_Stat	poor
51	Self_emp_	Bachelors	13	Married	...	Exec_man	Husband	White	Male	13	United_Stat	poor
39	Private	HS_grad	9	Divorced	...	Handlers_c	Not_in_fan	White	Male	40	United_Stat	poor
54	Private	11th	7	Married	...	Handlers_c	Husband	Black	Male	40	United_Stat	poor
28	Private	Bachelors	13	Married	...	Prof_speci	Wife	Black	Female	40	Cuba	poor
38	Private	Masters	14	Married	...	Exec_man	Wife	White	Female	40	United_Stat	poor
50	Private	9th	5	Married_sp	...	Other_serv	Not_in_fan	Black	Female	16	Jamaica	poor
52	Self_emp_	HS_grad	9	Married	...	Exec_man	Husband	White	Male	45	United_Stat	rich
31	Private	Masters	14	Never_mar	...	Prof_speci	Not_in_fan	White	Female	50	United_Stat	rich
42	Private	Bachelors	13	Married	...	Exec_man	Husband	White	Male	40	United_Stat	rich
37	Private	Some_coll	10	Married	...	Exec_man	Husband	Black	Male	80	United_Stat	rich
30	State_gov	Bachelors	13	Married	...	Prof_speci	Husband	Asian	Male	40	India	rich
24	Private	Bachelors	13	Never_mar	...	Adm_cleric	Own_child	White	Female	30	United_Stat	poor
33	Private	Assoc_acc	12	Never_mar	...	Sales	Not_in_fan	Black	Male	50	United_Stat	poor
41	Private	Assoc_voc	11	Married	...	Craft_repa	Husband	Asian	Male	40	*MissingV	rich
34	Private	7th_8th	4	Married	...	Transport_	Husband	Amer_India	Male	45	Mexico	poor
26	Self_emp_	HS_grad	9	Never_mar	...	Farming_fi	Own_child	White	Male	35	United_Stat	poor
33	Private	HS_grad	9	Never_mar	...	Machine_c	Unmarried	White	Male	40	United_Stat	poor
38	Private	11th	7	Married	...	Sales	Husband	White	Male	50	United_Stat	poor
44	Self_emp_	Masters	14	Divorced	...	Exec_man	Unmarried	White	Female	45	United_Stat	rich
41	Private	Doctorate	16	Married	...	Prof_speci	Husband	White	Male	60	United_Stat	rich
:	:	:	:	:	:	:	:	:	:	:	:	:

