Modelos de Clasificación

Dr. Álvaro Pardo
Universidad Católica del Uruguay
apardo@ucu.edu.uy
@AlvaroPardoUy

Presentación

Director del Departamento de Ingeniería de la Universidad Católica del Uruguay.

Profesor de las asignaturas de IA en la Business School de la UCU.

Socio fundador de Digital Sense empresa dedicada a brindar servicios de I+D en Machine Learning, Computer Vision y Remote Sensing.

Integrante del Directorio y Vicepresidente de la Agencia Nacional de Investigación en Innovación de Uruguay.

Aprobación del curso

Actividad	**Instrumento**	**Peso**
Resolución de ejercicios y problemas	Rúbrica	40%
Cuestionarios	Test	40%
Autoevaluación	Diana	20%

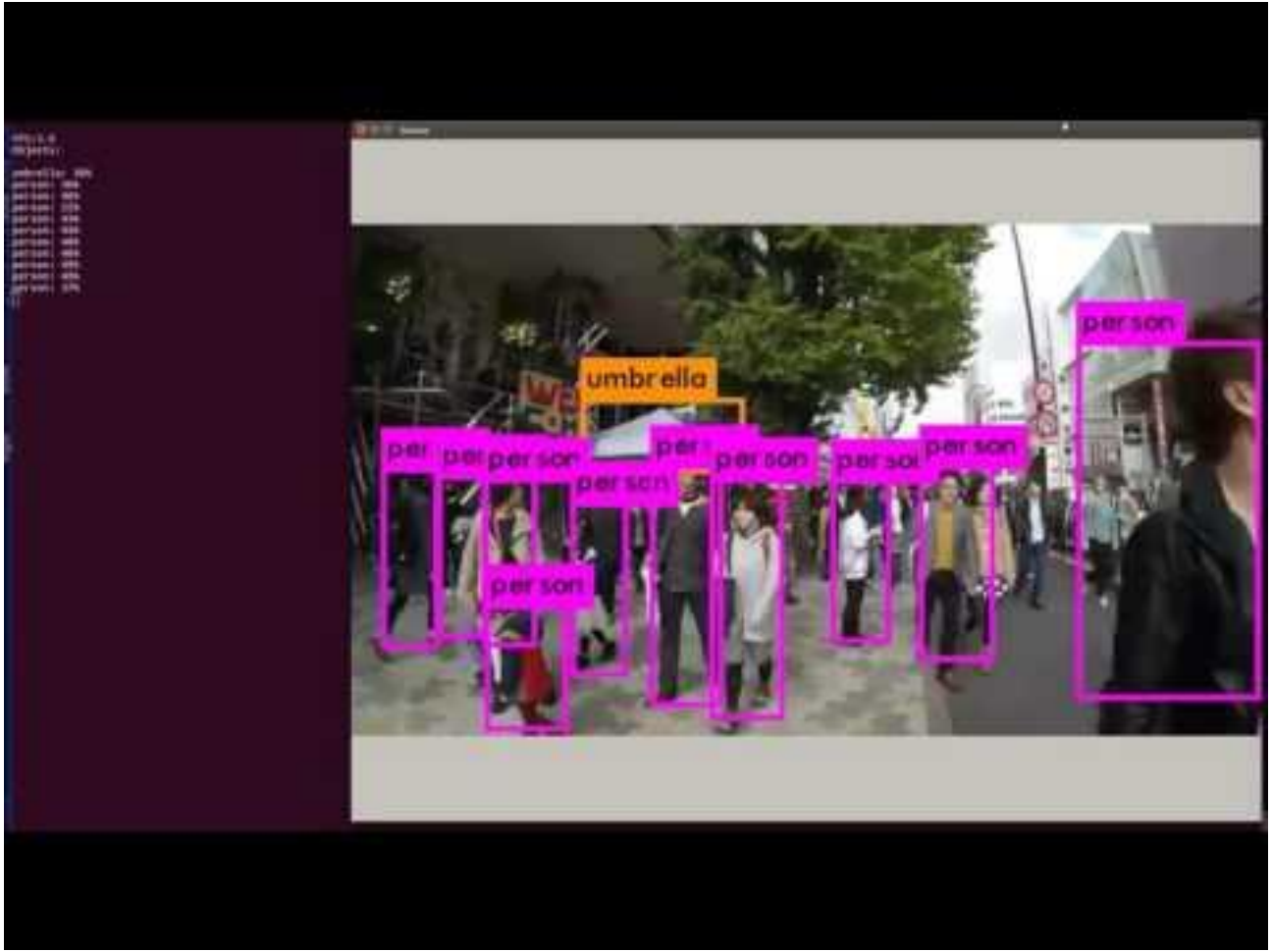
Inteligencia Artificial: Introducción

simulation of intelligent behaviour by computers

McCarthy (1956)

