

Arboles de Decisión (I)

Carlos Hurtado L.

Depto de Ciencias de la Computación,
Universidad de Chile

Contenido

- Motivación de Clasificación
- Motivación de Árboles de Decisión
- Noción de Split
- Algoritmo de Hunt

Aprendizaje Supervisado

- Dado un conjunto de datos:

$$\{(x_1, f(x_1)), (x_2, f(x_2)), \dots, (x_n, f(x_n))\}.$$

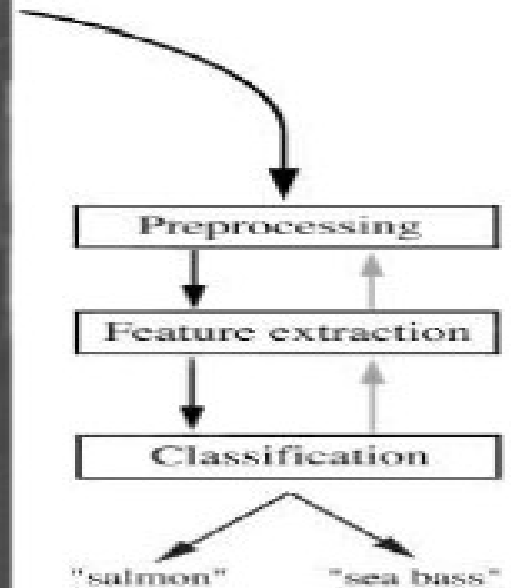
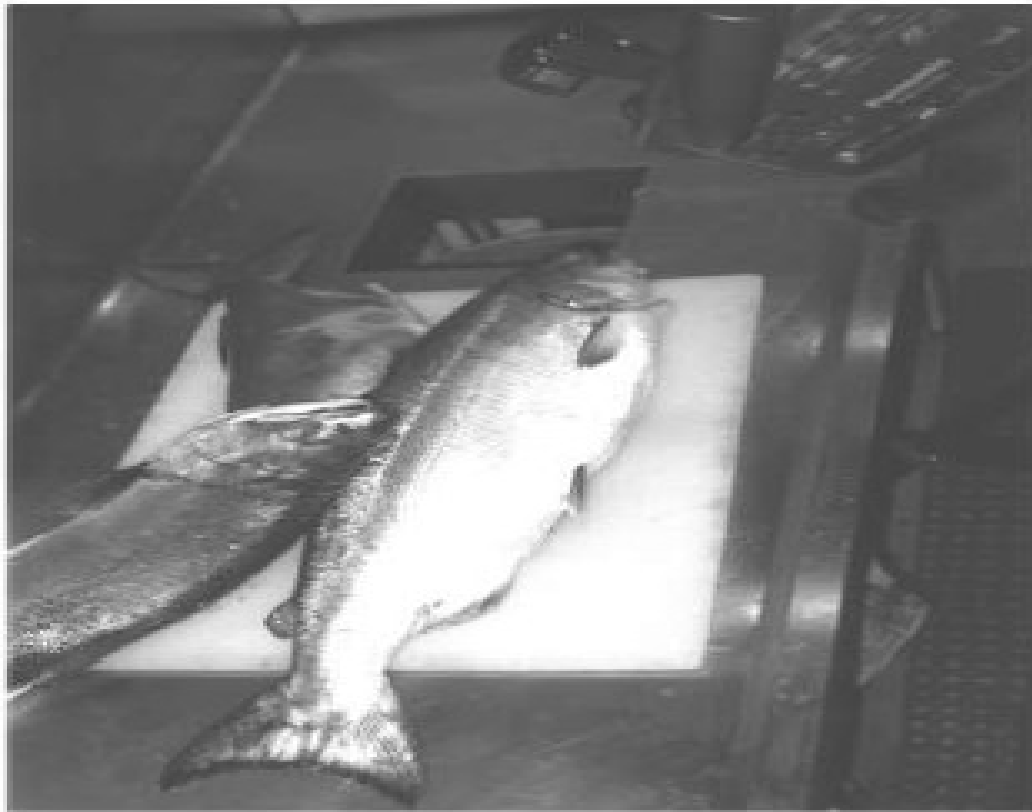
Donde $y=f(x)$ es una función (sistema).

