Árboles de Decisión

Aprendizaje Supervisado

Se parte de un conjunto de datos clasificados

Aprendizaje o Entrenamiento

- Generar o aprender un modelo a partir de un conjunto de datos llamado “datos de entrenamiento”

Prueba

- Probar/evaluar los modelos generados usando conjuntos de “datos de prueba”

Introducción

Un árbol de decisión es una forma gráfica y analítica de representar todos los eventos (sucesos) que pueden surgir a partir de una decisión asumida en cierto momento.

Nos ayudan a tomar la decisión “más acertada”, desde un punto de vista probabilístico, ante un abanico de posibles decisiones.

Permite desplegar visualmente un problema y organizar el trabajo de cálculos que deben realizarse.

Ventajas

Resume los ejemplos de partida, permitiendo la clasificación de nuevos casos.

Facilita la interpretación de una decisión.

Proporciona un alto grado de comprensión.

Explica el comportamiento respecto a una determinada tarea.

Reduce el número de variables independientes.

Ejemplo

Los árboles de decisión se utilizan en sistemas expertos, aunque en algunas ocasiones puede llegar a ser más lento.

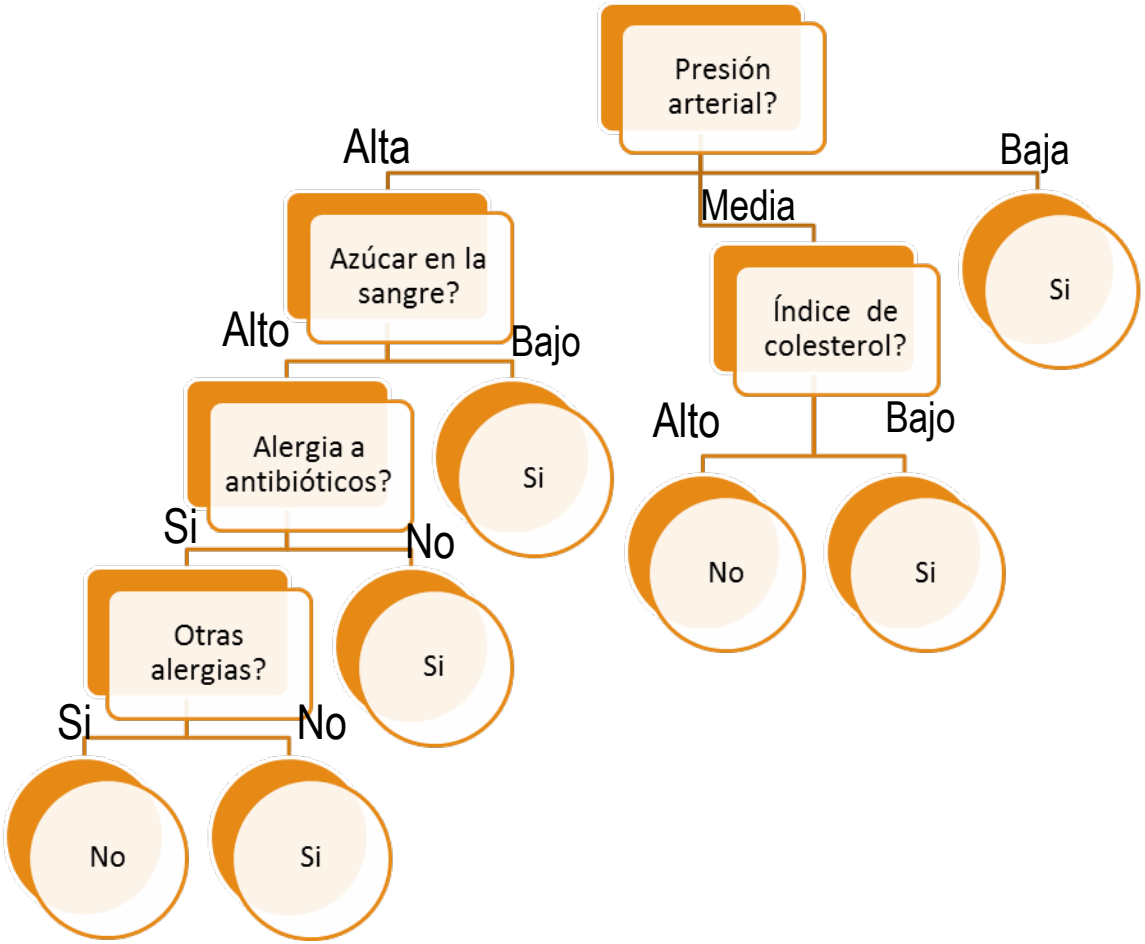
El siguiente ejemplo es de un sistema experto que ayuda a **diagnosticar que se emplee un fármaco X** en una persona con presión arterial.

