Aprendizaje de Conceptos
Introducción a los Árboles de Decisión

Miguel A. Gutiérrez Naranjo

Departamento de Ciencias de la Computación e Inteligencia Artificial
Universidad de Sevilla

3 de abril, 2017

¿Qué es aprender?

Definiciones de *aprendizaje*:

- Cualquier *cambio* en un sistema que le permita realizar la misma tarea de manera *más eficiente* la próxima vez (*H. Simon*)
- *Modificar la representación* del mundo que se está percibiendo (*R. Michalski*)
- Realizar cambios *útiles* en nuestras mentes (*M. Minsky*)
- Se dice que aprendemos de la *experiencia* a realizar alguna *tarea* si la realización de la tarea mejora con la experiencia respecto a alguna *medida de rendimiento* (*T. M. Mitchell*)

¿Qué es aprender?

Definiciones de *aprendizaje*:

- Cualquier *cambio* en un sistema que le permita realizar la misma tarea de manera *más eficiente* la próxima vez (*H. Simon*)
- *Modificar la representación* del mundo que se está percibiendo (*R. Michalski*)
- Realizar cambios *útiles* en nuestras mentes (*M. Minsky*)
- Se dice que aprendemos de la *experiencia* a realizar alguna *tarea* si la realización de la tarea mejora con la experiencia respecto a alguna *medida de rendimiento* (*T. M. Mitchell*)

¿Qué es aprender?

Definiciones de *aprendizaje*:

- Cualquier *cambio* en un sistema que le permita realizar la misma tarea de manera *más eficiente* la próxima vez (*H. Simon*)
- *Modificar la representación* del mundo que se está percibiendo (*R. Michalski*)
- Realizar cambios *útiles* en nuestras mentes (*M. Minsky*)
- Se dice que aprendemos de la *experiencia* a realizar alguna *tarea* si la realización de la tarea mejora con la experiencia respecto a alguna *medida de rendimiento* (*T. M. Mitchell*)

¿Qué es aprender?

Definiciones de *aprendizaje*:

- Cualquier *cambio* en un sistema que le permita realizar la misma tarea de manera *más eficiente* la próxima vez (*H. Simon*)
- *Modificar la representación* del mundo que se está percibiendo (*R. Michalski*)
- Realizar cambios *útiles* en nuestras mentes (*M. Minsky*)
- Se dice que aprendemos de la *experiencia* a realizar alguna *tarea* si la realización de la tarea mejora con la experiencia respecto a alguna *medida de rendimiento* (*T. M. Mitchell*)

Aprendizaje Automático

- Aprender a conocer y gestionar emociones.
- Aprender a atarse los zapatos.
- Aprender habilidades sociales.
- ...
- Aprendizaje de conceptos.

Aprendizaje Automático

- Aprender a conocer y gestionar emociones.
- Aprender a atarse los zapatos.
- Aprender habilidades sociales.
- ...
- Aprendizaje de conceptos.

A Neural Algorithm of Artistic Style

Leon A. Gatys,[1,2,3*] Alexander S. Ecker,[1,2,4,5] Matthias Bethge[1,2,4]

[1]Werner Reichardt Centre for Integrative Neuroscience
and Institute of Theoretical Physics, University of Tübingen, Germany
[2]Bernstein Center for Computational Neuroscience, Tübingen, Germany

https://arxiv.org/pdf/1508.06576v1.pdf

A Neural Algorithm of Artistic Style

Imagen del *Neckarfront* en Tübingen, Alemania. La misma imagen al estilo de *El Hundimiento del Minotauro* de J.M.W. Turner, 1805; de *La noche estrellada* de V. van Gogh, 1889; y de *El grito* de E. Munch, 1893.

Otro ejemplo de Aprendizaje

29 de Octubre, 2012

Noticias > Ciencia y tecnología

Imprimir

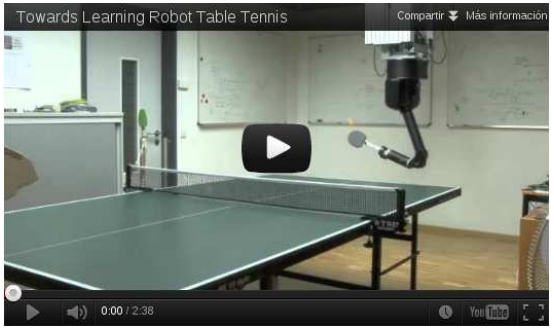
Un robot capaz de aprender a jugar al ping-pong

- Un operador enseña a un brazo articulado los movimientos básicos del juego
- Una vez ha aprendido, puede devolver casi el 90% de las pelotas
- El sistema cuenta con cuatro cámaras y es capaz de improvisar

Recomendar 38 personas han recomendado esto. Sign Up recomiendan tus amigos. Twittear 18 +1 0 Enviar

ÁLVARO IBÁÑEZ 'ALVY' (ESPECIAL PARA RTVE.ES)

29.10.2012



Aprendizaje de conceptos

Un concepto es el conjunto de todas sus instancias.
G.W. Leibniz

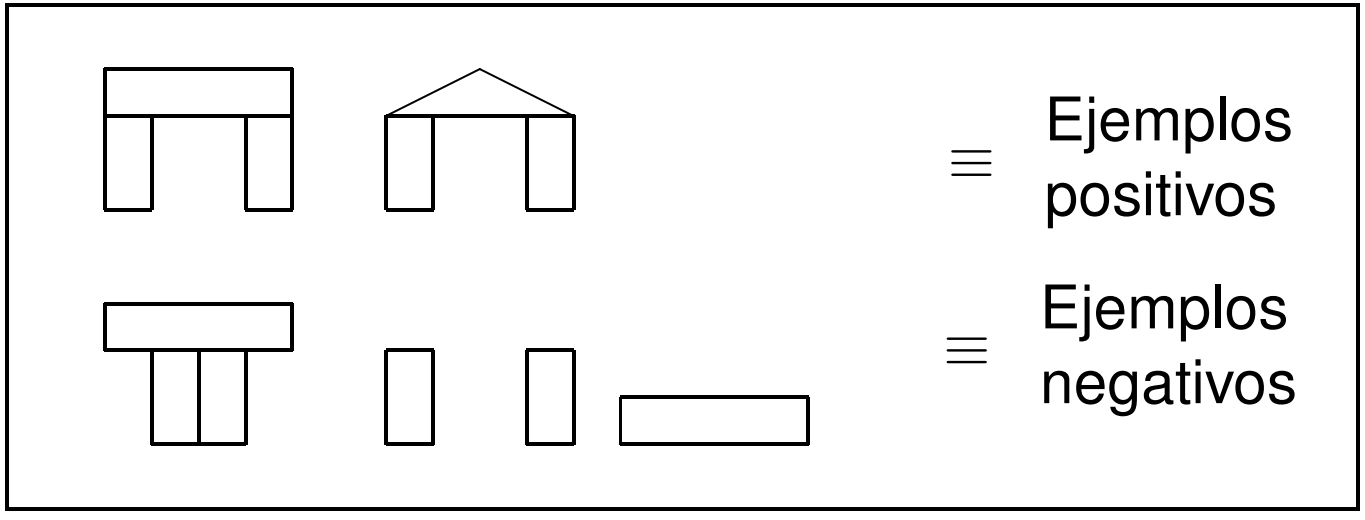
- El estilo de Van Gogh es el conjunto de sus cuadros.
- La humanidad es el conjunto de todos los hombres.
- La relación mayor_que definida sobre el conjunto de los números reales es el conjunto de pares (x, y) tales que x es mayor que y.