- Construir un modelo para predecir $f(x')$, donde x' no está en el conjunto de entrenamiento.
- Clasificación: f es una función discreta.
- En algunos casos no hay una relación funcional entre variables de entrada y de salida
 - Aprendizaje de probabilidades condicionales $P(y|x)$.

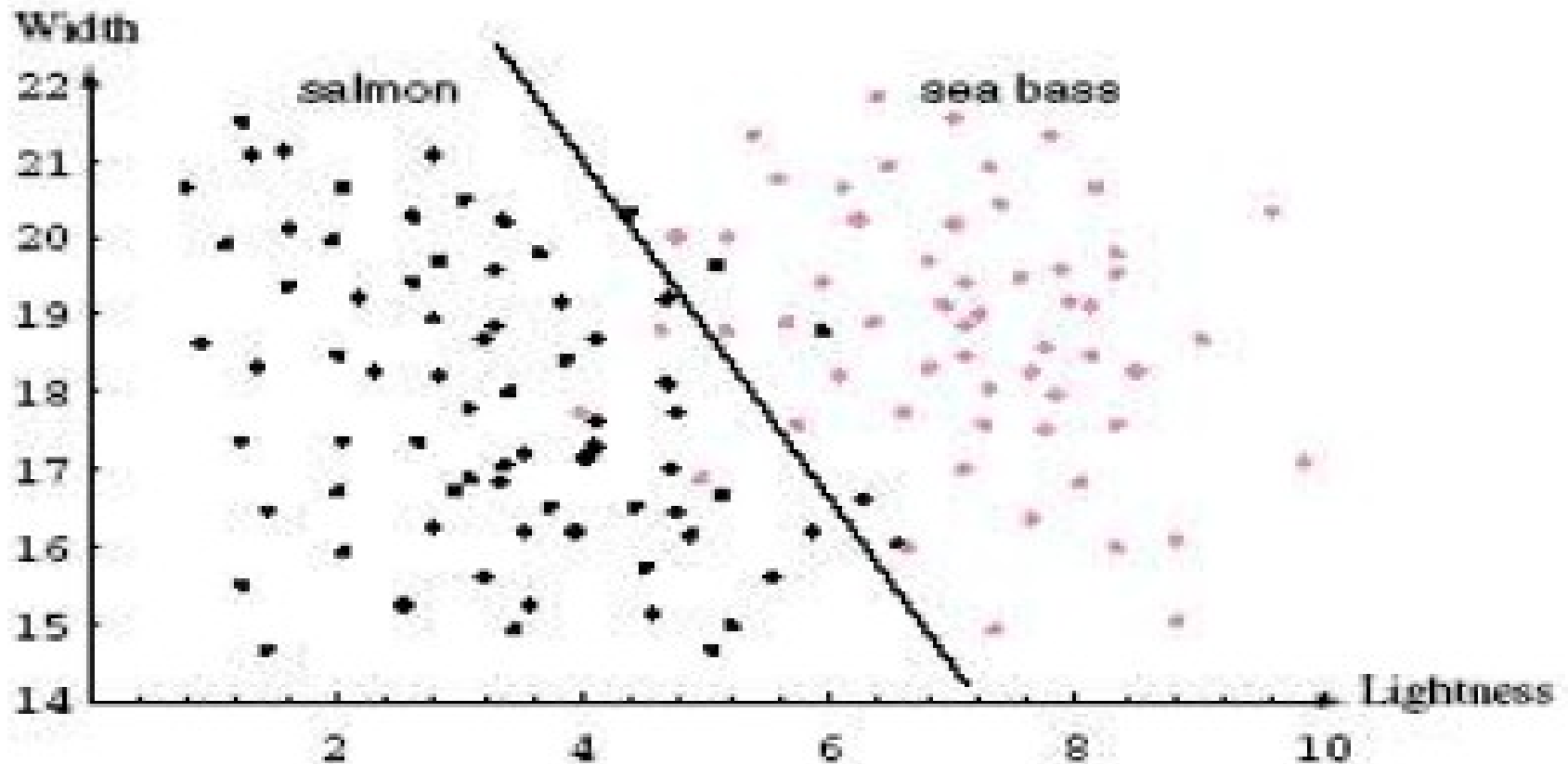
Clasificación

- Dado un conjunto de datos, donde cada dato pertenece a una clase.
- Construir un modelo que permita predecir la clase de un nuevo dato.
- En general: el modelo es "mejor" si el error de predicción es menor.