Ejemplo

Los árboles de decisión se utilizan en sistemas expertos, aunque en algunas ocasiones puede llegar a ser más lento.

El siguiente ejemplo es de un sistema experto que ayuda a **diagnosticar que se emplee un fármaco X** en una persona con presión arterial.

Ejemplo (continuación)



Pasos para el Análisis del Árbol de Decisión

Definir el problema.

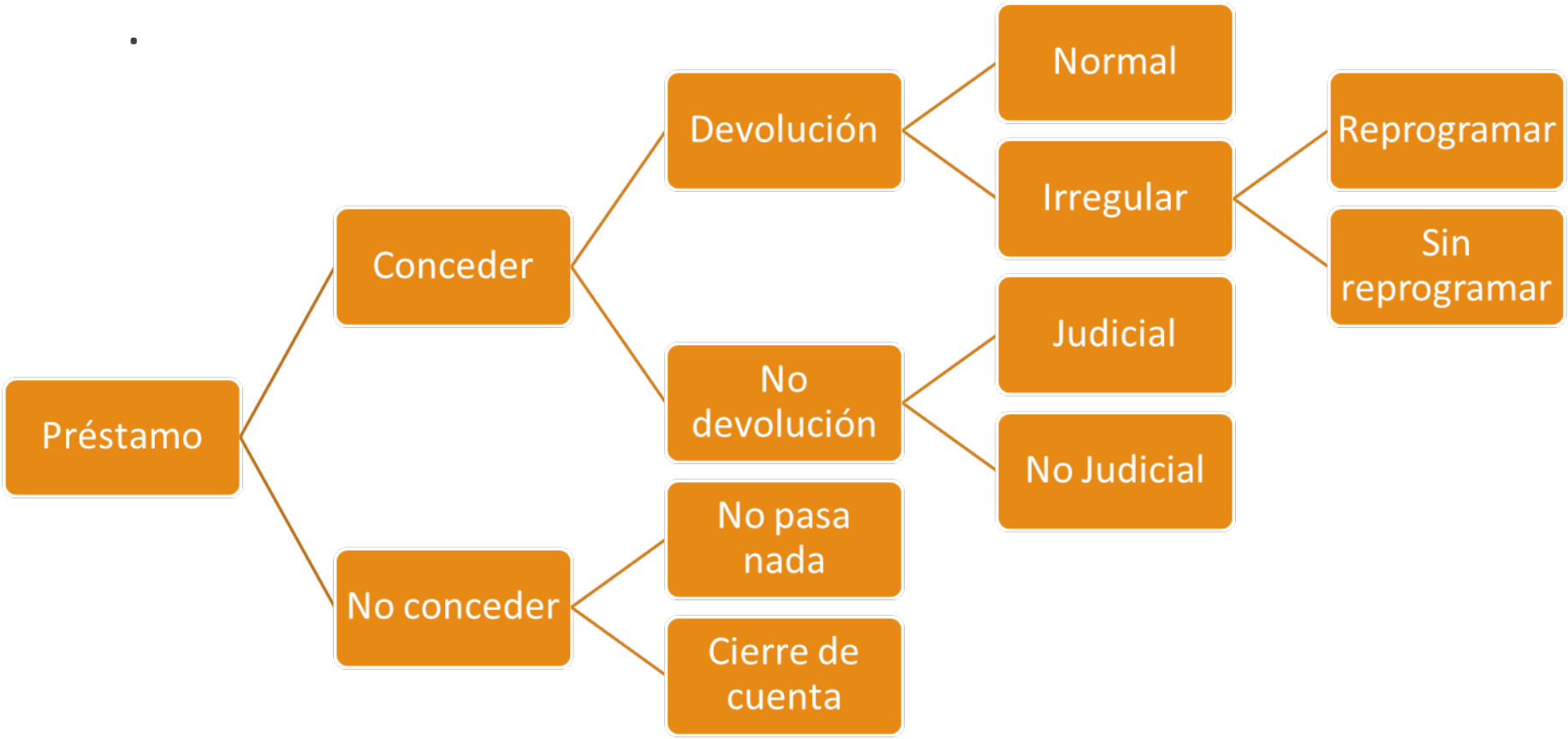
Dibujar el árbol de decisión.