Gottfried Wilhelm Leibniz

Aprendizaje

ARCHES - P. Winston 1975

Ejemplos

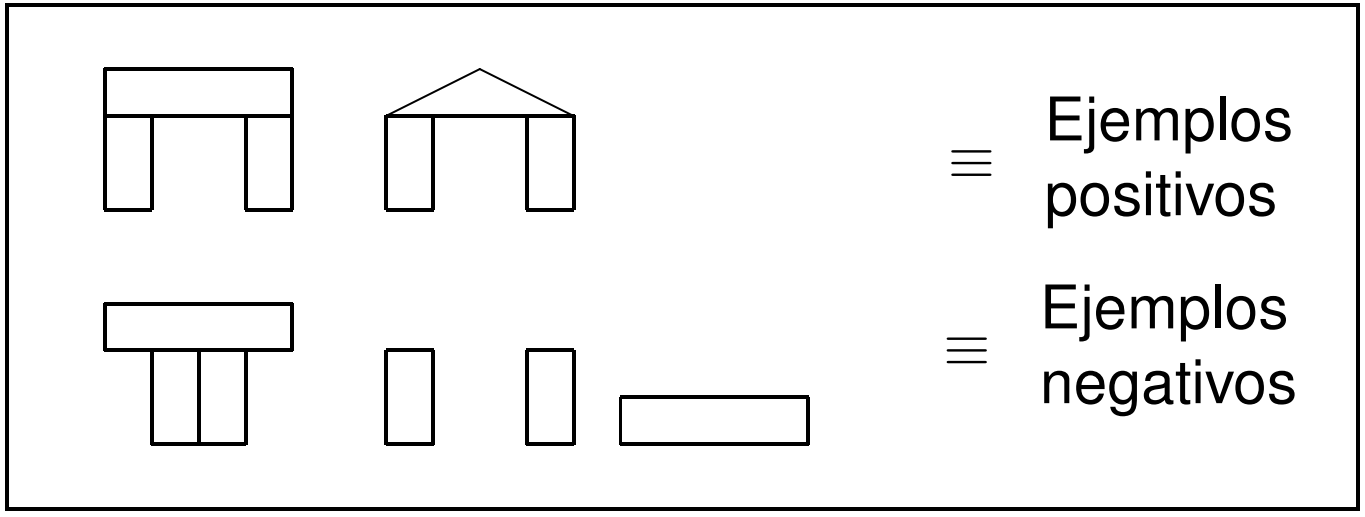


Aprendizaje

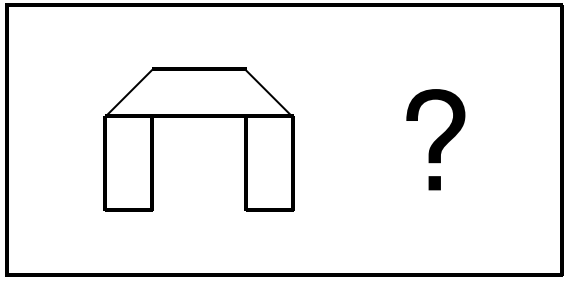
Aprendizaje

ARCHES - P. Winston 1975

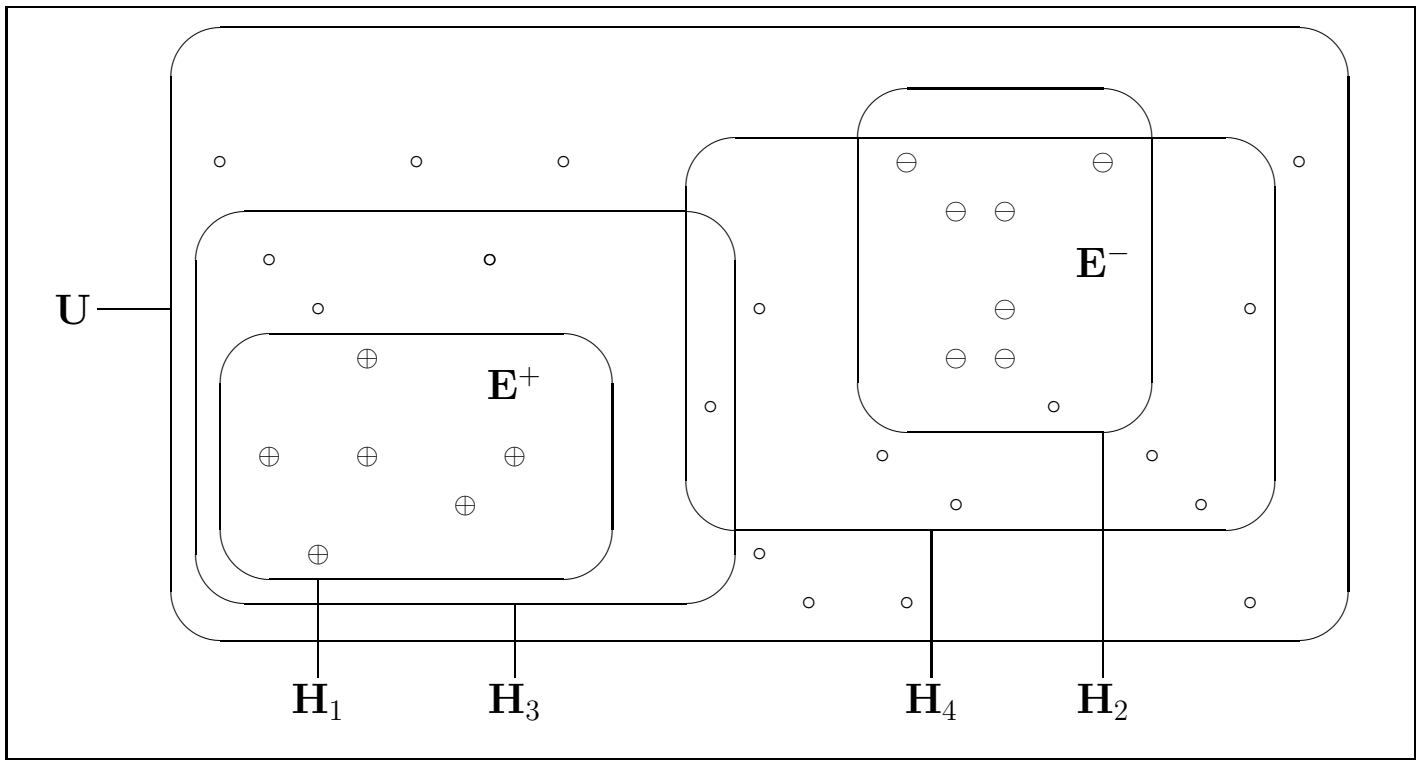
Ejemplos



Aprendizaje



Gráficamente



Un ejemplo

- $\mathbf{E}^+ = \{2, 4, 6, 8\}$
- $\mathbf{E}^- = \{11, 17\}$

- $\mathbf{H}_1 = \mathbf{E}^+$
- $\mathbf{H}_2 = \mathbb{N} - \mathbf{E}^-$
- $\mathbf{H}_3 = \{n \in \mathbb{N} \mid n \text{ es par } \}$
- $\mathbf{H}_4 = \{n \in \mathbb{N} \mid n \leq 10\}$

Un ejemplo

- $\mathbf{E}^+ = \{2, 4, 6, 8\}$
- $\mathbf{E}^- = \{11, 17\}$

- $\mathbf{H}_1 = \mathbf{E}^+$
- $\mathbf{H}_2 = \mathbb{N} - \mathbf{E}^-$
- $\mathbf{H}_3 = \{n \in \mathbb{N} \mid n \text{ es par } \}$
- $\mathbf{H}_4 = \{n \in \mathbb{N} \mid n \leq 10\}$

Inductivismo ingenuo

- El aprendizaje comienza con la observación.
- El observador usa sus sentidos, debe registrar su observación de modo fidedigno y debe hacerlo sin prejuicios

- ¿Cómo podemos obtener enunciados generales a partir de enunciados singulares? Aprendizaje
- ¿Cómo podemos justificar las afirmaciones generales? ¿Qué grado de credibilidad tienen? Estadística

Inductivismo ingenuo

- El aprendizaje comienza con la observación.
- El observador usa sus sentidos, debe registrar su observación de modo fidedigno y debe hacerlo sin prejuicios

- ¿Cómo podemos obtener enunciados generales a partir de enunciados singulares? Aprendizaje
- ¿Cómo podemos justificar las afirmaciones generales? ¿Qué grado de credibilidad tienen? Estadística

Inductivismo ingenuo

- El aprendizaje comienza con la observación.
- El observador usa sus sentidos, debe registrar su observación de modo fidedigno y debe hacerlo sin prejuicios

- ¿Cómo podemos obtener enunciados generales a partir de enunciados singulares? Aprendizaje
- ¿Cómo podemos justificar las afirmaciones generales? ¿Qué grado de credibilidad tienen? Estadística

Inductivismo ingenuo

Generalización

- El número de enunciados singulares base de la generalización debe ser grande.
- Las observaciones se deben repetir en una amplia variedad de observaciones
- Ningún enunciado observacional debe entrar en contradicción con la ley universal derivada.

Inductivismo ingenuo

Principio de inducción

Si en una amplia variadad de condiciones se observa una gran cantidad de *A* y si todos los *A* observados poseen sin excepción la propiedad *B*, entonces todos los *A* tienen la propiedad *B*.

Justificación

¿Por qué el razonamiento *inductivo* conduce al conocimiento científico fiable e incluso verdadero?

Inductivismo ingenuo

Principio de inducción

Si en una amplia variadad de condiciones se observa una gran cantidad de *A* y si todos los *A* observados poseen sin excepción la propiedad *B*, entonces todos los *A* tienen la propiedad *B*.

Justificación

¿Por qué el razonamiento *inductivo* conduce al conocimiento científico fiable e incluso verdadero?

Problemas con la inducción

La paradoja de Nelson Goodman (1955)

- Un objeto (en particular, una esmeralda) es *verdul* si y sólo si es verde hasta el tiempo t, y azul a partir del tiempo t, donde t podría ser, por ejemplo, el año 3045.
- Todas las observaciones de esmeraldas verdes hechas hasta el presente, sirven tanto para apoyar la conclusión de que todas las esmeraldas son verdes, como que todas las esmeraldas son verdules.
- Las observaciones no nos permiten discriminar entre verdes y verdules. Necesitamos asumir *a priori* que hay regularidad en la naturaleza.

Problemas con la inducción

La paradoja de Nelson Goodman (1955)

- Un objeto (en particular, una esmeralda) es *verdul* si y sólo si es verde hasta el tiempo t, y azul a partir del tiempo t, donde t podría ser, por ejemplo, el año 3045.
- Todas las observaciones de esmeraldas verdes hechas hasta el presente, sirven tanto para apoyar la conclusión de que todas las esmeraldas son verdes, como que todas las esmeraldas son verdules.
- Las observaciones no nos permiten discriminar entre verdes y verdules. Necesitamos asumir *a priori* que hay regularidad en la naturaleza.

Problemas con la inducción

La paradoja de Nelson Goodman (1955)

- Un objeto (en particular, una esmeralda) es *verdul* si y sólo si es verde hasta el tiempo t, y azul a partir del tiempo t, donde t podría ser, por ejemplo, el año 3045.
- Todas las observaciones de esmeraldas verdes hechas hasta el presente, sirven tanto para apoyar la conclusión de que todas las esmeraldas son verdes, como que todas las esmeraldas son verdules.
- Las observaciones no nos permiten discriminar entre verdes y verdules. Necesitamos asumir *a priori* que hay regularidad en la naturaleza.

Problemas con la inducción

La paradoja de los cuervos. Carl Hempel (1945)

- Considérese la siguiente oración S_1

  a es un cuervo y *a* es negro.

- S_1 brinda apoyo inductivo a la siguiente generalización G_1:

  Todos los cuervos son negros.

- En lógica de primer orden, esta G_1 es equivalente a G_2:

  Todo lo que no es negro no es un cuervo.

- Cualquier cosa que no sea un cuervo y no sea negro brinda apoyo inductivo a G_2, que es equivalente a G_1.

Problemas con la inducción

La paradoja de los cuervos. Carl Hempel (1945)

- Considérese la siguiente oración S_1

  a es un cuervo y *a* es negro.

- S_1 brinda apoyo inductivo a la siguiente generalización G_1:

  Todos los cuervos son negros.

- En lógica de primer orden, esta G_1 es equivalente a G_2:

  Todo lo que no es negro no es un cuervo.

- Cualquier cosa que no sea un cuervo y no sea negro brinda apoyo inductivo a G_2, que es equivalente a G_1.

Problemas con la inducción

La paradoja de los cuervos. Carl Hempel (1945)

- Considérese la siguiente oración S_1

  a es un cuervo y *a* es negro.

- S_1 brinda apoyo inductivo a la siguiente generalización G_1:

  Todos los cuervos son negros.

- En lógica de primer orden, esta G_1 es equivalente a G_2:

  Todo lo que no es negro no es un cuervo.

- Cualquier cosa que no sea un cuervo y no sea negro brinda apoyo inductivo a G_2, que es equivalente a G_1.

Problemas con la inducción

La paradoja de los cuervos. Carl Hempel (1945)

- Considérese la siguiente oración S_1

  a es un cuervo y *a* es negro.

- S_1 brinda apoyo inductivo a la siguiente generalización G_1:

  Todos los cuervos son negros.

- En lógica de primer orden, esta G_1 es equivalente a G_2:

  Todo lo que no es negro no es un cuervo.

- Cualquier cosa que no sea un cuervo y no sea negro brinda apoyo inductivo a G_2, que es equivalente a G_1.

Inductivismo ingenuo

David Hume

- El principio de inducción tuvo éxito en la ocasión 1.
- El principio de inducción tuvo éxito en la ocasión 2.
- ...
- El principio de inducción tuvo éxito en la ocasión *n*.
- Por tanto, el principio de inducción funciona siempre

David Hume

Nuestras creencias en las teorías no son más que hábitos psicológicos que adquirimos como resultado de las repeticiones de las observaciones relevantes.

Inductivismo ingenuo

David Hume

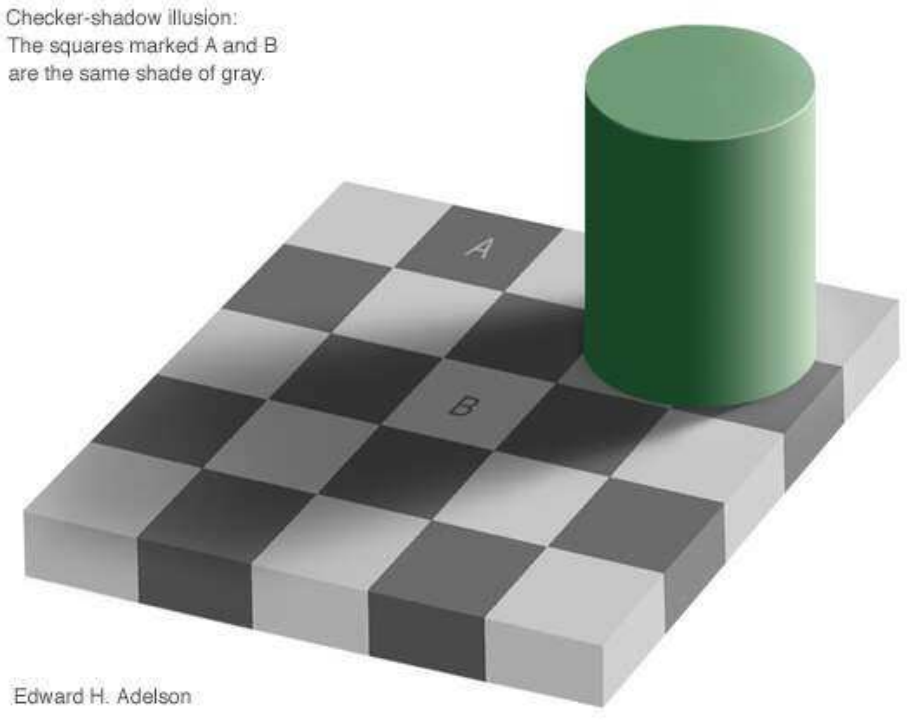
- El principio de inducción tuvo éxito en la ocasión 1.
- El principio de inducción tuvo éxito en la ocasión 2.
- ...
- El principio de inducción tuvo éxito en la ocasión n.
- Por tanto, el principio de inducción funciona siempre

David Hume

Nuestras creencias en las teorías no son más que hábitos psicológicos que adquirimos como resultado de las repeticiones de las observaciones relevantes.