Clasificación: Ejemplo



Clasificación: ejemplo (cont.)



Aplicaciones de Clasificación (I)

- Detección de Fraude:
 - **Objetivo:** predecir uso fraudulento en tarjetas de crédito
 - **Método:**
 - Usar transacciones de compras en un determinado período e info. en cuentas.
 - Definir atributos de entrada como: cuándo se compra, frecuencia de compra, frecuencia de pagos, etc.
 - Inducir un modelo para predecir clase de nuevas transacciones de compra.

Aplicaciones de Clasificación (II)

- Predicción de Deserciones ("churn"):
 - **Objetivo:** en un determinado mes, predecir qué clientes abandonarán un determinado servicio (ej., teléfono celular)
 - **Método:**
 - Definir atributos de entrada a partir de transacciones del cliente (ej., llamados telefónicos, frecuencia, último llamado, interrupciones, etc.) e info. descriptiva de usuarios.
 - Usar info. de meses anteriores y definir atributo de clase como abandono o no abandono del servicio.
 - Inducir un clasificador.

Aplicaciones de Clasificación (III)

- Clasificación de Documentos:
 - **Objetivo:** Predecir tópico de un nuevo documento
 - **Método:**
 - Definir atributos de entrada a partir del contenido del documento (e.g., vector de términos)
 - Usar documentos pre-clasificados como datos de entrenamiento.
 - Inducir un clasificador.

Tarea

- Machine Learning in Automated Text Categorization. Fabrizio Sebastiani. ACM Computing Surveys, Vol 34, No 1, March 2002.
- Control de Lectura 26 de Octubre.

Etapas en Clasificación

- Construcción del modelo
 - Datos de entrenamiento
- Prueba del modelo
 - Datos de prueba
- Uso del modelo
 - Datos de uso

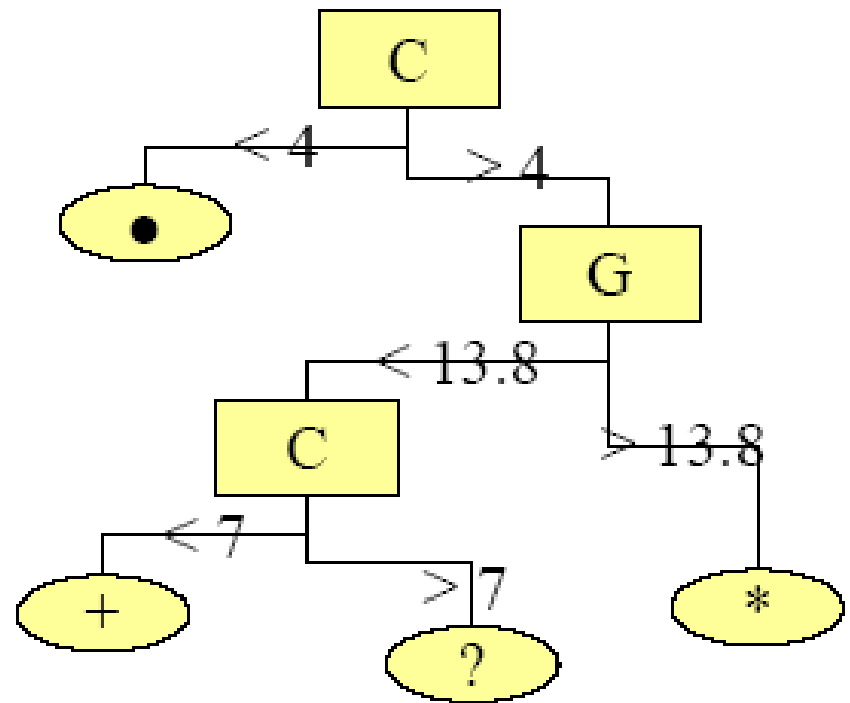
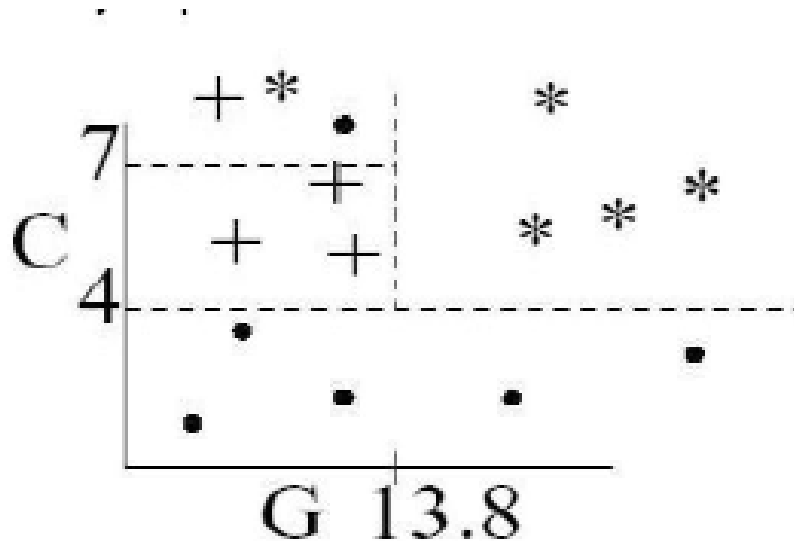
Clasificación: Tipos de Modelos