Asignar probabilidades a los eventos aleatorios.

Estimar los resultados para cada combinación posible de alternativas.

Resolver el problema obteniendo como solución la ruta que proporcione la política óptima

Conceder préstamo

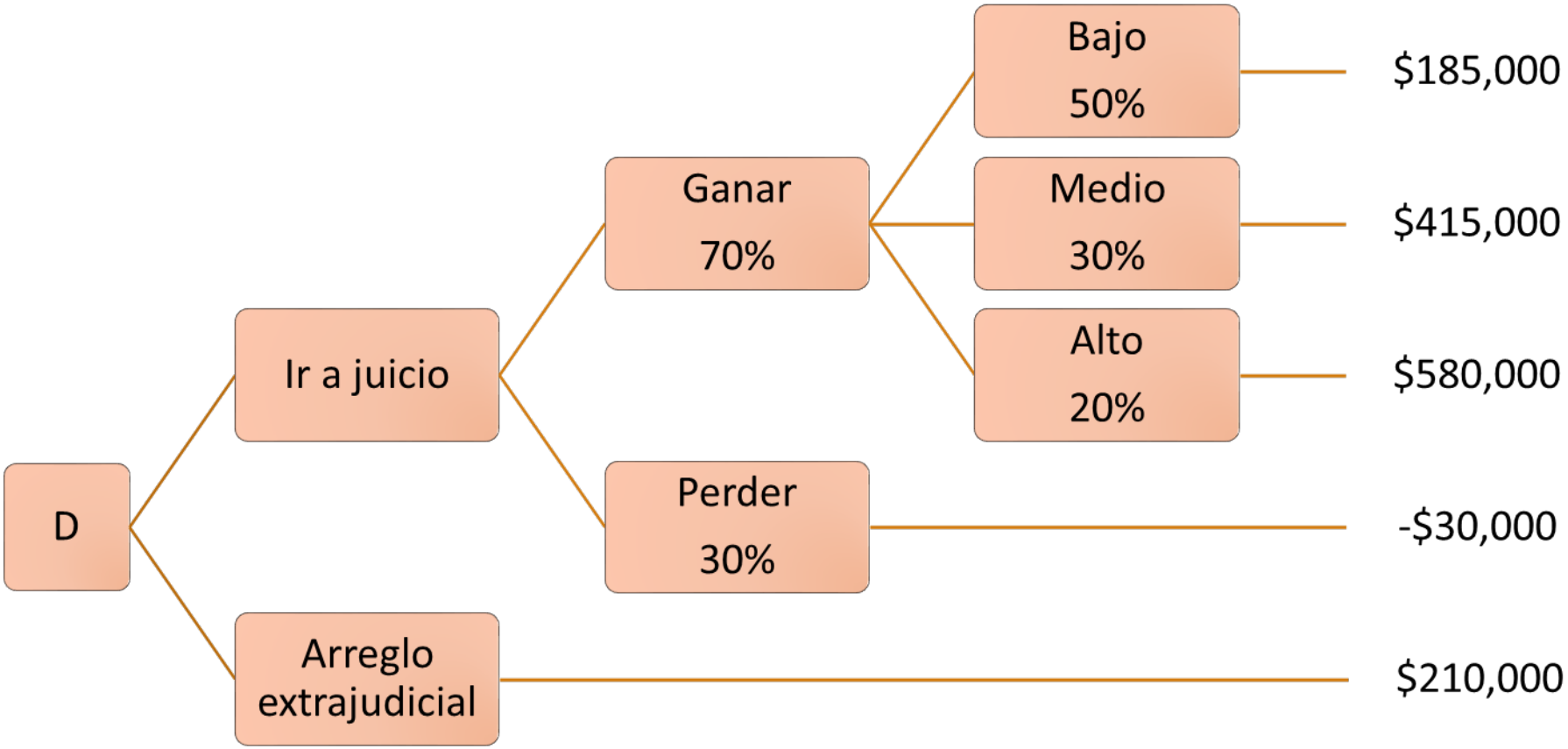


Problema

Una compañía de seguros nos ofrece una indemnización por accidente de $210.000. Si no aceptamos la oferta y decidimos ir a juicio podemos obtener $185.000, $415.000 o $580.000 dependiendo de las alegaciones que el juez considere aceptables. Si perdemos el juicio, debemos pagar los gastos que ascienden a $30.000.

Sabiendo que el 70% de los juicios se gana, y de éstos, en el 50% se obtiene la menor indemnización, en el 30% la intermedia y en el 20% la más alta, determinar la decisión más acertada.

Árbol


D
Ir a juicio
Arreglo extrajudicial
Ganar 70%
Perder 30%
Bajo 50%
Medio 30%
Alto 20%
$185,000
$415,000
$580,000