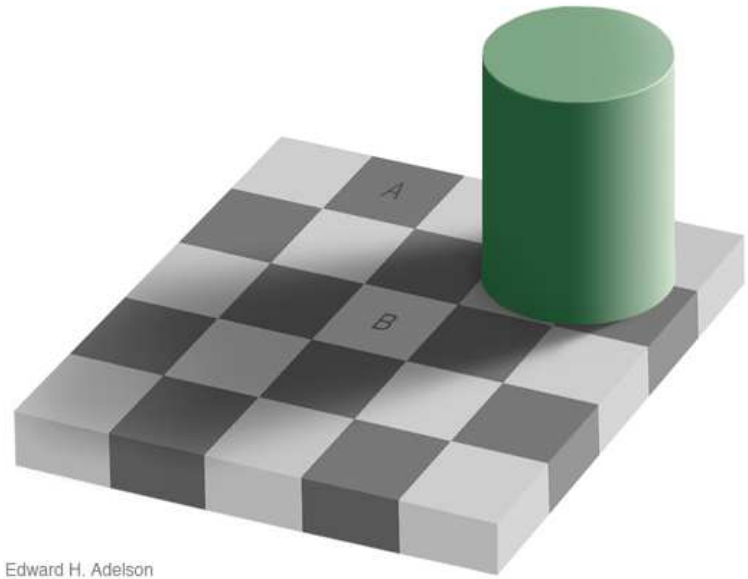
Inductivismo ingenuo

¿Podemos fiarnos de nuestros sentidos?

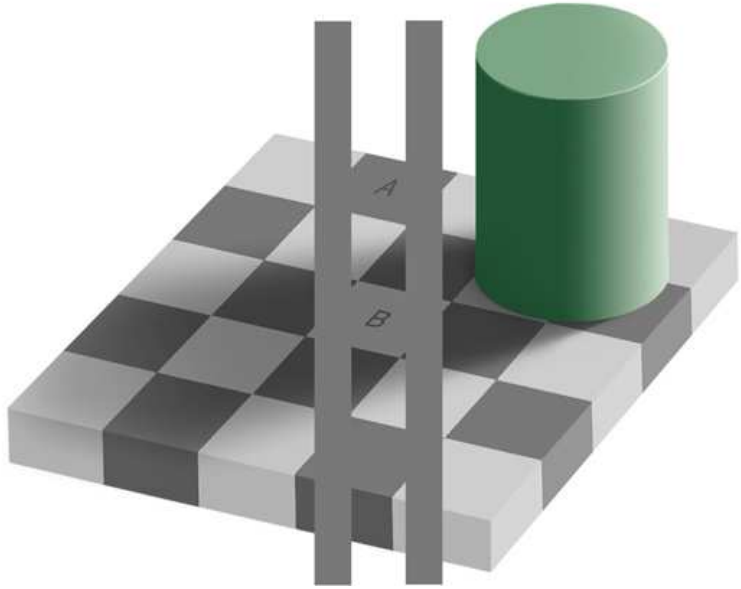


Inductivismo ingenuo

¿Podemos fiarnos de nuestros sentidos?



Edward H. Adelson



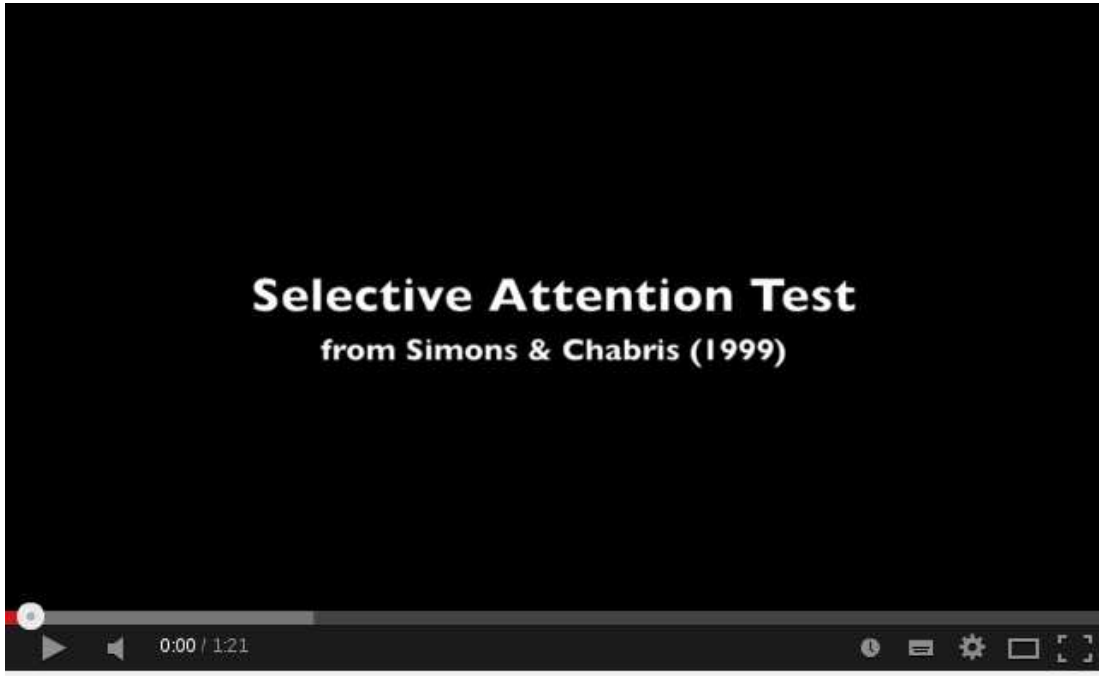
Inductivismo ingenuo

¿Podemos fiarnos de nuestros sentidos?

TRAS LEER LA
LA FRASE TE HAS DADO
CUENTA DE QUE LA
LA MENTE HUMANA
A MENUDO NO TE
INFORMA DE QUE LA
LA PALABRA "LA" SE HA
REPETIDO EN CADA
OCASIÓN.

Inductivismo ingenuo

¿Podemos fiarnos de nuestros sentidos?



Atención selectiva (Simons & Chabris, 1999)

Jean Piaget

Jean Piaget (1896-1880)

Uno no sabe lo que ve, sino que ve lo que sabe.

Inductivismo ingenuo

Lenguaje

- Necesitamos un lenguaje para expresar las observaciones.
- La elección del lenguaje es previa a la realización de la observación.
- La realidad es demasiado compleja para representarla mediante un lenguaje.
- Para representar un fenómeno, necesitamos elegir algunos atributos.

Lenguaje

En 1888, *Heinrich Hertz* efectuó un experimento eléctrico para producir y detectar por primera vez las ondas de radio predichas por la teoría electromagnética de Maxwell.

Lenguaje

¿Cómo representamos el experimento?

- Lectura de los contadores
- Presencia de chispas en el circuito eléctrico
- Dimensiones del circuito
- Dimensiones del laboratorio
- Color de las paredes
- ...

Las ondas de radio deben tener una velocidad igual a la velocidad de la luz, pero en las mediciones de Hertz, era claramente distinta. Nunca resolvió ese problema.

Lenguaje

Hasta después de su muerte no se solucionó. Las ondas de radio se reflejaban en las paredes interfiriendo en las mediciones. Las dimensiones del laboratorio sí eran muy relevantes.

El problema de la representación

Un ejemplo sencillo

- Pares atributo–valor

Cielo: Soleado, lluvioso.
Viento: Fuerte, débil,sin_viento.
Temperatura: Alta, templada, fría.
Agua: Caliente,templada, fría.
Humedad: Alta, normal, baja.
Previsión: Cambio, igual.

- Con este lenguaje podemos expresar

$$2 \times 3 \times 3 \times 3 \times 3 \times 2 = 324$$

- Por tanto hay 2^{324} conceptos posibles

Número de conceptos

- 2^{324} es un número grande, pero . . . ¿Cómo de grande?
-

Número de conceptos

- 2^{324} es un número grande, pero . . . ¿Cómo de grande?
- $2^{324} \sim 10^{97}$

Número de conceptos

- 2^{324} es un número grande, pero . . . ¿Cómo de grande?
- $2^{324} \sim 10^{97}$

En el universo conocido hay entre 10^{77} y 10^{80} átomos.

Sesgo inductivo

- No podemos considerar todos los posibles conceptos consistentes con los datos en igualdad de condiciones. Necesitamos introducir un sesgo.
- *Sesgo inductivo:* cualquier medio que el sistema de aprendizaje pueda usar para tener preferencia entre dos hipótesis consistentes con los ejemplos
- Tipos de sesgo inductivo:
  - Sesgo en el lenguaje: el lenguaje disponible para expresar las hipótesis define un espacio de hipótesis que excluye conceptos (por ejemplo, profundidad del árbol de decisión, número de literales en las reglas,...)
  - Sesgo preferencial: el algoritmo de búsqueda en el espacio de hipótesis incorpora implícitamente alguna preferencia de algunas hipótesis sobre otras.

Sesgo inductivo

- No podemos considerar todos los posibles conceptos consistentes con los datos en igualdad de condiciones. Necesitamos introducir un sesgo.
- *Sesgo inductivo:* cualquier medio que el sistema de aprendizaje pueda usar para tener preferencia entre dos hipótesis consistentes con los ejemplos
- Tipos de sesgo inductivo:
  - Sesgo en el lenguaje: el lenguaje disponible para expresar las hipótesis define un espacio de hipótesis que excluye conceptos (por ejemplo, profundidad del árbol de decisión, número de literales en las reglas,...)
  - Sesgo preferencial: el algoritmo de búsqueda en el espacio de hipótesis incorpora implícitamente alguna preferencia de algunas hipótesis sobre otras.

Sesgo inductivo

- No podemos considerar todos los posibles conceptos consistentes con los datos en igualdad de condiciones. Necesitamos introducir un sesgo.
- *Sesgo inductivo:* cualquier medio que el sistema de aprendizaje pueda usar para tener preferencia entre dos hipótesis consistentes con los ejemplos
- Tipos de sesgo inductivo:
  - Sesgo en el lenguaje: el lenguaje disponible para expresar las hipótesis define un espacio de hipótesis que excluye conceptos (por ejemplo, profundidad del árbol de decisión, número de literales en las reglas,...)
  - Sesgo preferencial: el algoritmo de búsqueda en el espacio de hipótesis incorpora implícitamente alguna preferencia de algunas hipótesis sobre otras.

Sesgo inductivo

- No podemos considerar todos los posibles conceptos consistentes con los datos en igualdad de condiciones. Necesitamos introducir un sesgo.
- *Sesgo inductivo:* cualquier medio que el sistema de aprendizaje pueda usar para tener preferencia entre dos hipótesis consistentes con los ejemplos
- Tipos de sesgo inductivo:
  - Sesgo en el lenguaje: el lenguaje disponible para expresar las hipótesis define un espacio de hipótesis que excluye conceptos (por ejemplo, profundidad del árbol de decisión, número de literales en las reglas,...)
  - Sesgo preferencial: el algoritmo de búsqueda en el espacio de hipótesis incorpora implícitamente alguna preferencia de algunas hipótesis sobre otras.

Sesgo inductivo

- No podemos considerar todos los posibles conceptos consistentes con los datos en igualdad de condiciones. Necesitamos introducir un sesgo.
- *Sesgo inductivo:* cualquier medio que el sistema de aprendizaje pueda usar para tener preferencia entre dos hipótesis consistentes con los ejemplos
- Tipos de sesgo inductivo:
  - Sesgo en el lenguaje: el lenguaje disponible para expresar las hipótesis define un espacio de hipótesis que excluye conceptos (por ejemplo, profundidad del árbol de decisión, número de literales en las reglas,...)
  - Sesgo preferencial: el algoritmo de búsqueda en el espacio de hipótesis incorpora implícitamente alguna preferencia de algunas hipótesis sobre otras.

Necesidad de sesgo

- Teorema:
- Sea X un universo y D un conjunto de entrenamiento sobre X. Sean $E^+ = \{x \in X \mid (x, 1) \in D\}$ y $E^- = \{x \in X \mid (x, 0) \in D\}$. Sea H el conjunto de conceptos que contiene a *todos* los conceptos sobre X ($H = \mathcal{P}(X) = 2^X$). Sea VS (espacio de versiones) el subconjunto de H formado por los conceptos consistentes con D. Sea $x_0 \in X$ tal que $x_0 \notin E^+ \cup E^-$ y sea $h \in VS$ tal que $h(x_0) = 0$. Entonces existe $h' \in VS$ tal que $h'(x_0) = 1$.

Necesidad de sesgo

- Demostración:
- Sea h' la siguiente hipótesis

$$h'(x) = \begin{cases} 1 & \text{si } x = x_0 \\ h(x) & \text{si } x \in X - \{x_0\} \end{cases}$$

Por definición, $h'(x_0) = 1$. Por tanto sólo hay que ver que $h' \in VS$. Puesto que en H están todas las hipótesis, $h' \in H$. Además, para todo $x \neq x_0$, tenemos que $h'(x) = h(x)$. En particular para todo $x \in E^+ \cup E^-$ tenemos que $h'(x) = h(x)$. Por tanto, si $h \in VS$ entonces $h' \in VS$.