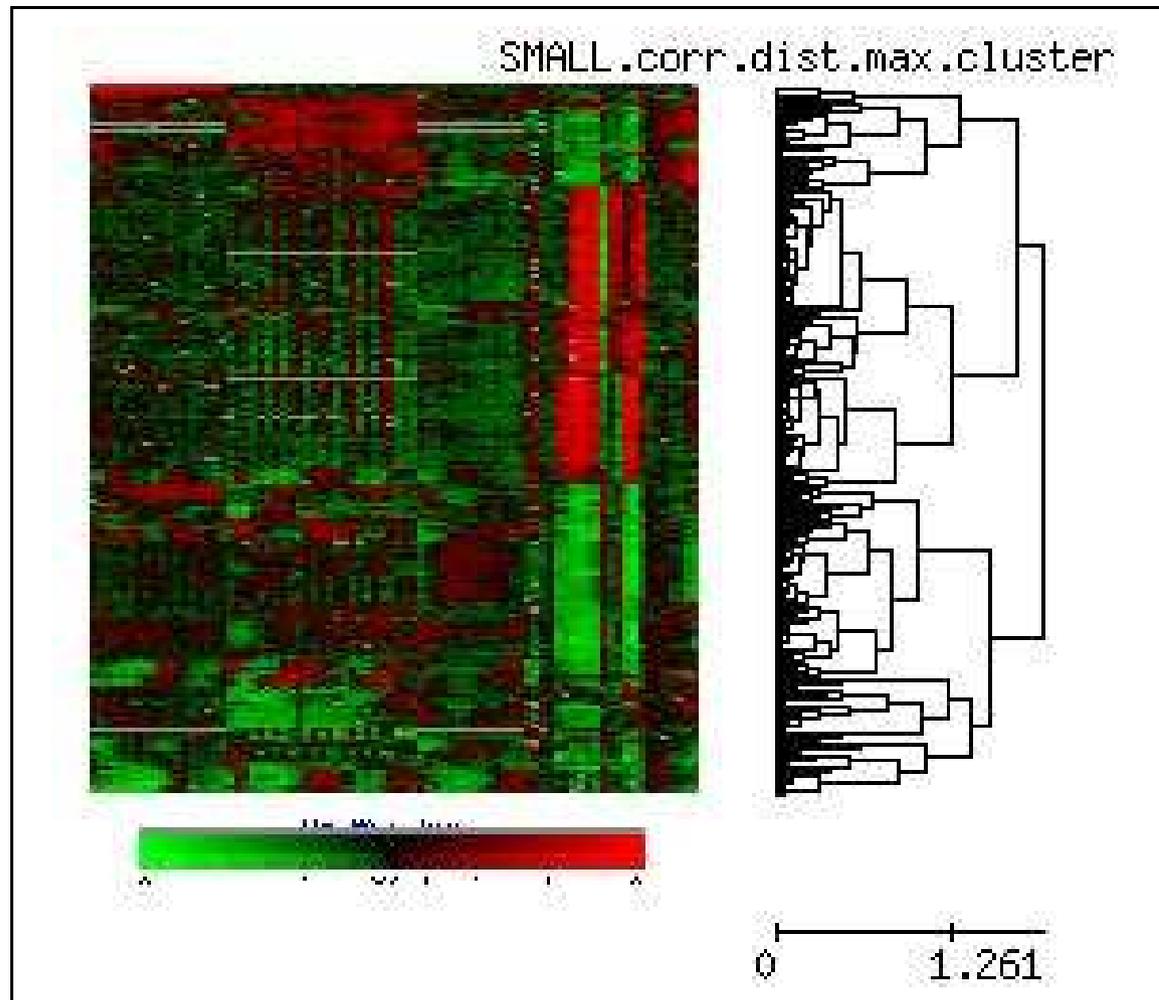
Datos en un problema de clasificación supervisada

PARTE II: RECONOCIMIENTO DE PATRONES (SUPERVISADO)



Reconocimiento de caracteres escritos a mano

PARTE II: RECONOCIMIENTO DE PATRONES (SUPERVISADO)