Inteligencia Artificial: Historia

1950 Turing
1951 Minsky
Red Neuronal
1958 Rosenblatt
Perceptron
1967 Vecinos
Cercanos
1970 - 1980 Redes
Neuronales
~90s SVM, RF,...
1997 IBM Deep Blue
2006 Netflix
2009 ImageNet


umbrella
person
person
person
person
person
person
person
person
person

Inteligencia Artificial
Machine Learning
Speech Recognition
...

Abordaje Simbólico

Inteligencia Artificial basada en sentencias del tipo **SI … ENTONCES ...** (Weak AI)

Ejercicio: Ta Te Ti

Machine Learning

Se instruye a la máquina mediante datos y no mediante instrucciones programas explícitamente.

Abordaje Simbólico vs Machine Learning

Introducción a Machine Learning

La capacidad de las máquinas para resolver problemas sin ser explícitamente programadas

“**Una máquina aprende** respecto a una **tarea T**, una métrica de **performance P** y un tipo de **experiencia E**, si **el sistema mejora su performance P para la tarea T en base a la experiencia E**” Mitchell (2006)

Ejemplo:

- **TAREA (T):** Clasificar transacciones fraudulentas y legítimas en función de los atributos MONTO y HORA.
- **MÉTRICA DE PERFORMANCE (P):** Cantidad de transacciones mal clasificadas.
- **EXPERIENCIA (E):** Histórico de transacciones etiquetadas como fraudulentas o legítimas.
- **CLASIFICADOR (MÁQUINA):** Ejemplo: aprende los umbrales de decisión para MONTO y HORA para resolver T, minimizando P en función de los datos E.

APRENDIZAJE

SUPERVISADO

Se entrenan algoritmos en base a datos de entrenamiento con entradas y salidas (típicamente etiquetadas por humanos)

NO SUPERVISADO

Los datos de entrenamiento no están etiquetados. Se busca estrucutura en ellos.

SUPERVISADO

- Conjunto de entrenamiento X con muestras etiquetadas.
- Para cada muestra de X se conoce el resultado esperado (clase, etiqueta)
- El objetivo es aprender a partir de X para poder predecir la salida para nuevas muestras.
- Ejemplo: Reconocer si un correo spam.

NO SUPERVISADO

- Los datos de entrenamiento X no están etiquetados.
- El objetivo es encontrar estructuras de interés en los datos.
- Ejemplos: Segmentación de clientes, encontrar las variables de interés para un problema (que muestran variaciones estadísticas interesantes)

PROBLEMAS

Clasificación / Regresión / Agrupamiento (Clustering, Segmentación) /Asociación / Recomendación

CLASIFICACIÓN

SALIDA

Es una categoría (clase) que puede estar acompañada por una probabilidad o score.

EJEMPLOS

Spam / No Spam

Fraude / No Fraude

Perro / Gato

Hombre / Mujer

Clasificador

- **Problema de clasificación:** Clasificar un objeto x en un conjunto de clases {w1,...,wn} en función de atributos {x1,...,xn}.
- **Entrada:** atributos {x1,...,xn}
- **Salida:** clase w estimada a partir de {x1,...,xn}, probabilidad P(w|x1,...,xn) (no siempre se estima)

- **Entrenamiento:** El clasificador se construye mediante un conjunto de entrenamiento.
- **Testeo:** con un conjunto de testeo se pueden estimar los errores cometidos.

Medidas de Performance

- VP: Verdadero Positivo (TP)
- VN: Verdadero Negativo (TN)
- FN: Falso Negativo
- FP: Falso Positivo

$$Sensibilidad = \frac{VP}{VP + FN} \quad Especificidad = \frac{VN}{VN + FP}$$

Ejemplo: Detección de Fraude

	Predicted type		
	Legitimate	Fraudulent	Total
Observed type			
Legitimate	12,302	377	12,679
Fraudulent	582	1838	2420
Total	12,884	2215	15,099

Predicted type was fraud for estimated probability of fraud exceeding 0.5.