- Enfoque Discriminante
 - Árboles de Decisión
 - Reglas de Decisión
 - Discriminantes lineales
- Enfoque Probabilístico
 - Redes Bayesianas
- Enfoque de Regresión
 - Redes Neuronales
 - Regresión Logística

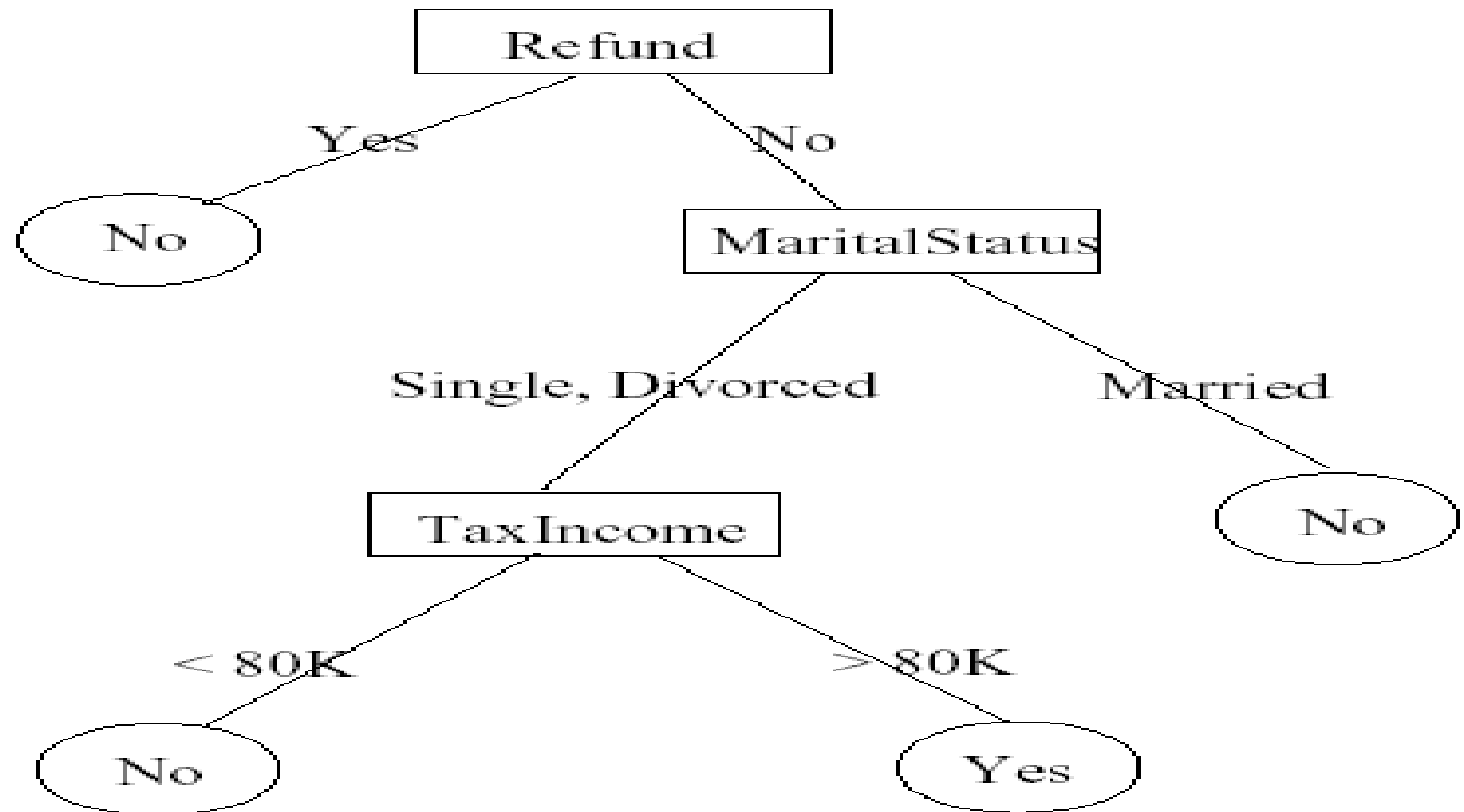
Arboles de Decisión

- Fáciles de construir
- Fáciles de interpretar
 - No sólo sirven para clasificar nuevos datos, sino que para qué determina cada clase.
- Buena precisión en muchos casos

Arboles de Decisión



Arbol de Decisión para Detección de Fraude



Construcción de árboles de decisión