Microarray de ADN

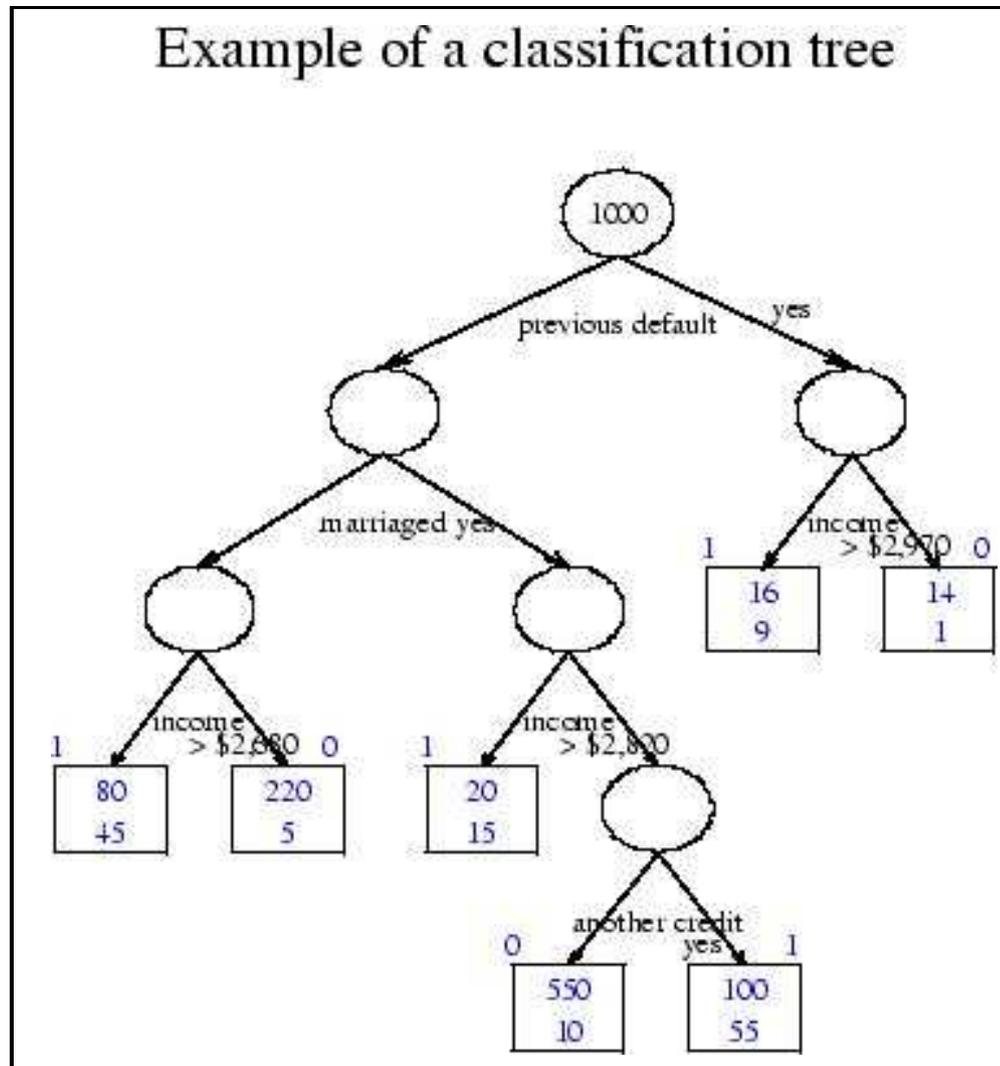
Clasificación Supervisada

	X_1	\dots	X_n	C
$(\mathbf{x}^{(1)}, c^{(1)})$	$x_1^{(1)}$	\dots	$x_n^{(1)}$	$c^{(1)}$
$(\mathbf{x}^{(2)}, c^{(2)})$	$x_1^{(2)}$	\dots	$x_n^{(2)}$	$c^{(2)}$
\dots		\dots		\dots
$(\mathbf{x}^{(N)}, c^{(N)})$	$x_1^{(N)}$	\dots	$x_n^{(N)}$	$c^{(N)}$
$\mathbf{x}^{(N+1)}$	$x_1^{(N+1)}$	\dots	$x_n^{(N+1)}$???

Paradigmas de Clasificación Supervisada

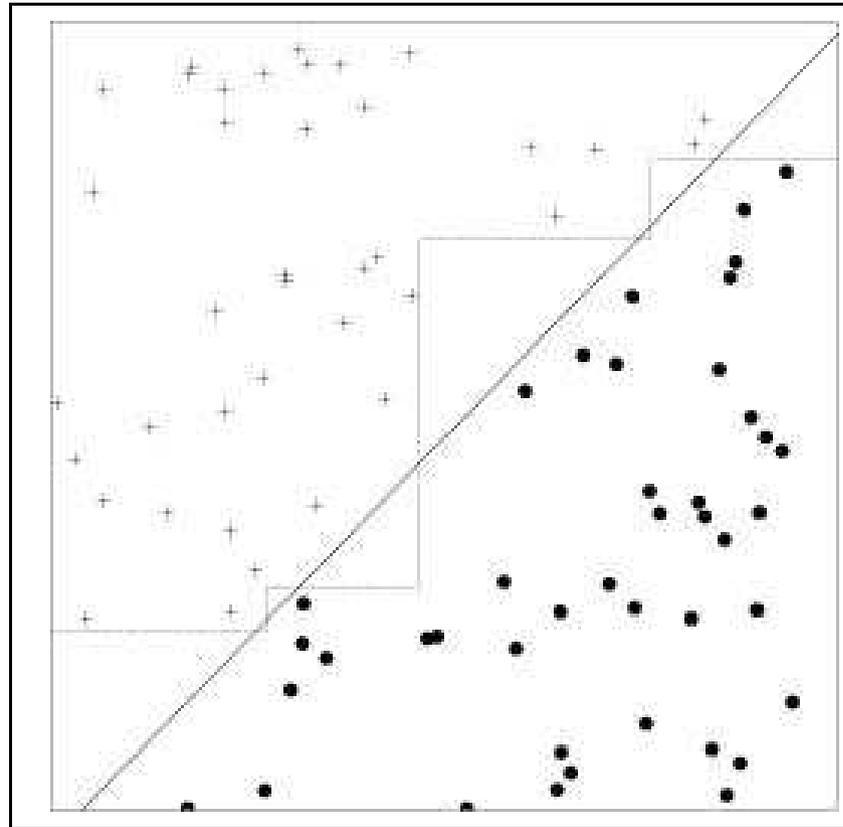
- Análisis discriminante
- Árboles de clasificación
- Clasificadores basados en redes Bayesianas
- Clasificadores k -NN
- Inducción de reglas
- Redes Bayesianas
- Redes neuronales
- Regresión logística
- Sistemas clasificadores
- ...