Árboles de decisión

- Pares atributo–valor

Cielo	Temperatura	Humedad	Viento	Agua	Previsión	HacerDeporte
Soleado	Templada	Normal	Fuerte	Templada	Igual	Sí
Soleado	Templada	Alta	Fuerte	Templada	Igual	Sí
Lluvia	Fría	Alta	Fuerte	Templada	Cambio	No
Soleado	Templada	Alta	Fuerte	Fría	Cambio	Sí
...	...	...	...	...	...	...

- Un árbol de decisión es un grafo etiquetado que representa un concepto.

Árboles de decisión

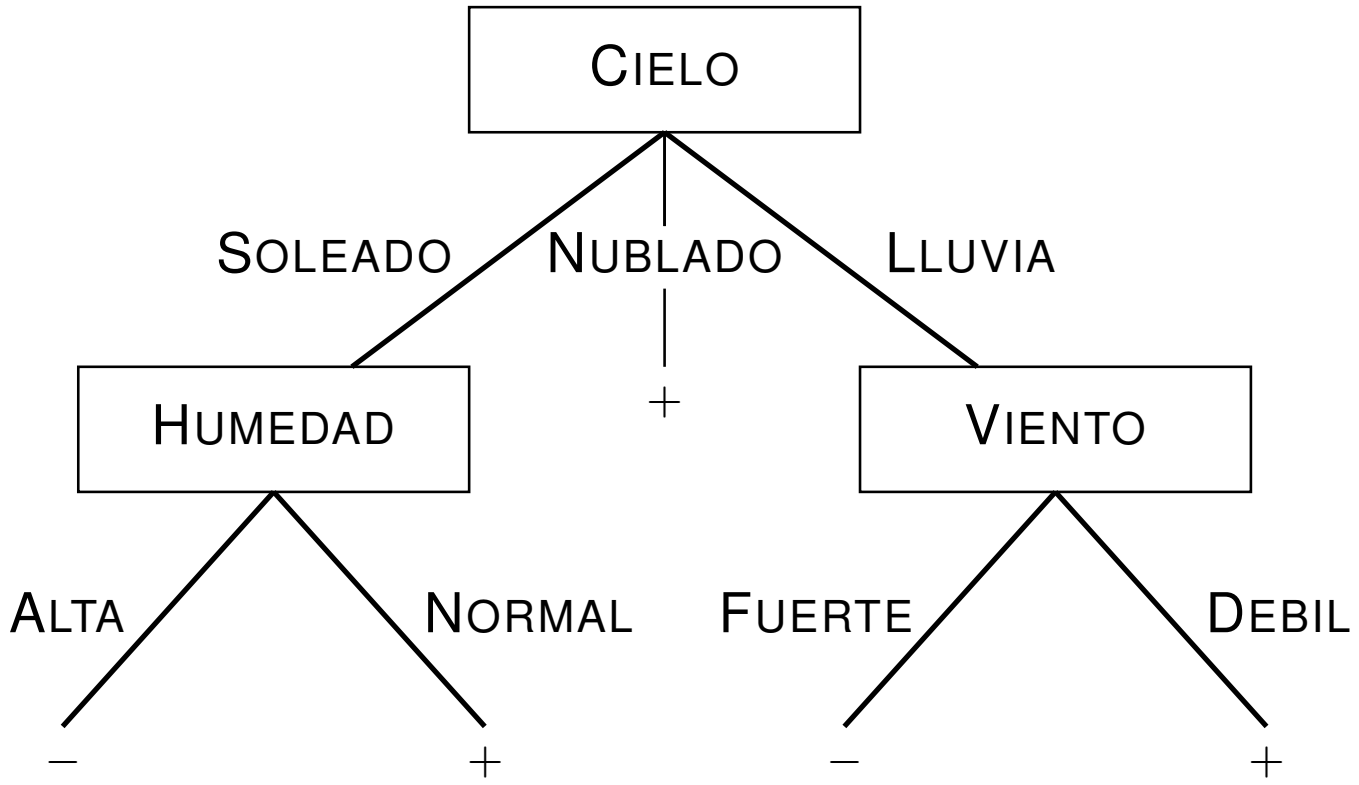
- Pares atributo–valor

Cielo	Temperatura	Humedad	Viento	Agua	Previsión	HacerDeporte
Soleado	Templada	Normal	Fuerte	Templada	Igual	Sí
Soleado	Templada	Alta	Fuerte	Templada	Igual	Sí
Lluvia	Fría	Alta	Fuerte	Templada	Cambio	No
Soleado	Templada	Alta	Fuerte	Fría	Cambio	Sí
...	...	...	...	...	...	...

- Un árbol de decisión es un grafo etiquetado que representa un concepto.