MEDIDAS DE PERFORMANCE:

Sensibilidad = VP/(VP+FN) = 1838 / 2420 = **75,95%**

Especificidad = VN/(VN+FP) = 12302 / 12679 = **97,03%**

% Corr. Clas. = (12302+1838)/15099 = **93,64%**

Employing transaction aggregation strategy to detect credit card fraud, Jha, Guillen, Westland, Expert Systems with Applications 39 (2012)

	Predicted type		
	Legitimate	Fraudulent	Total
Observed type			
Legitimate	12,302	377	12,679
Fraudulent	582	1838	2420
Total	12,884	2215	15,099

Predicted type was fraud for estimated probability of fraud exceeding 0.5.

Observación:

Sensibilidad = VP/(VP+FN) = VP/P, Especificidad = VN/(VN+FP) = VN/N

Si P = 100 y N = 100.000 entonces VP = 75,95% * 100 = 76 **PERO** VN = 97,03% * 100.000 = 97.030 y FP = 100.000 – 97.030 = 2.970.

Por cada VP hay 39 FP! (FP:VP= 39:1)

COSTOS

- Probabilidad de error
- Costo de los errores
- Asimetría en los costos (FP / FN / %)
- Costos operativos
- Costos de reputación

Clasificación

Vecinos más cercanos

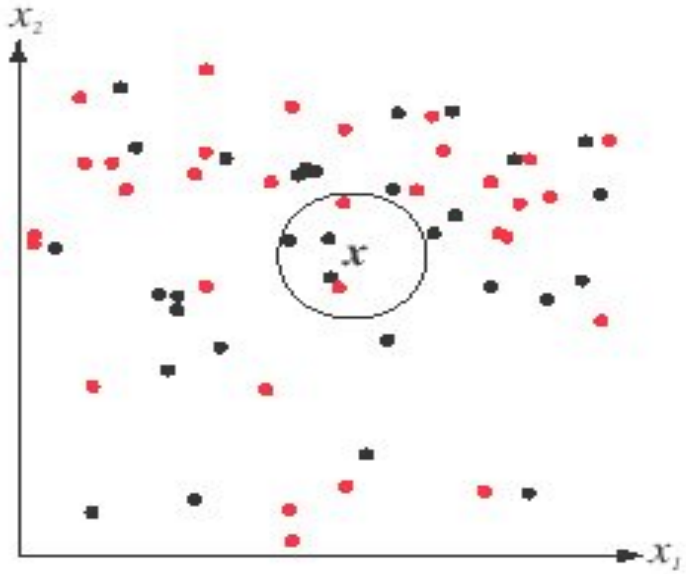


FIGURE 4.15. The k-nearest-neighbor query starts at the test point **x** and grows a spherical region until it encloses k training samples, and it labels the test point by a majority vote of these samples. In this $k = 5$ case, the test point **x** would be labeled the category of the black points. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Árboles de Decisión