PARTE II: RECONOCIMIENTO DE PATRONES (SUPERVISADO)



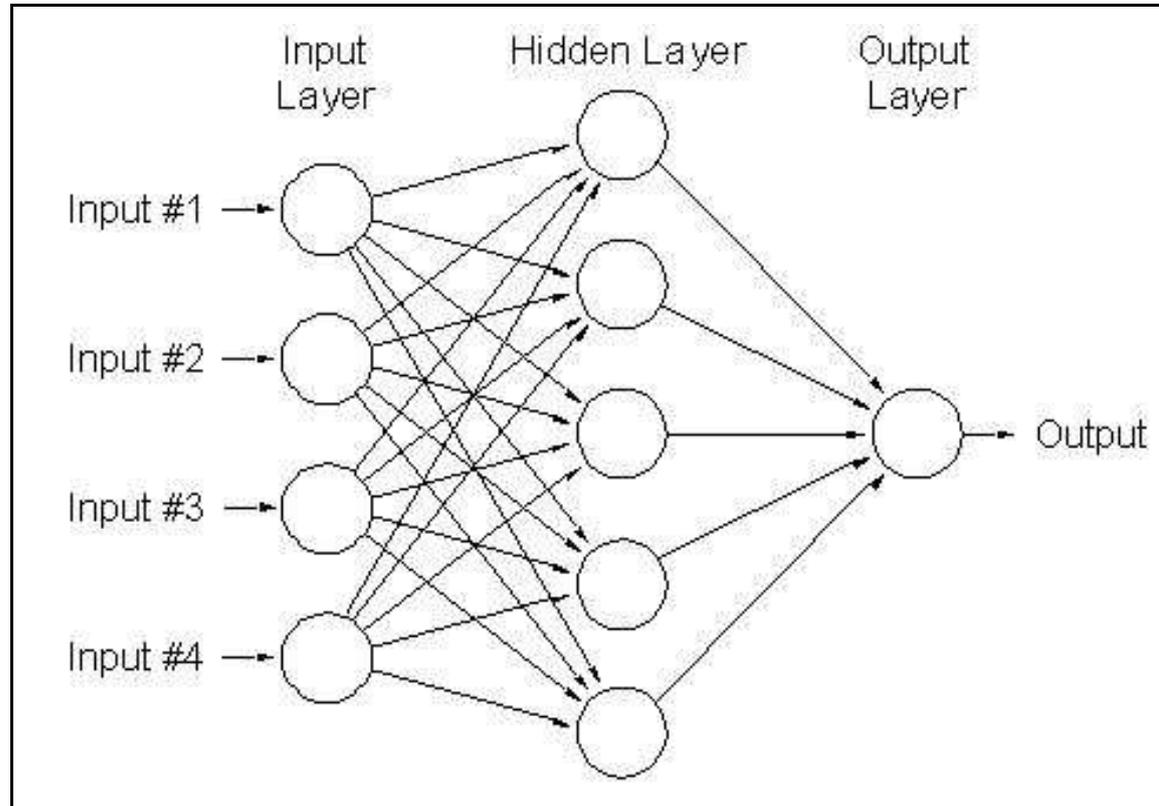
Árbol de clasificación

PARTE II: RECONOCIMIENTO DE PATRONES (SUPERVISADO)