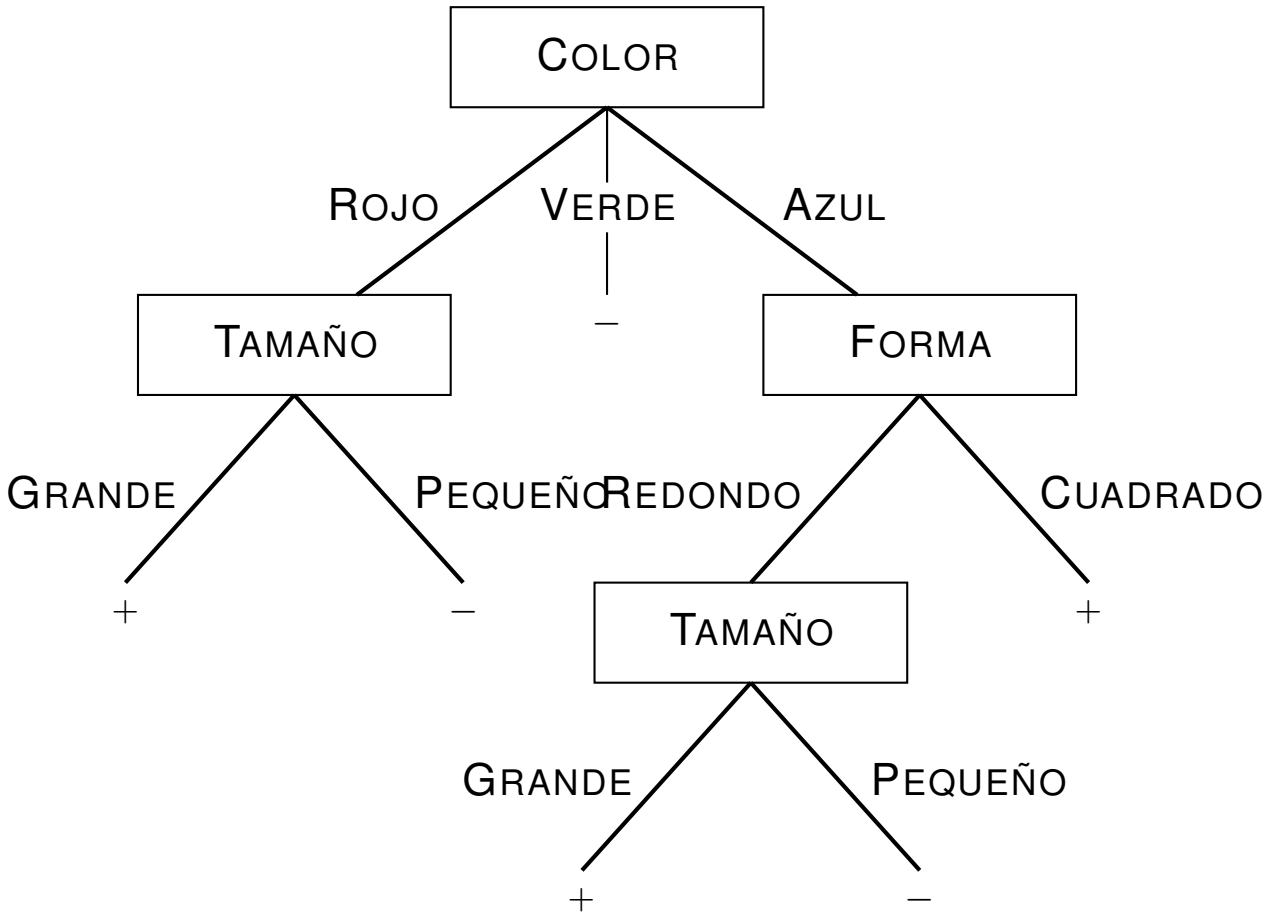
Árboles de decisión

- Ejemplos de árboles de decisión



Árboles de decisión

- Ejemplos de árboles de decisión



Árboles de decisión

$$D = [34^{+}, 27^{-}]$$

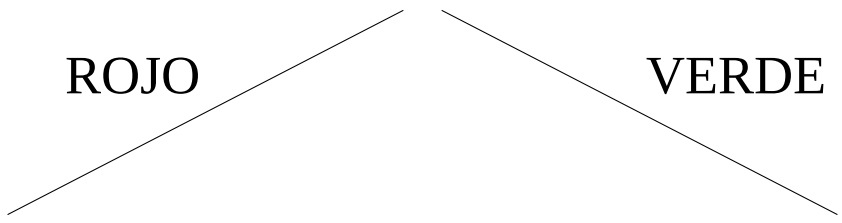
Árboles de decisión

$$D = [34^{+}, 27^{-}]$$

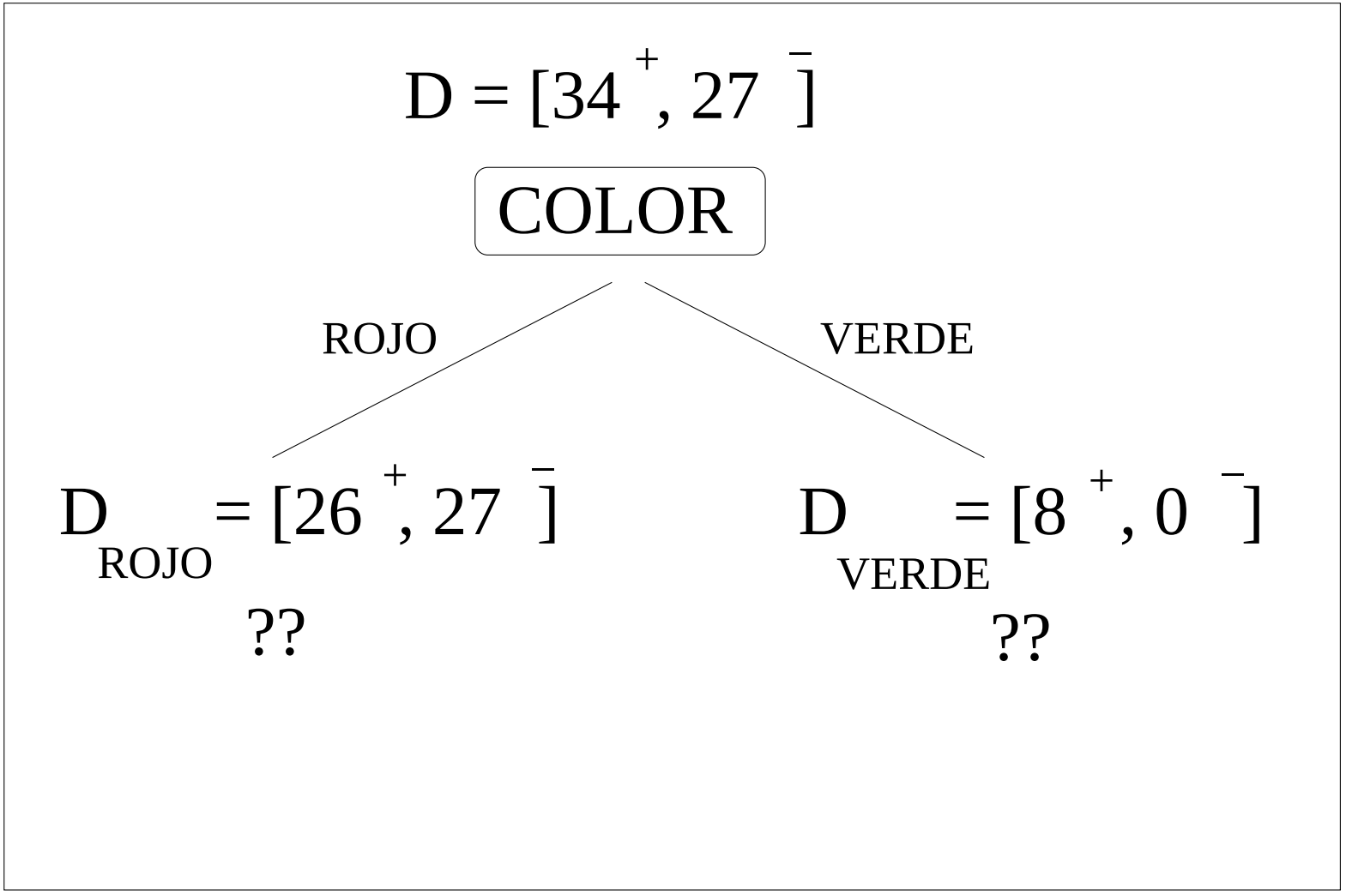
COLOR

Árboles de decisión

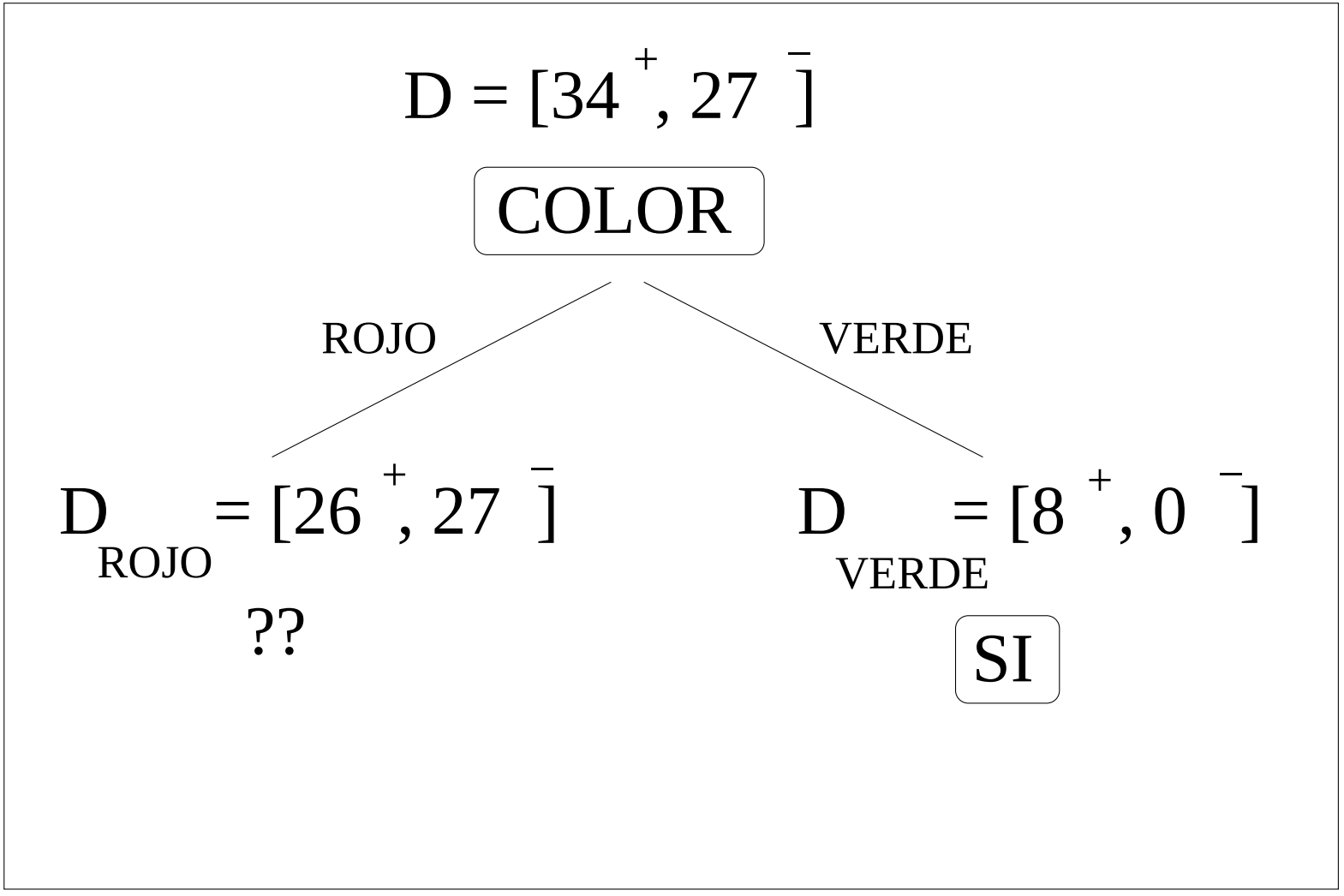
$$D = [34^{+}, 27^{-}]$$



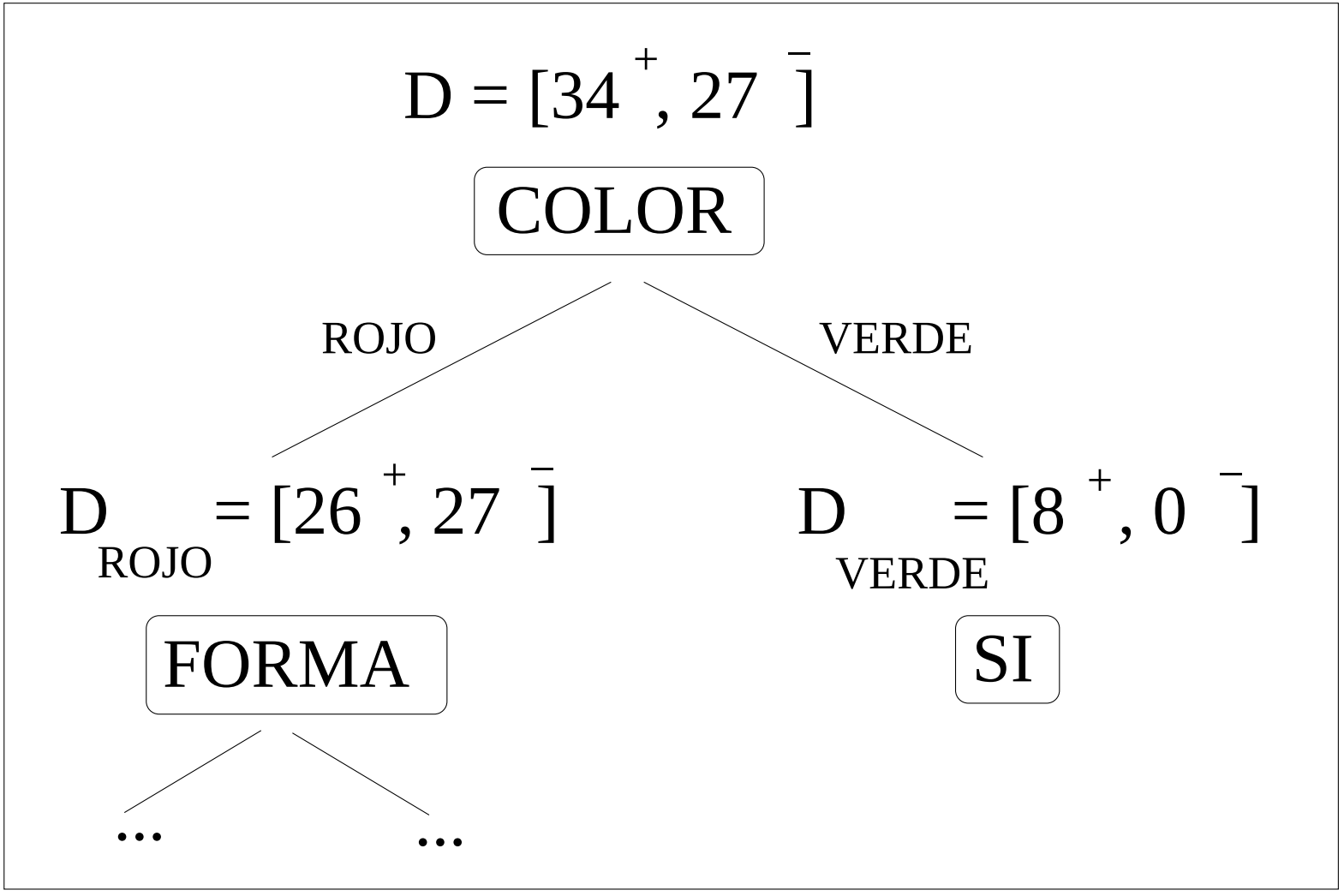
Árboles de decisión



Árboles de decisión



Árboles de decisión



Algoritmo ID3

ID3(Ejemplos, Atributo-objetivo, Atributos)

```
1. Si todos los Ejemplos son positivos, devolver un nodo etiquetado con +
2. Si todos los Ejemplos son negativos, devolver un nodo etiquetado con -
3. Si Atributos está vacío, devolver un nodo etiquetado con el valor más
   frecuente de Atributo-objetivo en Ejemplos.
4. En otro caso:
4.1. Sea A el atributo de Atributos que MEJOR clasifica Ejemplos
4.2. Crear Árbol, con un nodo etiquetado con A.
4.3. Para cada posible valor v de A, hacer:
   * Añadir un arco a Árbol, etiquetado con v.
   * Sea Ejemplos(v) el subconjunto de Ejemplos con valor
    del atributo A igual a v.
   * Si Ejemplos(v) es vacío:
    - Entonces colocar debajo del arco anterior un nodo etiquetado con
     el valor más frecuente de Atributo-objetivo en Ejemplos.
    - Si no, colocar debajo del arco anterior el subárbol
     ID3(Ejemplos(v), Atributo-objetivo, Atributos-A).

4.4 Devolver Árbol
```

Árboles de decisión

- Árboles de decisión
  - Nodos interiores: atributos
  - Arcos: posibles valores del nodo origen
  - Hojas: valor de clasificación (usualmente $+$ ó $-$, aunque podría ser cualquier conjunto de valores, no necesariamente binario)

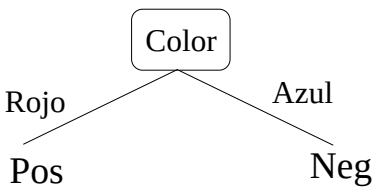
- Capaz de representar cualquier concepto.
- Necesitamos sesgo.

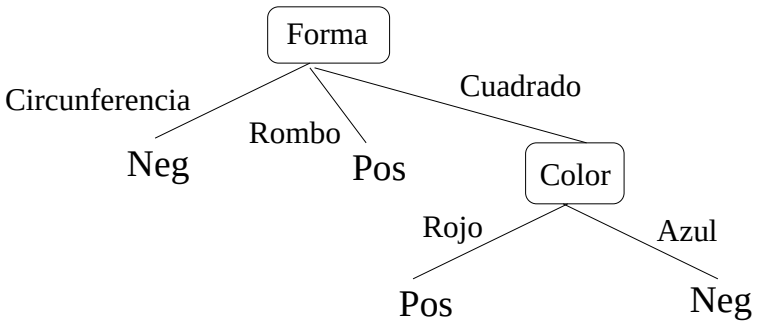
Árboles de decisión

- Árboles de decisión
  - Nodos interiores: atributos
  - Arcos: posibles valores del nodo origen
  - Hojas: valor de clasificación (usualmente $+$ ó $-$, aunque podría ser cualquier conjunto de valores, no necesariamente binario)
- Capaz de representar cualquier concepto.
- Necesitamos sesgo.

Clasificadores

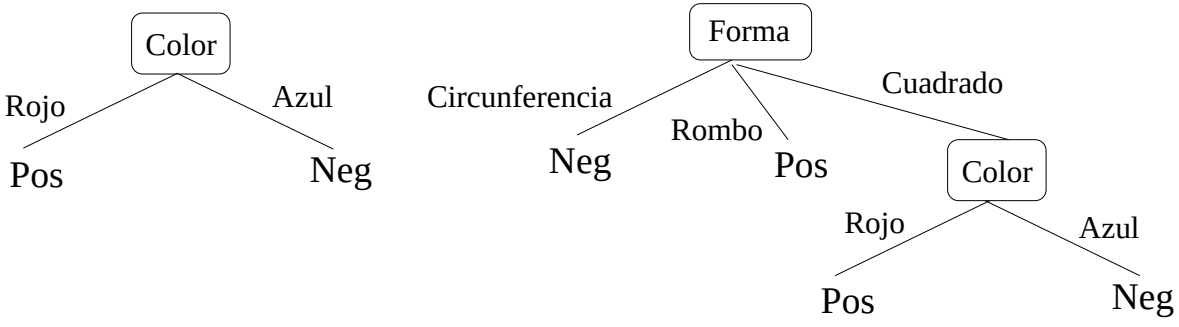
Color	Forma	Clasificación
Rojo	Cuadrado	Pos
Rojo	Rombo	Pos
Azul	Circunferencia	Neg
Azul	Cuadrado	Neg





Clasificadores

Color	Forma	Clasificación
Rojo	Cuadrado	Pos
Rojo	Rombo	Pos
Azul	Circunferencia	Neg
Azul	Cuadrado	Neg



$\langle Azul, Rombo \rangle$???