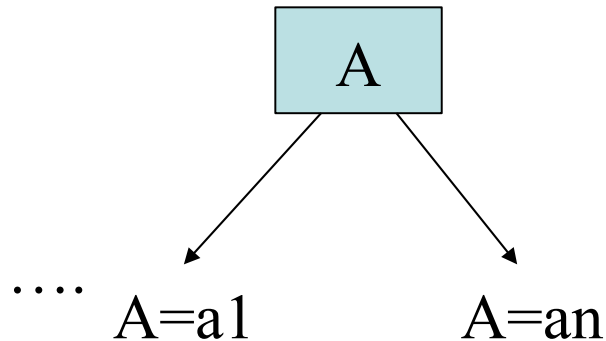
- Etapas
 - Construcción
 - Poda
 - Es necesario “podar” el árbol para evitar **sobre-ajuste**
 - Prueba
 - Medimos la calidad del modelo: error, tamaño, etc.

Construcción de árboles de decisión: algoritmos

- Memoria Principal
 - Algoritmo de Hunt (CLS, 1960's)
 - ID3 (Quinlan 70's and 80's), C4.5 (Quinlan 90's)
- Escalables
 - SLIQ, SPRINT

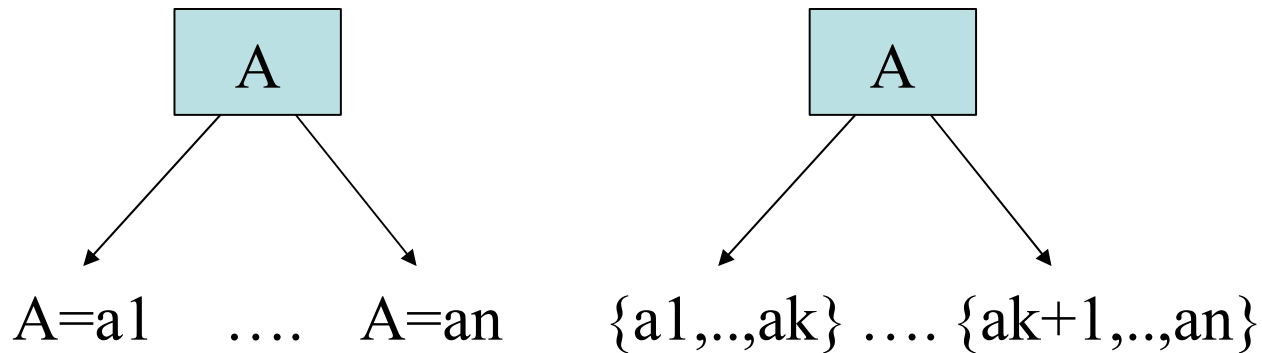
Split

- **Definición:** un *split* es una variable (atributo) más una lista de condiciones sobre la variable.