Superficie de decisión de un árbol de clasificación

PARTE II: RECONOCIMIENTO DE PATRONES (SUPERVISADO)



Red neuronal

Criterios de Comparación de Clasificadores

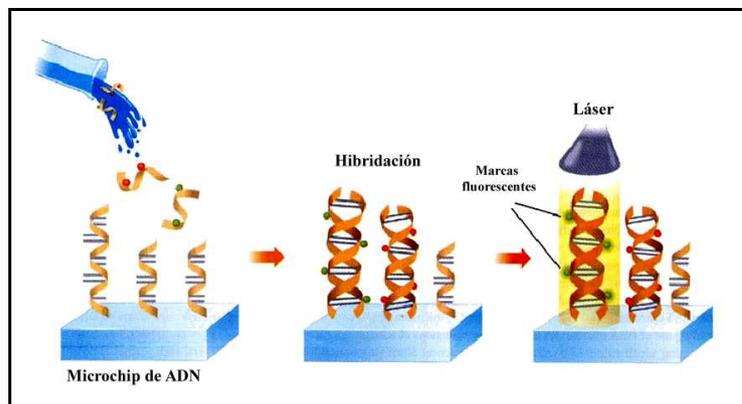
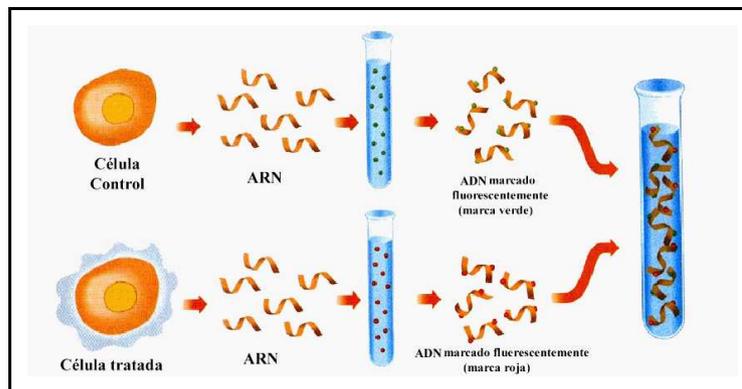
- Tasa de acierto
- Área bajo la curva ROC
- Complejidad algorítmica del inductor
- Interpretabilidad del modelo
- Simplicidad del modelo
- ...

*Selección de Genes en el Diagnóstico del Cáncer a partir de
Microchips de ADN*

- Bioinformática desarrollo espectacular
- Microchips de ADN transformación de la biología
- Posible medir el nivel de expresión de miles de genes al mismo tiempo
- Análisis de la información recogida para la comprensión de aspectos involucrados en los procesos biológicos

PARTE II: RECONOCIMIENTO DE PATRONES (SUPERVISADO)