Clasificadores

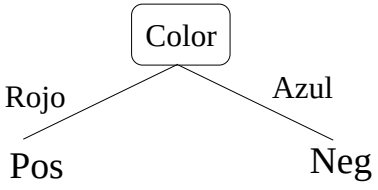
	Cuadrado	Circunferencia	Rombo
Rojo	$\oplus$		$\oplus$
Azul	$\ominus$	$\ominus$	

(1)

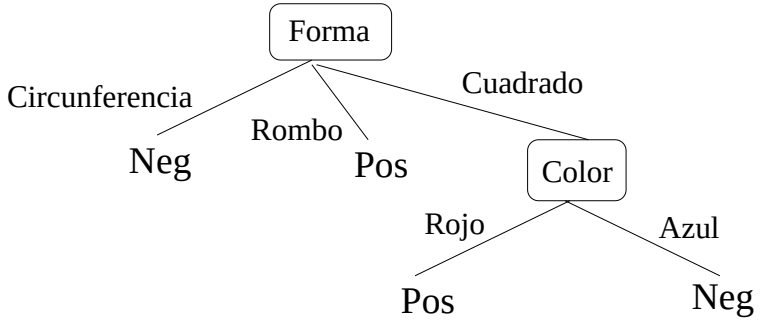
	Cuadrado	Circunferencia	Rombo
Rojo	$\oplus$		$\oplus$
Azul	$\ominus$	$\ominus$	

(2)

(1)



(2)



$\langle Azul, Rombo \rangle$???

Necesidad de sesgo

- Todos las hipótesis son expresables (i.e., no hay sesgo en el lenguaje).
- Dado un conjunto de entrenamiento, el espacio de versiones *VS* es el conjunto de árboles que lo clasifica correctamente.
- Si el conjunto de entrenamiento no contiene errores, siempre podremos obtener un árbol de decisión que clasifique correctamente todos los ejemplos del conjunto de entrenamiento.
- Según vimos, puesto que *todas* las hipótesis son expresables, dada una instancia que no pertenezca al conjunto de entrenamiento, la mitad de los árboles la clasificará correctamente (con clasificación booleana).
- ¿Qué árbol elegimos?

Necesidad de sesgo

- Todos las hipótesis son expresables (i.e., no hay sesgo en el lenguaje).
- Dado un conjunto de entrenamiento, el espacio de versiones *VS* es el conjunto de árboles que lo clasifica correctamente.
- Si el conjunto de entrenamiento no contiene errores, siempre podremos obtener un árbol de decisión que clasifique correctamente todos los ejemplos del conjunto de entrenamiento.
- Según vimos, puesto que *todas* las hipótesis son expresables, dada una instancia que no pertenezca al conjunto de entrenamiento, la mitad de los árboles la clasificará correctamente (con clasificación booleana).
- ¿Qué árbol elegimos?

Necesidad de sesgo

- Todos las hipótesis son expresables (i.e., no hay sesgo en el lenguaje).
- Dado un conjunto de entrenamiento, el espacio de versiones *VS* es el conjunto de árboles que lo clasifica correctamente.
- Si el conjunto de entrenamiento no contiene errores, siempre podremos obtener un árbol de decisión que clasifique correctamente todos los ejemplos del conjunto de entrenamiento.
- Según vimos, puesto que *todas* las hipótesis son expresables, dada una instancia que no pertenezca al conjunto de entrenamiento, la mitad de los árboles la clasificará correctamente (con clasificación booleana).
- ¿Qué árbol elegimos?

Necesidad de sesgo

- Todos las hipótesis son expresables (i.e., no hay sesgo en el lenguaje).
- Dado un conjunto de entrenamiento, el espacio de versiones *VS* es el conjunto de árboles que lo clasifica correctamente.
- Si el conjunto de entrenamiento no contiene errores, siempre podremos obtener un árbol de decisión que clasifique correctamente todos los ejemplos del conjunto de entrenamiento.
- Según vimos, puesto que *todas* las hipótesis son expresables, dada una instancia que no pertenezca al conjunto de entrenamiento, la mitad de los árboles la clasificará correctamente (con clasificación booleana).
- ¿Qué árbol elegimos?

Necesidad de sesgo

- Todos las hipótesis son expresables (i.e., no hay sesgo en el lenguaje).
- Dado un conjunto de entrenamiento, el espacio de versiones *VS* es el conjunto de árboles que lo clasifica correctamente.
- Si el conjunto de entrenamiento no contiene errores, siempre podremos obtener un árbol de decisión que clasifique correctamente todos los ejemplos del conjunto de entrenamiento.
- Según vimos, puesto que *todas* las hipótesis son expresables, dada una instancia que no pertenezca al conjunto de entrenamiento, la mitad de los árboles la clasificará correctamente (con clasificación booleana).
- ¿Qué árbol elegimos?

La navaja de Occam

Guillermo de Occam (1288-1349)

Lex parsimoniae

Entia non sunt multiplicanda praeter necessitatem (No ha de presumirse la existencia de más cosas que las absolutamente necesarias)

Guillermo de Occam

La navaja de Occam

En igualdad de condiciones la solución más sencilla es probablemente la correcta

La navaja de Occam

- Queremos el árbol *más pequeño*.
- *Más pequeño* significa que el número de preguntas que hay que realizar para decidir la clasificación de una instancia sea la menor posible. Técnicamente, queremos el árbol de menor profundidad posible.
- ¿Cómo lo encontramos?
- **Idea:** Generamos todo el espacio de versiones y nos quedamos con el árbol de menor profundidad.
- Quizá no sea tan buena idea . . .

La navaja de Occam

- Queremos el árbol *más pequeño*.
- *Más pequeño* significa que el número de preguntas que hay que realizar para decidir la clasificación de una instancia sea la menor posible. Técnicamente, queremos el árbol de menor profundidad posible.
- ¿Cómo lo encontramos?
- **Idea:** Generamos todo el espacio de versiones y nos quedamos con el árbol de menor profundidad.
- Quizá no sea tan buena idea . . .

La navaja de Occam

- Queremos el árbol *más pequeño*.
- *Más pequeño* significa que el número de preguntas que hay que realizar para decidir la clasificación de una instancia sea la menor posible. Técnicamente, queremos el árbol de menor profundidad posible.
- ¿Cómo lo encontramos?
- **Idea:** Generamos todo el espacio de versiones y nos quedamos con el árbol de menor profundidad.
- Quizá no sea tan buena idea . . .

La navaja de Occam

- Queremos el árbol *más pequeño*.
- *Más pequeño* significa que el número de preguntas que hay que realizar para decidir la clasificación de una instancia sea la menor posible. Técnicamente, queremos el árbol de menor profundidad posible.
- ¿Cómo lo encontramos?
- **Idea:** Generamos todo el espacio de versiones y nos quedamos con el árbol de menor profundidad.
- Quizá no sea tan buena idea . . .

La navaja de Occam

- Queremos el árbol *más pequeño*.
- *Más pequeño* significa que el número de preguntas que hay que realizar para decidir la clasificación de una instancia sea la menor posible. Técnicamente, queremos el árbol de menor profundidad posible.
- ¿Cómo lo encontramos?
- **Idea:** Generamos todo el espacio de versiones y nos quedamos con el árbol de menor profundidad.
- Quizá no sea tan buena idea . . .

Algoritmo ID3

```
ID3(Ejemplos, Atributo-objetivo, Atributos)
```

```
1. Si todos los Ejemplos son positivos, devolver un nodo etiquetado con +
2. Si todos los Ejemplos son negativos, devolver un nodo etiquetado con -
3. Si Atributos está vacío, devolver un nodo etiquetado con el valor más
   frecuente de Atributo-objetivo en Ejemplos.
4. En otro caso:
4.1. Sea A el atributo de Atributos que MEJOR clasifica Ejemplos
4.2. Crear Árbol, con un nodo etiquetado con A.
4.3. Para cada posible valor v de A, hacer:
   * Añadir un arco a Árbol, etiquetado con v.
   * Sea Ejemplos(v) el subconjunto de Ejemplos con valor
     del atributo A igual a v.
   * Si Ejemplos(v) es vacío:
     - Entonces colocar debajo del arco anterior un nodo etiquetado con
       el valor más frecuente de Atributo-objetivo en Ejemplos.
     - Si no, colocar debajo del arco anterior el subárbol
       ID3(Ejemplos(v), Atributo-objetivo, Atributos-A).

4.4 Devolver Árbol
```

- ¿Cómo elegimos el **mejor** atributo?

¿Qué es la entropía?

Es un concepto transversal en ciencia.

- En termodinámica, mide el número de microestados compatibles con el macroestado de equilibrio.
- En informática es la aleatoriedad recogida por un sistema operativo.
- En teoría de la información, es el grado de incertidumbre que existe sobre un conjunto de datos.
- La entropía de un agujero negro es la cuarta parte del área del horizonte de sucesos.
- ...
- En general, la entropía de un sistema mide su grado de dispersión.

¿Qué es la entropía?

Es un concepto transversal en ciencia.

- En termodinámica, mide el número de microestados compatibles con el macroestado de equilibrio.
- En informática es la aleatoriedad recogida por un sistema operativo.
- En teoría de la información, es el grado de incertidumbre que existe sobre un conjunto de datos.
- La entropía de un agujero negro es la cuarta parte del área del horizonte de sucesos.
- ...
- En general, la entropía de un sistema mide su grado de dispersión.

¿Qué es la entropía?

Es un concepto transversal en ciencia.

- En termodinámica, mide el número de microestados compatibles con el macroestado de equilibrio.
- En informática es la aleatoriedad recogida por un sistema operativo.
- En teoría de la información, es el grado de incertidumbre que existe sobre un conjunto de datos.
- La entropía de un agujero negro es la cuarta parte del área del horizonte de sucesos.
- ...
- En general, la entropía de un sistema mide su grado de dispersión.

¿Qué es la entropía?

Es un concepto transversal en ciencia.

- En termodinámica, mide el número de microestados compatibles con el macroestado de equilibrio.
- En informática es la aleatoriedad recogida por un sistema operativo.
- En teoría de la información, es el grado de incertidumbre que existe sobre un conjunto de datos.
- La entropía de un agujero negro es la cuarta parte del área del horizonte de sucesos.
- ...
- En general, la entropía de un sistema mide su grado de dispersión.

¿Qué es la entropía?

Es un concepto transversal en ciencia.

- En termodinámica, mide el número de microestados compatibles con el macroestado de equilibrio.
- En informática es la aleatoriedad recogida por un sistema operativo.
- En teoría de la información, es el grado de incertidumbre que existe sobre un conjunto de datos.
- La entropía de un agujero negro es la cuarta parte del área del horizonte de sucesos.
- ...
- En general, la entropía de un sistema mide su grado de dispersión.

¿Qué es la entropía?

Es un concepto transversal en ciencia.

- En termodinámica, mide el número de microestados compatibles con el macroestado de equilibrio.
- En informática es la aleatoriedad recogida por un sistema operativo.
- En teoría de la información, es el grado de incertidumbre que existe sobre un conjunto de datos.
- La entropía de un agujero negro es la cuarta parte del área del horizonte de sucesos.
- ...
- En general, la entropía de un sistema mide su grado de dispersión.

Entropía: Fundamentos matemáticos

- Consideremos una distribución de probabilidad **P** y un evento E que puede ocurrir o no.
- Si al realizar el experimento, obtenemos E (o $\neg E$), entonces podemos decir que tenemos *información* que no teníamos antes.
- $I(E)$ es la medida cuantitativa de la *información* que hemos obtenido.
- De este modo, la *información* que hemos ganado al obtener E es también una medida de la *incertidumbre* que teníamos sobre el resultado del experimento.