-$30,000
$210,000

Problema

Una fábrica está evaluada en 150 millones. La fábrica desea incorporar un nuevo producto al mercado. Existen tres estrategias para incorporar el nuevo producto:

- Alternativa 1 Hacer un estudio de mercado del producto de forma de determinar si se introduce o no al mercado.
- Alternativa 2 Introducir inmediatamente el producto al mercado (sin estudio).
- Alternativa 3 No lanzar inmediatamente el producto al mercado (sin estudio).

Problema

En ausencia de estudio de mercado, la fábrica estima que el producto tiene un 55% de posibilidades de ser exitoso y de 45% de ser un fracaso.

Si el producto es exitoso, la fábrica aumentaría en 300 millones su valor, si el producto fracasa se devaluaría en 100 millones.

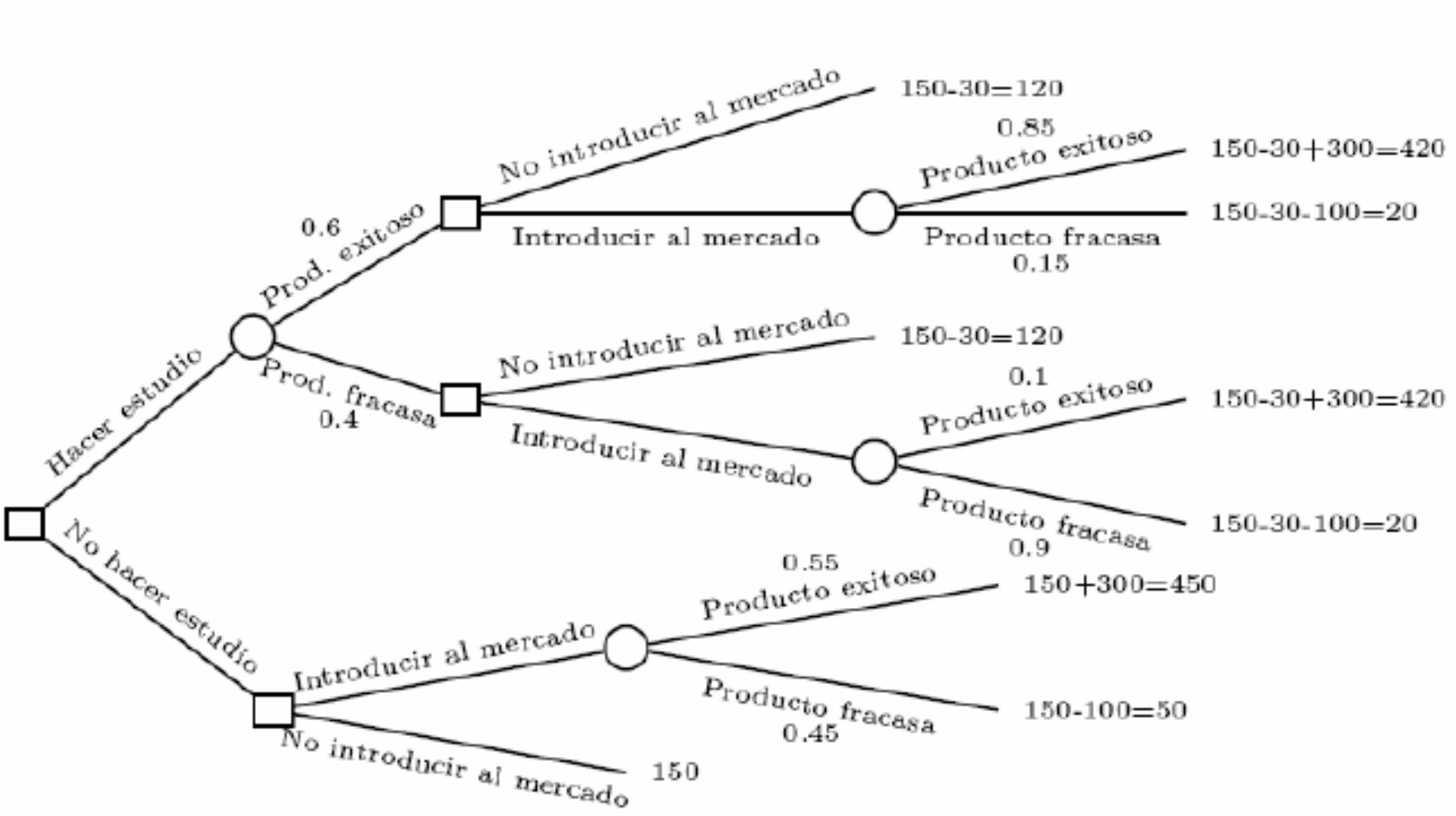
El estudio de mercado vale 30 millones. El estudio predice que existe un 60% de probabilidad de que el producto sea exitoso.