Split para variables categóricas

- Simple vs. Complejo



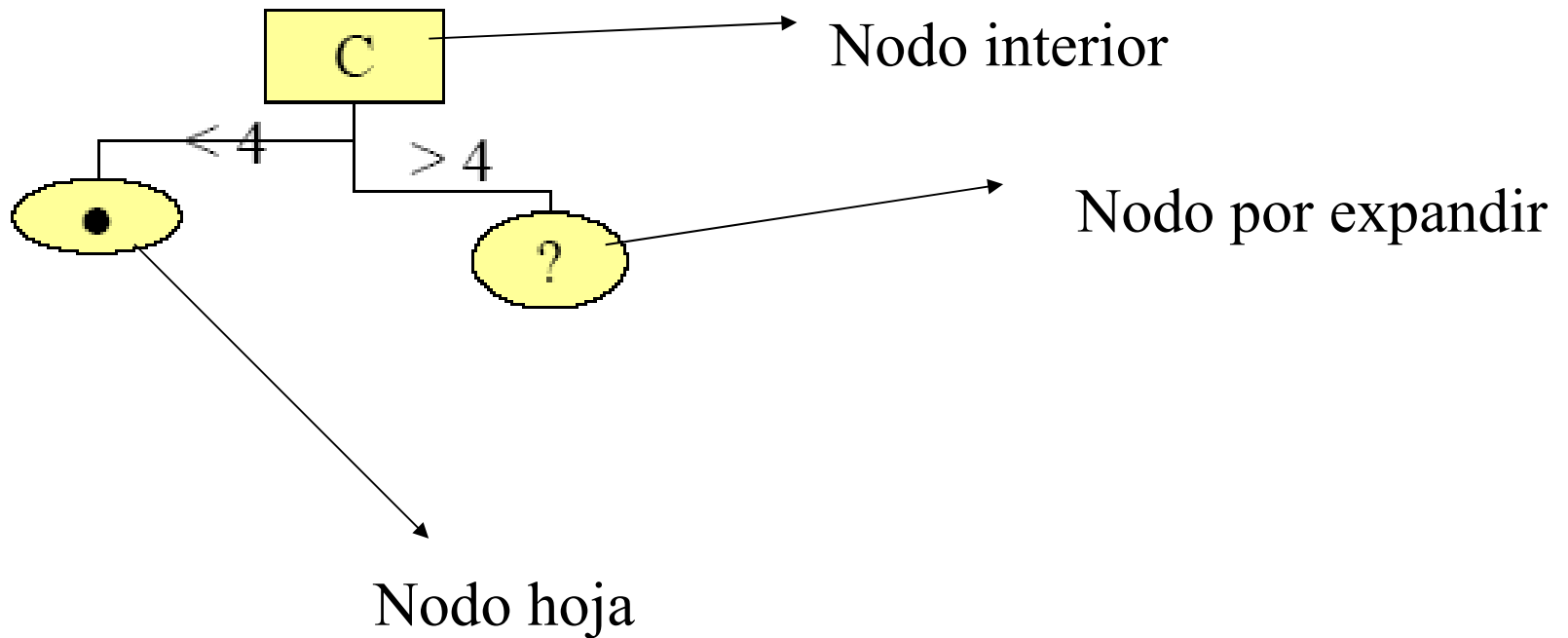
Split para variables numéricas

- Estáticos:
 - Discretizando en un inicio
- Dinámicos:
 - Decisión binaria: $A \leq v$, $A > v$
 - Rangos: $[0,15k)$, $[15k,60k)$, $[60k,100k)$, etc.
 - Combinación lineal de variables
 - Más complejo y poco común en árboles de decisión