Entropía: Fundamentos matemáticos

- Consideremos una distribución de probabilidad **P** y un evento E que puede ocurrir o no.
- Si al realizar el experimento, obtenemos E (o $\neg E$), entonces podemos decir que tenemos *información* que no teníamos antes.
- $I(E)$ es la medida cuantitativa de la *información* que hemos obtenido.
- De este modo, la *información* que hemos ganado al obtener E es también una medida de la *incertidumbre* que teníamos sobre el resultado del experimento.

Entropía: Fundamentos matemáticos

- Consideremos una distribución de probabilidad **P** y un evento E que puede ocurrir o no.
- Si al realizar el experimento, obtenemos E (o $\neg E$), entonces podemos decir que tenemos *información* que no teníamos antes.
- $I(E)$ es la medida cuantitativa de la *información* que hemos obtenido.
- De este modo, la *información* que hemos ganado al obtener E es también una medida de la *incertidumbre* que teníamos sobre el resultado del experimento.

Entropía: Fundamentos matemáticos

- Consideremos una distribución de probabilidad **P** y un evento E que puede ocurrir o no.
- Si al realizar el experimento, obtenemos E (o $\neg E$), entonces podemos decir que tenemos *información* que no teníamos antes.
- $I(E)$ es la medida cuantitativa de la *información* que hemos obtenido.
- De este modo, la *información* que hemos ganado al obtener E es también una medida de la *incertidumbre* que teníamos sobre el resultado del experimento.

Entropía: Fundamentos matemáticos

La *información* o *incertidumbre* es una función de valores reales que sólo depende de las probabilidades del evento y que satisface las siguientes propiedades:

- Un evento con probabilidad 1 (evento seguro) tiene incertidumbre (información) 0.
- Si un evento E_1 tiene menos probabilidad de ocurrir que otro E_2, entonces, E_1 proporciona más información que E_2 (o, en otras palabras, tiene mayor incertidumbre).
- La información proporcionada por dos eventos independientes que ocurren simultáneamente es la suma de sus informaciones individuales (o, en otras palabras, la suma de sus incertidumbres).

Entropía: Fundamentos matemáticos

La *información* o *incertidumbre* es una función de valores reales que sólo depende de las probabilidades del evento y que satisface las siguientes propiedades:

- Un evento con probabilidad 1 (evento seguro) tiene incertidumbre (información) 0.
- Si un evento E_1 tiene menos probabilidad de ocurrir que otro E_2, entonces, E_1 proporciona más información que E_2 (o, en otras palabras, tiene mayor incertidumbre).
- La información proporcionada por dos eventos independientes que ocurren simultáneamente es la suma de sus informaciones individuales (o, en otras palabras, la suma de sus incertidumbres).

Entropía: Fundamentos matemáticos

La *información* o *incertidumbre* es una función de valores reales que sólo depende de las probabilidades del evento y que satisface las siguientes propiedades:

- Un evento con probabilidad 1 (evento seguro) tiene incertidumbre (información) 0.
- Si un evento E_1 tiene menos probabilidad de ocurrir que otro E_2, entonces, E_1 proporciona más información que E_2 (o, en otras palabras, tiene mayor incertidumbre).
- La información proporcionada por dos eventos independientes que ocurren simultáneamente es la suma de sus informaciones individuales (o, en otras palabras, la suma de sus incertidumbres).

Entropía: Fundamentos matemáticos

- La información I sólo depende de la probabilidad del evento E, $P(E)$.
- Se puede probar que las únicas funciones que verifican las propiedades anteriores son

$$\Lambda(t) = \begin{cases} -b \ln t & 0 < t \leq 1 \\ \infty & t = 0 \end{cases}$$

  donde b es cualquier número real positivo.
- Formalmente, la información del evento E se define como la función Λ sobre la probablidad de E

$$I(E) = \Lambda(P(E))$$

Entropía: Fundamentos matemáticos

Definición:

Sea $P = \{p_1, p_2, \ldots, p_n\}$ una distribución de probabilidad. Llamamos **entropía** b-aria de la distribución P al valor esperado de la función de información asociada a la distribución de probabilidad.

$$H_b(p_1, p_2, \ldots, p_n) = -\sum p_i \log_b p_i$$

con $p_i \log_b p_i = 0$.

- El término *entropía* fue usado por primera vez por Clausius en 1864 e introducido en la teoría de la información por Shannon en 1948.

Entropía: Fundamentos matemáticos

Propiedades

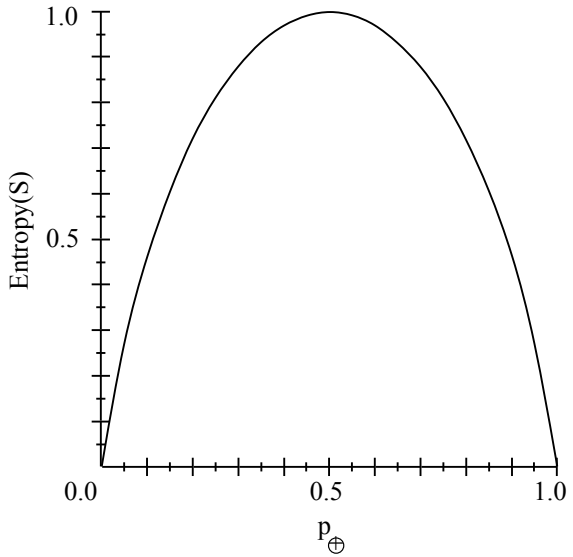
- $H_n(p_1, \ldots, p_n)$ es una función contínua y simétrica respecto a sus argumentos.
- $H_{n+1}(p_1, \ldots, p_n, 0) = H_n(p_1, \ldots, p_n)$
- $H_n(p_1, \ldots, p_n) \leq H_n(\frac{1}{n}, \ldots, \frac{1}{n})$
- $H_2(0, 1) = H_2(1, 0) = 0$
- $H_2(\frac{1}{2}, \frac{1}{2}) = 1$

Entropía: Fundamentos matemáticos

- La función de entropía con base 2 es muy común. Si tomamos base 2, la unidad de entropía es el *bit*. Si tomamos logatirmo neperiano (natural) la unidad es el *nat*. Si además la variable aleatoria sólo puede tomar dos valores, la fórmula para el cálculo de la entropía queda

$$H_2(p, 1-p) = p\, log_2 \frac{1}{p} + (1-p)\, log_2 \frac{1}{1-p}$$

Su gráfica es



¿Cómo saber qué atributo clasifica mejor?

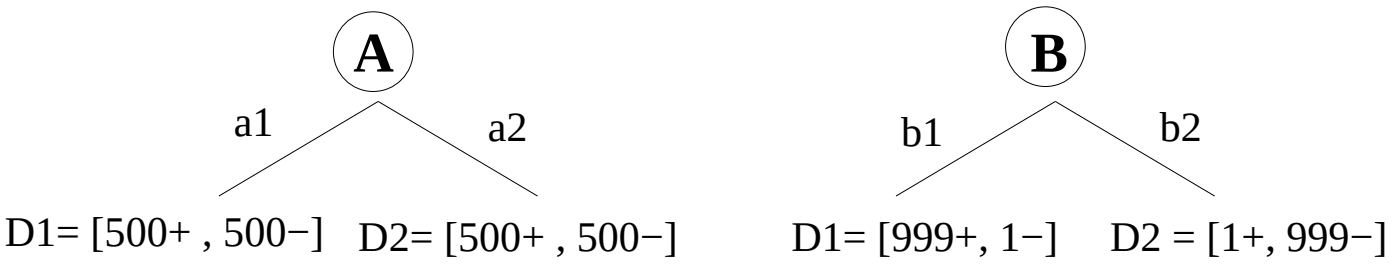
- Entropía de un conjunto de ejemplos D (resp. de una clasificación):

$$Ent(D) = -\frac{|P|}{|D|} \cdot log_2 \frac{|P|}{|D|} - \frac{|N|}{|D|} \cdot log_2 \frac{|N|}{|D|}$$

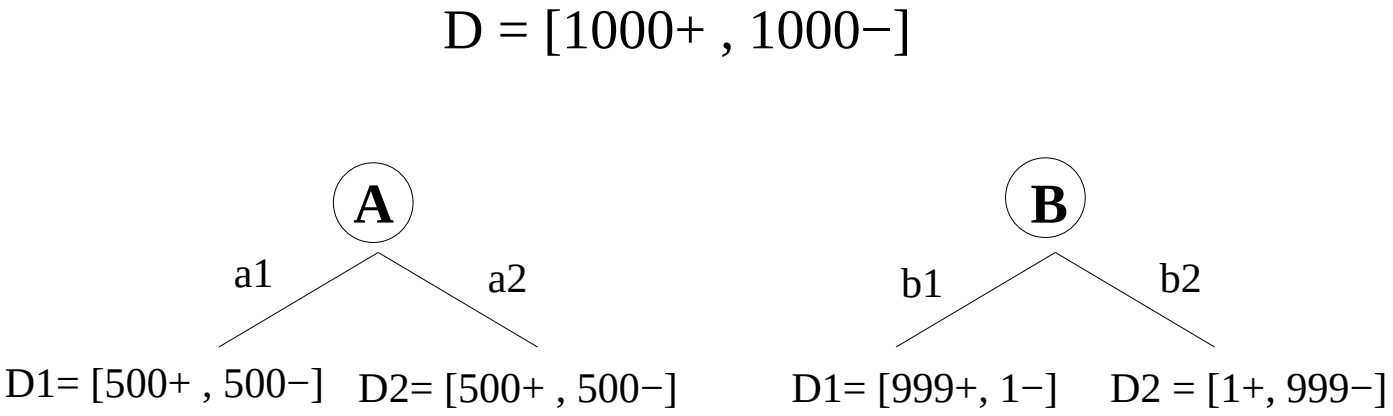
  donde P y N son, respectivamente, los subconjuntos de ejemplos positivos y negativos de D
  - Notación: $Ent([p^+, n^-])$, donde $p = |P|$ y $n = |N|$
- Intuición:
  - Mide el *grado de dispersión* de la clasificación.
  - Mide la ausencia de "homegeneidad" de la clasificación
  - Teoría de la Información: cantidad media de información (en bits) necesaria para codificar la clasificación de un ejemplo de D

¿Cómo saber qué atributo clasifica mejor?

$$D = [1000+ , 1000-]$$



¿Cómo saber qué atributo clasifica mejor?



Intuitivamente, el grado de dispersión (entropía) de los conjuntos *D*1 y *D*2 obtenidos usando el atributo **B** es menor que si usamos el atributo **A**.

¿Cómo saber qué atributo clasifica mejor?