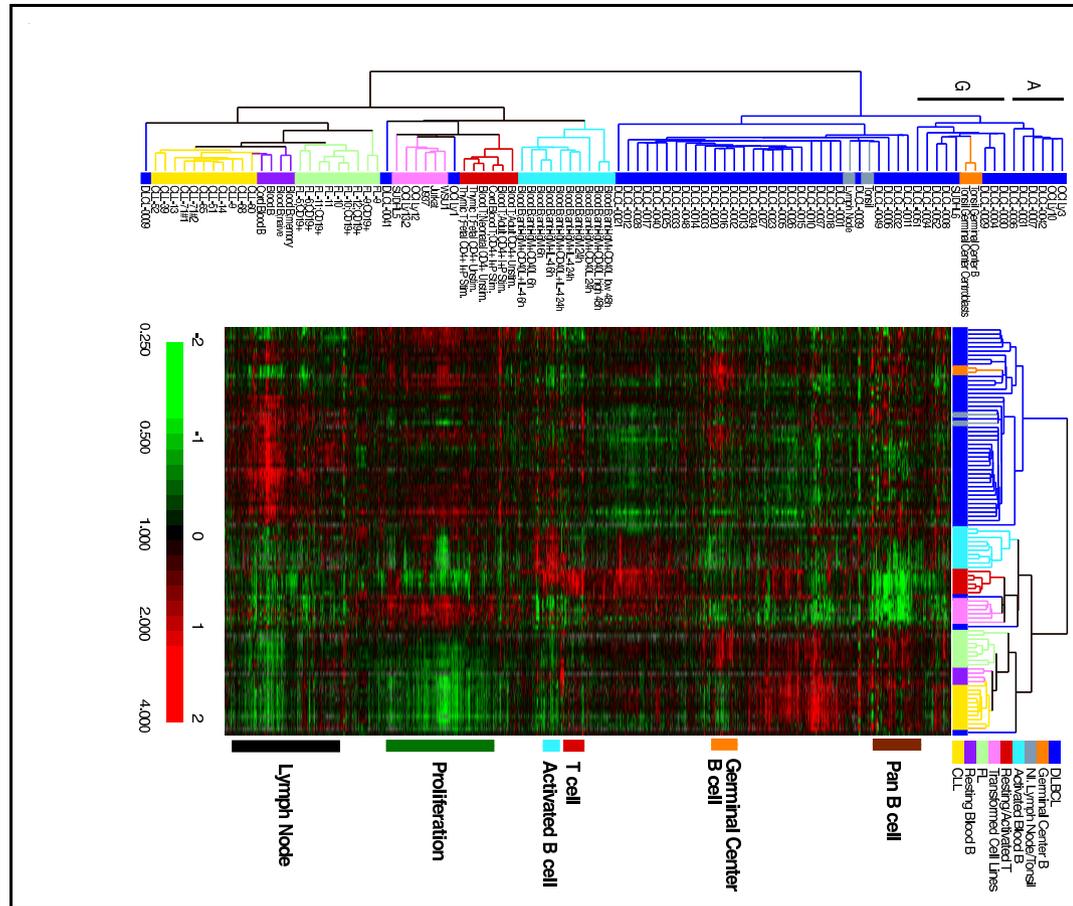
Selección de Genes en el Diagnóstico del Cáncer a partir de Microchips de ADN



Esquema representativo de la técnica de microchips de ADN

PARTE II: RECONOCIMIENTO DE PATRONES (SUPERVISADO)

Selección de Genes en el Diagnóstico del Cáncer a partir de Microchips de ADN



Niveles de expresión génica

PARTE II: RECONOCIMIENTO DE PATRONES (SUPERVISADO)

Selección de Genes en el Diagnóstico del Cáncer a partir de Microchips de ADN

Cáncer de colon 62 muestras (22 + 40), 2000 genes (Ben-Dor y col., 2000)

	Porcentaje	Número de variables
Naïve Bayes	88.70 ± 4.05	32
Seminaïve Bayes	87.09 ± 4.29	32
TAN	88.70 ± 4.05	32
k DB	83.87 ± 4.71	32

Cáncer de colon. Resultados con algoritmos de filtrado + filtrado

PARTE II: RECONOCIMIENTO DE PATRONES (NO SUPERVISADO)

- Cluster de documentos en la web (text mining)
- Segmentación de imágenes
- Taxonomía de seres vivos
- Árbol filogenético

PARTE II: RECONOCIMIENTO DE PATRONES (NO SUPERVISADO)