Problema

Si el estudio de mercado determina que el producto sería exitoso, existe un 85% de posibilidades de que efectivamente lo sea.

Si el estudio de mercado determina que el producto sería un fracaso, existe sólo un 15% de posibilidades de que el producto no sea exitoso. Si la empresa no desea correr riesgos (desea maximizar el valor esperado de la empresa).

¿Qué estrategia debería seguir?


Hacer estudio
No hacer estudio
Prod. exitoso
0.6
Prod. fracasa
0.4
No introducir al mercado
150-30=120
Introducir al mercado
0.85
Producto exitoso
150-30+300=420
Producto fracasa
0.15
150-30-100=20
No introducir al mercado
150-30=120
Introducir al mercado
0.1
Producto exitoso
150-30+300=420
Producto fracasa
0.9
150-30-100=20
Introducir al mercado
0.55
Producto exitoso
150+300=450
Producto fracasa
0.45
150-100=50
No introducir al mercado
150

Árboles de Decisión para Clasificacción

No son únicos

Queremos árboles sencillos y precisos

Encontrar el mejor árbol es un problema NP-difícil

Procedimientos heurísticos de construcción

¿Cómo elegir los atributos para separar los datos?

- El objetivo es reducir la impureza o incertidumbre en los datos (un conjunto de datos es puro si todos son de la misma clase)
- Elegir un atributo que maximice la “Ganancia de Información” (Teoría de la Información)

Árboles de Decisión

La entropía es una medida de incertidumbre asociada a una variable aleatoria:

- $H[D] = -\sum P(c_j)\log_2 P(c_j)$

donde C es el conjunto de datos y $\sum P(c_j) = 1$

Es el producto de la probabilidad por cantidad de información de cada uno

Árboles de Decisión

Dos eventos x_1 y x_2

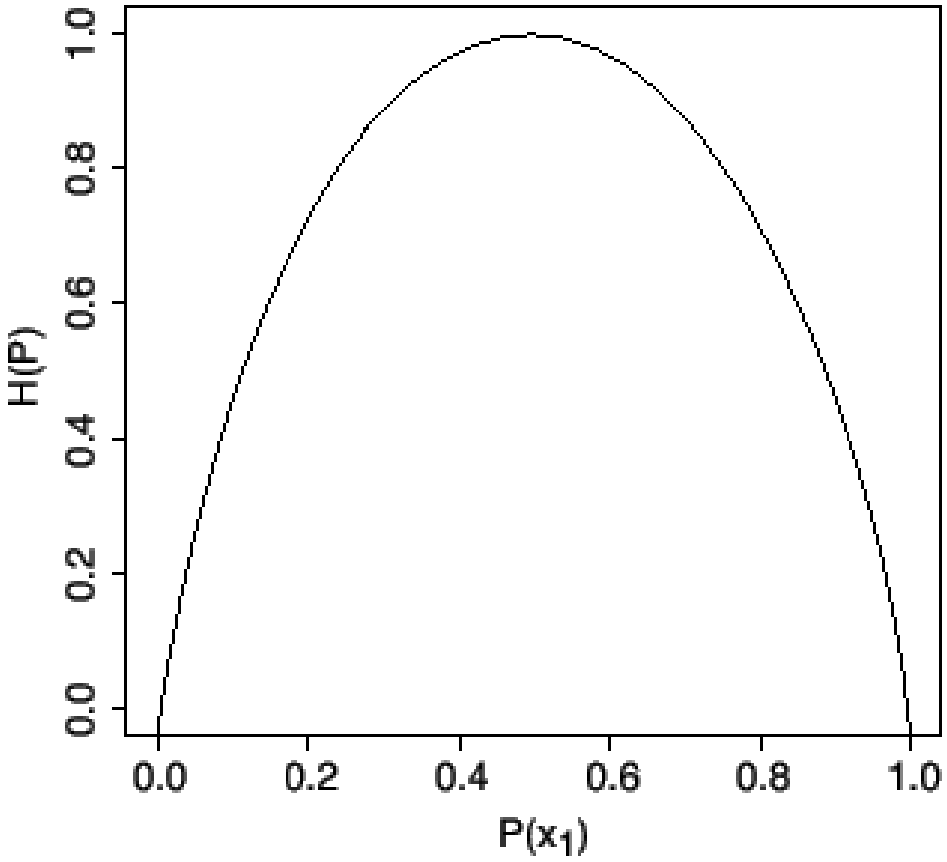
$X = \{x_1, x_2\}$

H(X) = 1, máxima

- $P(x_1) = P(x_2) = 0.5$
- Máxima incertidumbre

H(X) = 0, mínima

- $P(x_1) = 1$ y $P(x_2) = 0$ o viceversa
- Mínima incertidumbre



Árboles de Decisión

Dado un conjunto D de datos

Si se elige un atributo A_i como raíz, con v valores, esto llevará a una partición de D en D_1, D_2,…, D_v conjuntos

La entropía esperada seleccionando A_i como raíz es:

- $H_{Ai}[D] = \sum_v |D_j|/|D| * H[D_j]$

La ganancia de información está dada por la diferencia entre la entropía previa y la de la rama seleccionada:

- $gain(D,A_i) = H[D] - H_{Ai}[D]$

Eligir atributo con mayor ganancia para la partición

Árboles de Decisión

Podas

- No seguir dividiendo nodos con pocas instancias
- Construir árbol completo y luego subir desde las hojas aplicando test estdísticos en cada nodo
- Eliminar nodos internos, subiendo subárboles

- Ver parámetros en weka para J48, que es una implementación de C4.5 de Quinlan, que a su vez es es un refinamiento de ID3 de Quinlan

Árboles de Decisión

Prepoda

- Durante la construcción, en general es el criterio de parada a la hora de seguir especializando una rama

Pospoda

- Eliminar reglas/nodos de abajo hacía arriba para hacer árboles

Combinación

- Por ejemplo prepoda por cardinalidad y pospoda como C4.5
-

Discretización en Árboles de Decisión

¿Atributos numéricos?

64	65	68	69	70	71	72	75	80	81	83	85
yes	no	yes	yes	yes	no	no yes	yes yes	no	yes	yes	no

Infinitas posibilidades

Se intenta con el punto medio entre cada para de valores adyacentes

n-1 posibilidades

Discretización en Árboles de Decisión

64	65	68	69	70	71	72	75	80	81	83	85
yes	no	yes	yes	yes	no	no yes	yes yes	no	yes	yes	no

Calcular la ganancia de información en cada posible corte.

Ejemplo corte entre 70 y 71:

- Antes del corte 9 yes, 5 no, entropía 0,940
- Después del corte entropía 0,939
- Ganancia de Información 0,940-0,939=0,001