Ejemplos:

- $Ent([9^+, 5^-]) = -\frac{9}{14} \cdot log_2 \frac{9}{14} - \frac{5}{14} \cdot log_2 \frac{5}{14} = 0{,}94$
- $Ent([k^+, k^-]) = 1$ (ausencia total de homogeneidad)
- $Ent([p^+, 0^-]) = Ent([0^+, n^-]) = 0$ (homogeneidad total)

Ganancia de información

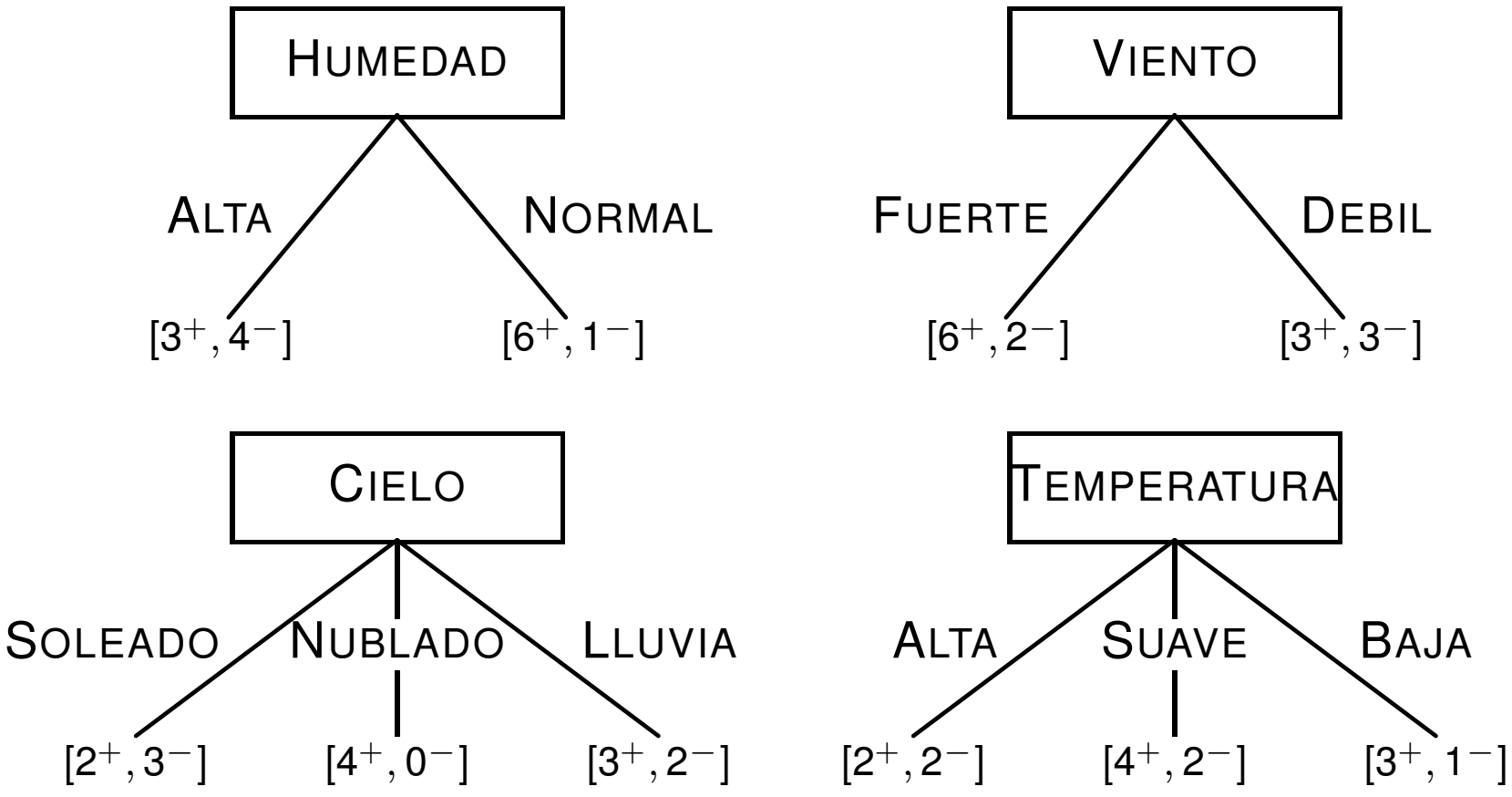
- Preferimos nodos con menos entropía (árboles pequeños)
- Entropía esperada después de usar un atributo A en el árbol:
$$\sum_{v \in Valores(A)} \frac{|D_v|}{|D|} \cdot Ent(D_v)$$
donde D_v es el subconjunto de ejemplos de D con valor del atributo A igual a v
- Ganancia de información esperada después de usar un atributo A:
$$Ganancia(D, A) = Ent(D) - \sum_{v \in Valores(A)} \frac{|D_v|}{|D|} \cdot Ent(D_v)$$
- En el algoritmo ID3, en cada nodo usamos el atributo con mayor ganancia de información (considerando los ejemplos correspondientes al nodo)

Algoritmo ID3 (ejemplo)

Conjunto de entrenamiento

Ej.	Cielo	Temperatura	Humedad	Viento	JugarTenis
D_1	Soleado	Alta	Alta	Débil	-
D_2	Soleado	Alta	Alta	Fuerte	-
D_3	Nublado	Alta	Alta	Débil	+
D_4	Lluvia	Suave	Alta	Débil	+
D_5	Lluvia	Baja	Normal	Débil	+
D_6	Lluvia	Baja	Normal	Fuerte	-
D_7	Nublado	Baja	Normal	Fuerte	+
D_8	Soleado	Suave	Alta	Débil	-
D_9	Soleado	Baja	Normal	Débil	+
D_{10}	Lluvia	Suave	Normal	Débil	+
D_{11}	Soleado	Suave	Normal	Fuerte	+
D_{12}	Nublado	Suave	Alta	Fuerte	+
D_{13}	Nublado	Alta	Normal	Débil	+
D_{14}	Lluvia	Suave	Alta	Fuerte	-

Algoritmo ID3 (ejemplo)

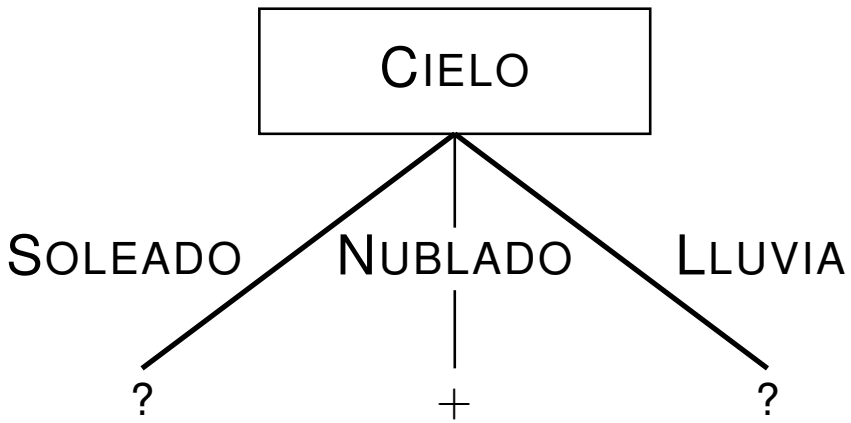


Algoritmo ID3 (ejemplo)

- Entropía inicial: $Ent([9^+, 5^-]) = 0{,}94$
- Selección del atributo para el nodo raíz:
  - $Ganancia(D, \text{HUMEDAD}) = 0{,}94 - \frac{7}{14} \cdot Ent([3^+, 4^-]) - \frac{7}{14} \cdot Ent([6^+, 1^-]) = 0{,}151$
  - $Ganancia(D, \text{VIENTO}) = 0{,}94 - \frac{8}{14} \cdot Ent([6^+, 2^-]) - \frac{6}{14} \cdot Ent([3^+, 3^-]) = 0{,}048$
  - $Ganancia(D, \text{CIELO}) = 0{,}94 - \frac{5}{14} \cdot Ent([2^+, 3^-]) - \frac{4}{14} \cdot Ent([4^+, 0^-]) - \frac{5}{14} \cdot Ent([3^+, 2^-]) = 0{,}246$ (mejor atributo)
  - $Ganancia(D, \text{TEMPERATURA}) = 0{,}94 - \frac{4}{14} \cdot Ent([2^+, 2^-]) - \frac{6}{14} \cdot Ent([4^+, 2^-]) - \frac{4}{14} \cdot Ent([3^+, 1^-]) = 0{,}02$
- El atributo seleccionado es CIELO

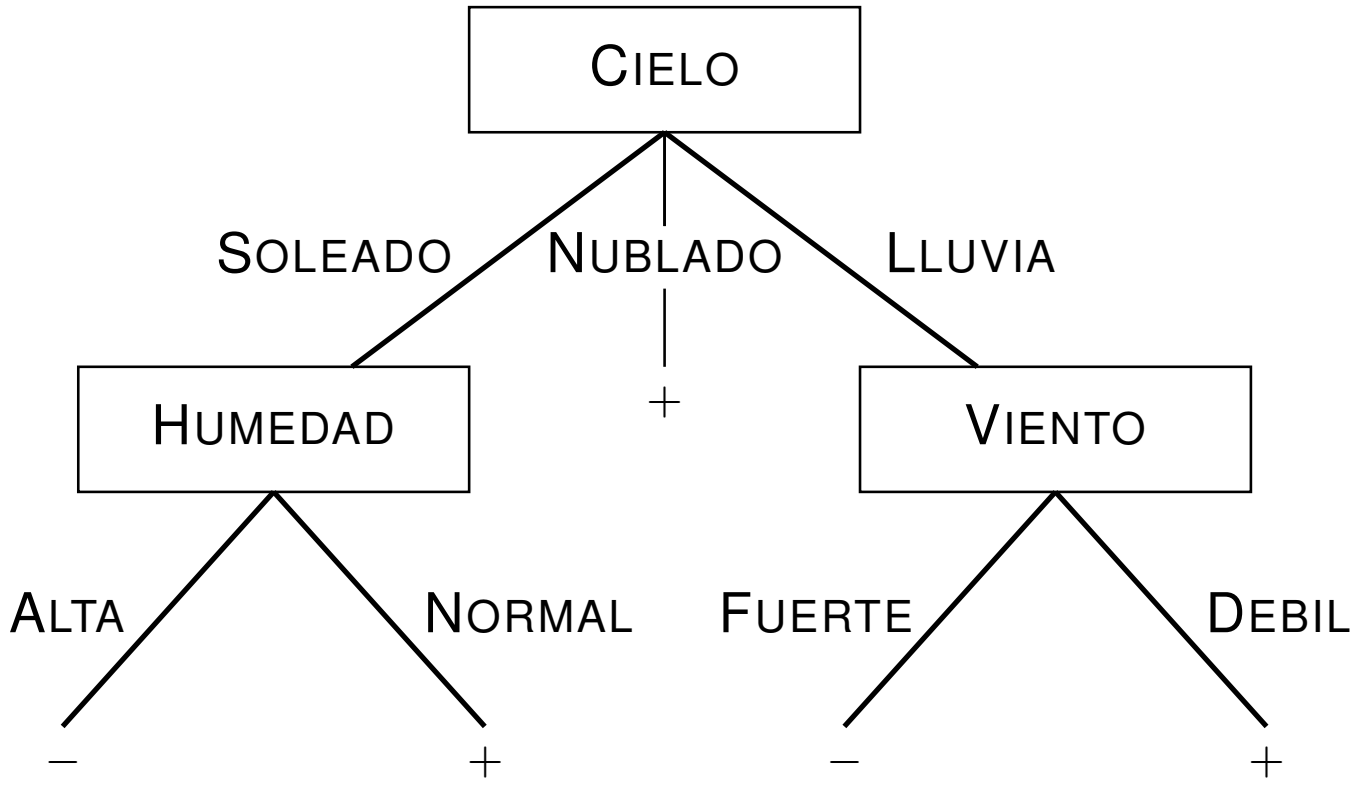
Algoritmo ID3 (ejemplo)

- Árbol parcialmente construido:



Algoritmo ID3 (ejemplo 1)

- Árbol finalmente aprendido:



. . . y mucho más

- Medida de rendimiento
- Conjunto de entrenamiento y conjunto de prueba
- Medida del rendimiento: proporción de ejemplos bien clasificados en el conjunto de prueba
- Validación cruzada
- Estratificación

- Concepto de ruido
- Concepto de sobreajuste
- Poda a priori
- Poda a posteriori
- . . .

. . . y mucho más

- Medida de rendimiento
- Conjunto de entrenamiento y conjunto de prueba
- Medida del rendimiento: proporción de ejemplos bien clasificados en el conjunto de prueba
- Validación cruzada
- Estratificación

- Concepto de ruido
- Concepto de sobreajuste
- Poda a priori
- Poda a posteriori
- . . .

. . . y mucho más

- Atributos con valores continuos
- Otras medidas para seleccionar atributos distintas a la entropía
- Otras estimaciones de error
- Atributos sin valores
- Atributos con coste
- . . .

- Algoritmos C4.5 y C5.0 (Quinlan)
- Algoritmo C4.8 (implementado en WEKA como J4.8)

. . . y mucho más

El error cometido se puede reducir de forma significativa aprendiendo muchos árboles de un mismo conjunto de entrenamiento y estableciendo algún sistema de votación entre ellos para clasificar una nueva instancia.

- **Bagging** (**B**ootstrap **agg**regating). Se generan K árboles diferentes a partir de K muestras tomada con reemplazamiento del conjunto de entrenamiento. Cada uno de esos árboles proporciona un *voto* sobre la clasificación de una nueva instancia. La clasificación con más votos es la clasificación asignada por el algorotmo.

. . . y mucho más

- **Boosting**. Se le asigna un peso a cada ejemplo del conjunto de entrenamiento. Se construyen varios árboles usando el peso de los ejemplos, uno tras otro, y en cada iteración el peso de los ejemplos cambia en función de la clasificación obtenida por le árbol construido.
- **Random Forests** El método combina la idea de Bagging la selección aleatoria de atributos
- . . .

Gracias

Miguel A. Gutiérrez Naranjo

http://www.cs.us.es/~naranjo/