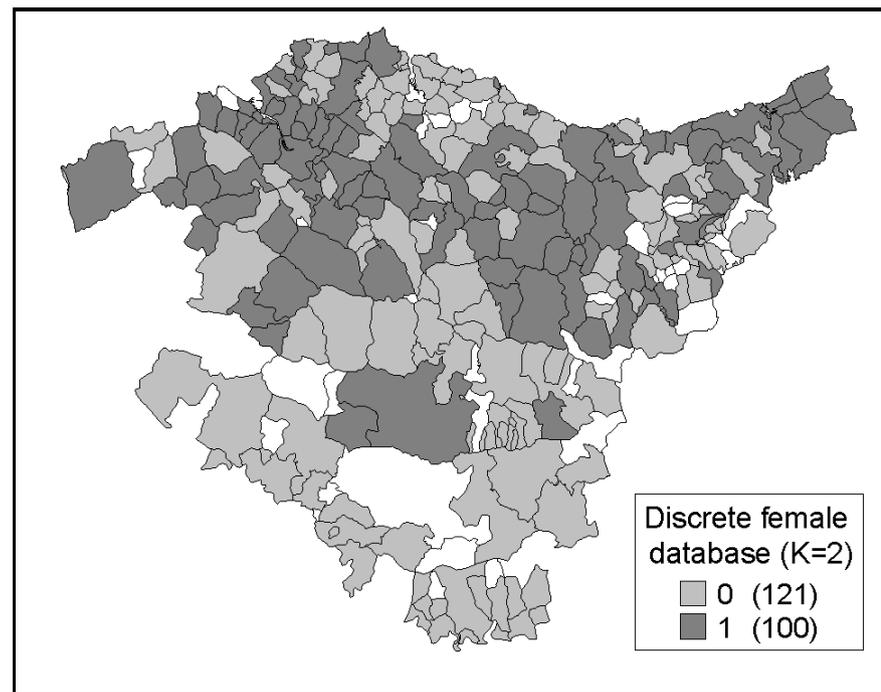
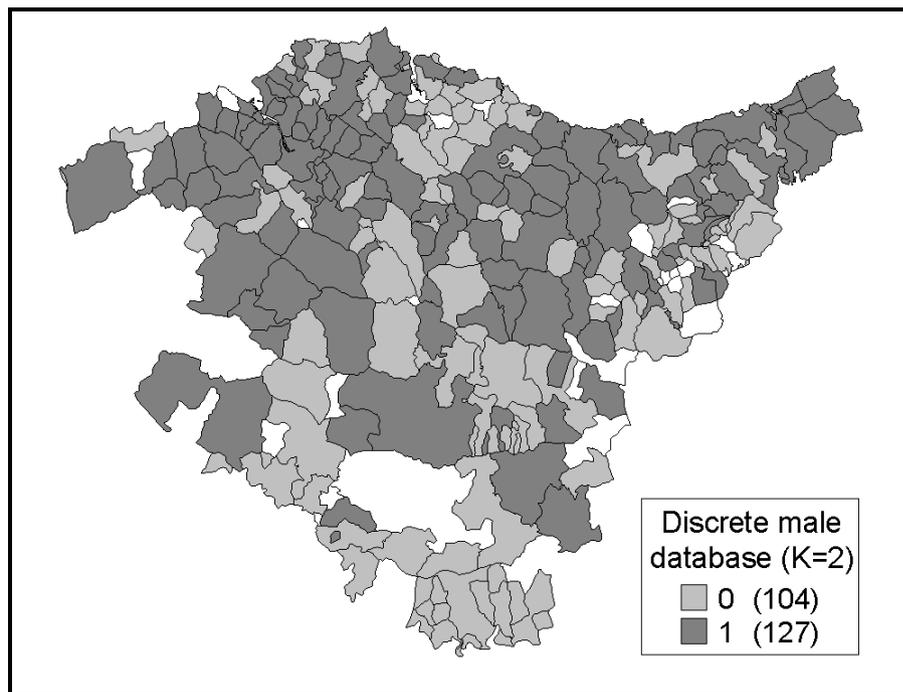
	Y_1	...	Y_i	...	Y_n	C
\mathbf{x}_1	y_1^1	...	y_i^1	...	y_n^1	$c^1?$
...
\mathbf{x}_j	y_1^j	...	y_i^j	...	y_n^j	$c^j?$
...
\mathbf{x}_N	y_1^N	...	y_i^N	...	y_n^N	$c^N?$

Clasificación no supervisada

Cluster de pueblos de la CAPV en base a la incidencia del cáncer

- Base de datos conteniendo la incidencia del cáncer estandarizada por la edad en cada pueblo de la Comunidad Autónoma del País Vasco
- Periodo de estudio: de 1986 a 1994
- Las 6 categorías de cáncer más frecuentes por sexo
- Analizar la distribución geográfica del tumor maligno
- Clusters de pueblos

PARTE II: RECONOCIMIENTO DE PATRONES (NO SUPERVISADO)



Mapas mostrando los clusters geográficos asumiendo $K = 2$