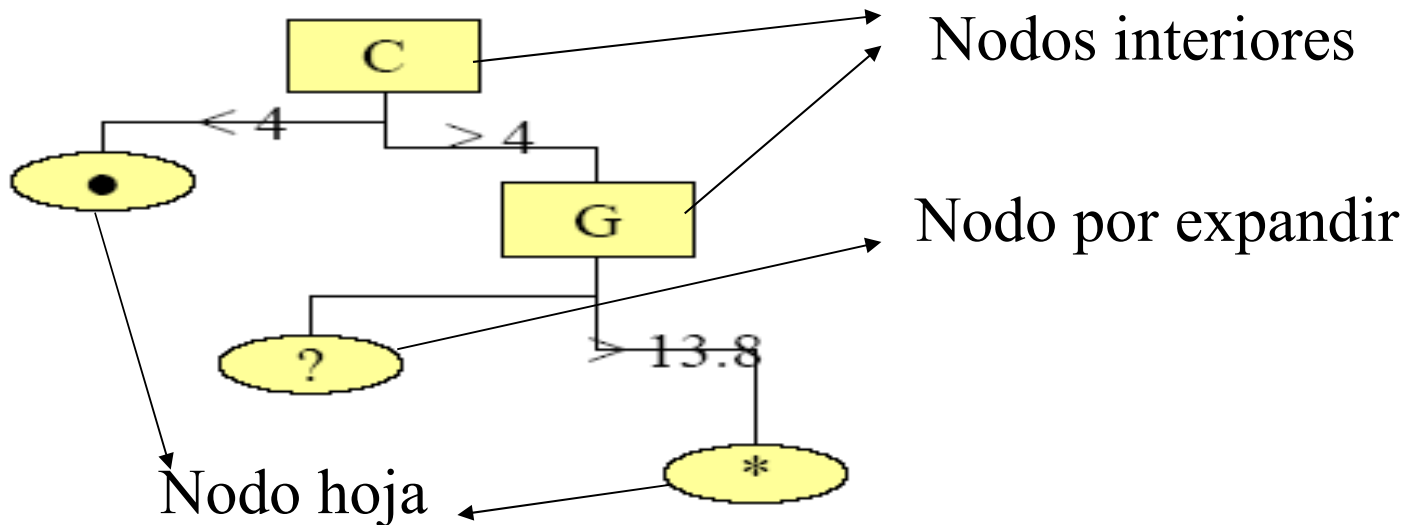
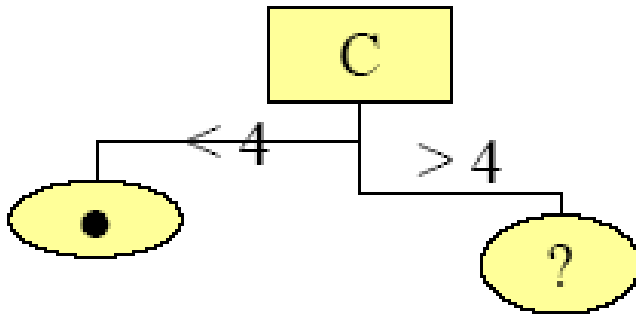
Algoritmo de Hunt

- Idea básica: cada nodo en el árbol de decisión tiene asociado un subconjunto de los datos de entrenamiento
- Inicialmente, el nodo raíz tiene asociado todo el conjunto de entrenamiento
- Construimos un árbol parcial que tiene tres tipos de nodos:
 - Expandidos (interiores)
 - Hojas: serán hojas en el árbol final y tienen asociada una clase
 - Nodos por expandir: son hojas en el árbol parcial, pero deben ser expandidos
- Operación de expansión de un nodo t :
 - Encontrar el mejor split para t
 - Particionar los datos de t en nodos hijos de acuerdo al split
 - Etiquetar t y sus nodos hijos con el mejor split