Árboles de Decisión

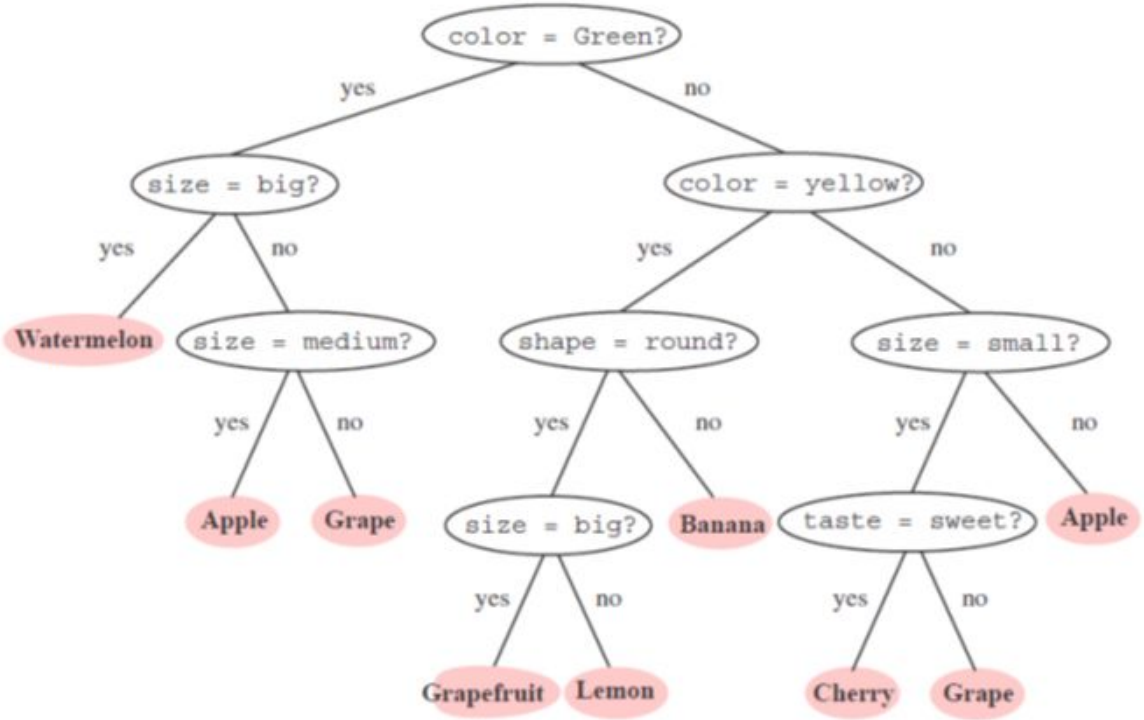
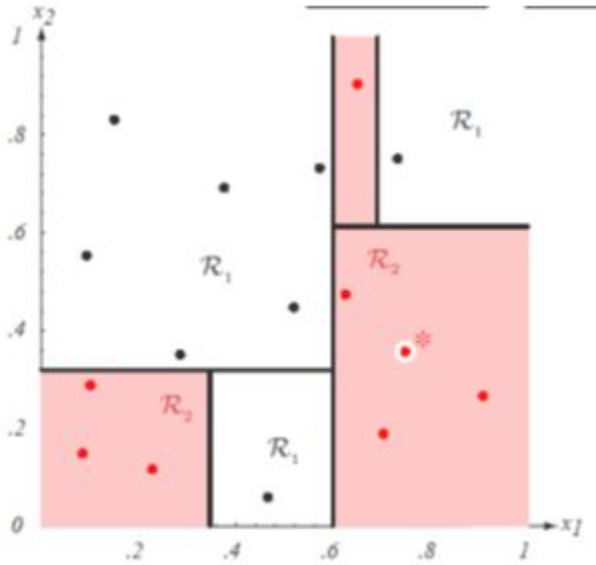
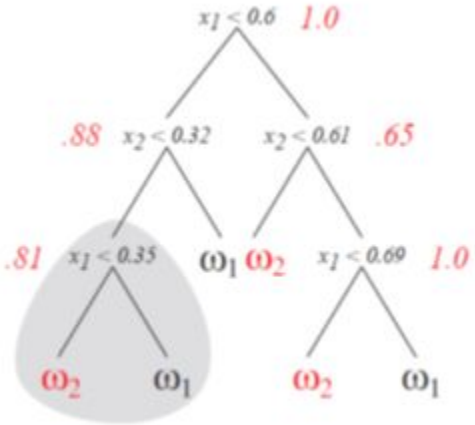


Fig. extraída de Pattern Classification de Duda, Hart & Stork.

Árboles de Decisión





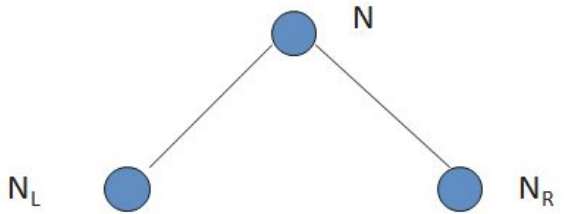
ω_1 (black)	
x_1	x_2
.15	.83
.09	.55
.29	.35
.38	.70
.52	.48
.57	.73
.73	.75
.47	.06

ω_2 (red)	
x_1	x_2
.10	.29
.08	.15
.23	.16
.70	.19
.62	.47
.91	.27
.65	.90
.75	.36* (.32†)

Fig. extraída de Pattern Classification de Duda, Hart & Stork.

Árboles de Decisión: Impureza de la división

- Para poder utilizar la impureza al momento de testear la calidad de la división de un nodo es necesario poder estimar la impureza de los nuevos nodos N_L y N_R.
- Si P_L es la proporción de elementos que van a N_L entonces:



$$i(N_L, N_R) = P_L i(N_L) + (1 - P_L) i(N_R)$$

Árboles de Decisión: Ejercicio