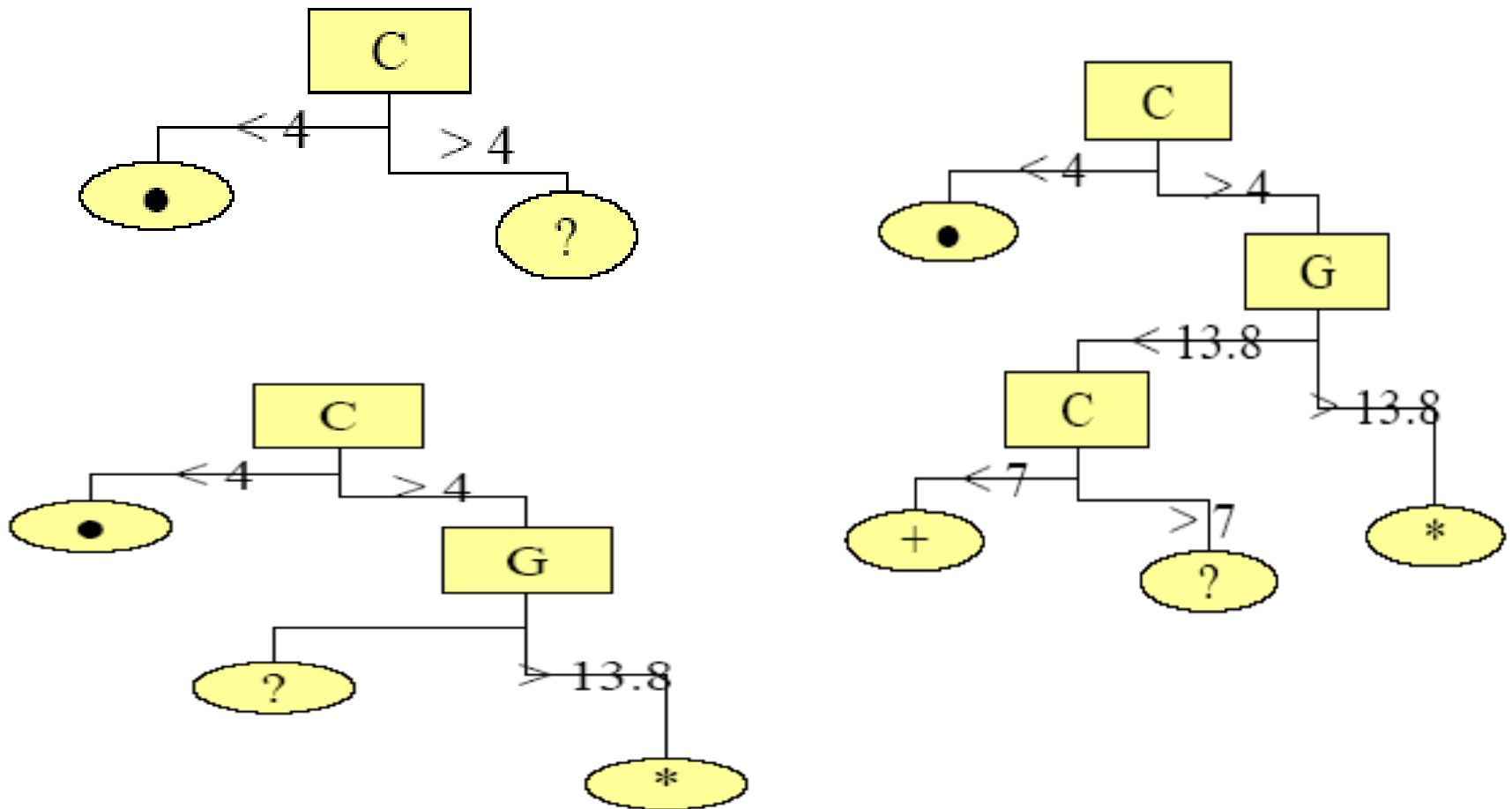
Algoritmo de Hunt (I)



Algoritmo de Hunt (II)



Algoritmo de Hunt (III)



Algoritmo de Hunt

- Main(Conjunto de Datos T)
 - Expandir(T)
- Expandir(Conjunto de Datos S)
 - If (todos los datos están en la misma clase)
then return
 - Encontrar el mejor split r
 - Usar r para particionar S en S_1 y S_2
 - Expandir(S_1)
 - Expandir(S_2)

Algoritmo de Hunt: observaciones

- Las operaciones de expansión se realizan "primero en profundidad"
- Lo complejo es encontrar el mejor split en cada operación de expansión
- Número de splits a buscar depende de si el atributo es categórico o no y del tipo de split (e.g., complejo vs. simple).