UNIVERSITAS Miguel Hernández

Tema 5. Árboles de Decisión.

José L. Sainz-Pardo Auñón

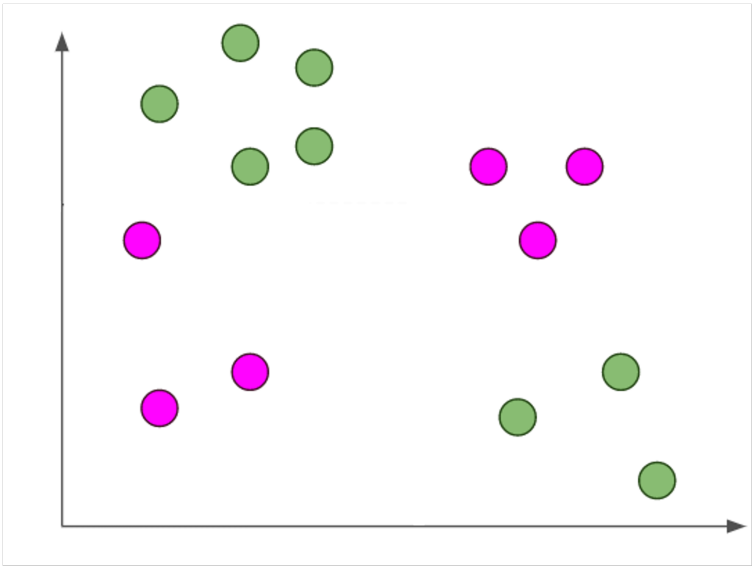
TÉCNICAS ESTADÍSTICAS PARA EL APRENDIZAJE II

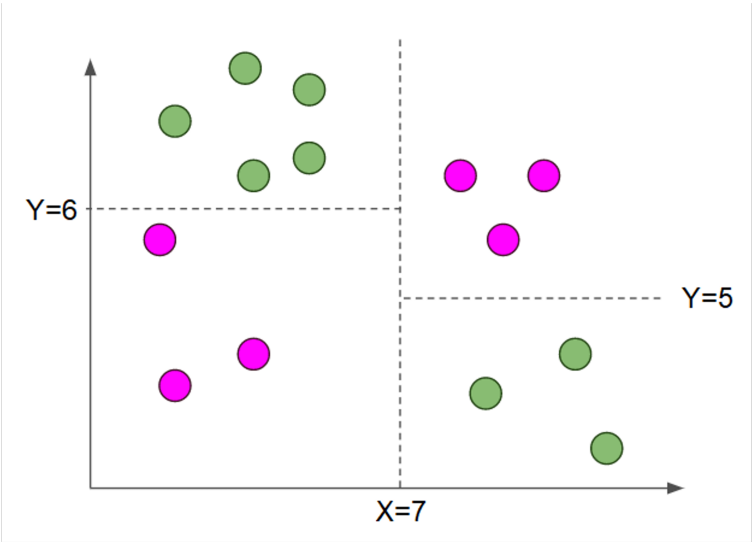
Máster Universitario en Estadística Computacional y Ciencia de Datos para la Toma de Decisiones.

Índice

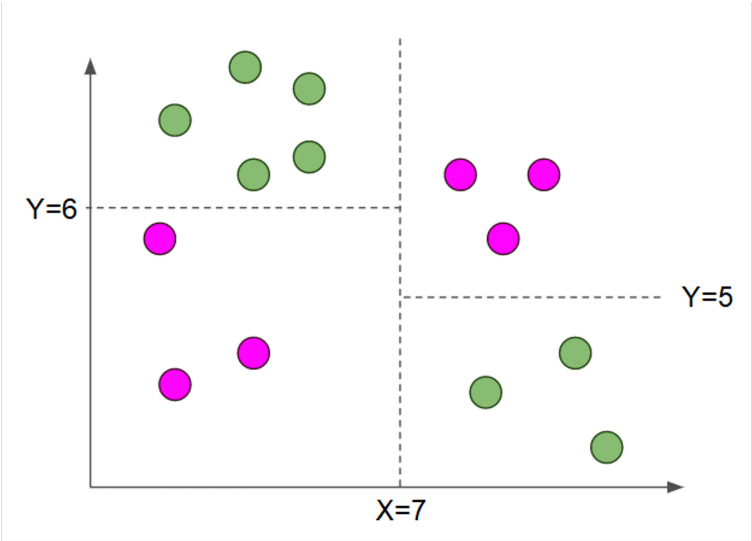
Ejemplo del problema

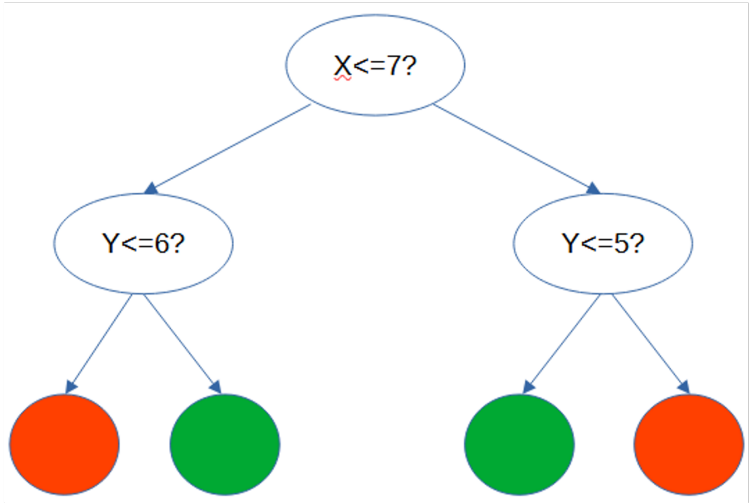
- No siempre la mejor manera de separar unos datos es mediante una recta o mediante su proximidad a los inviduos más cercanos.
- Los Árboles de Decisión permiten particionar de formas no lineales.



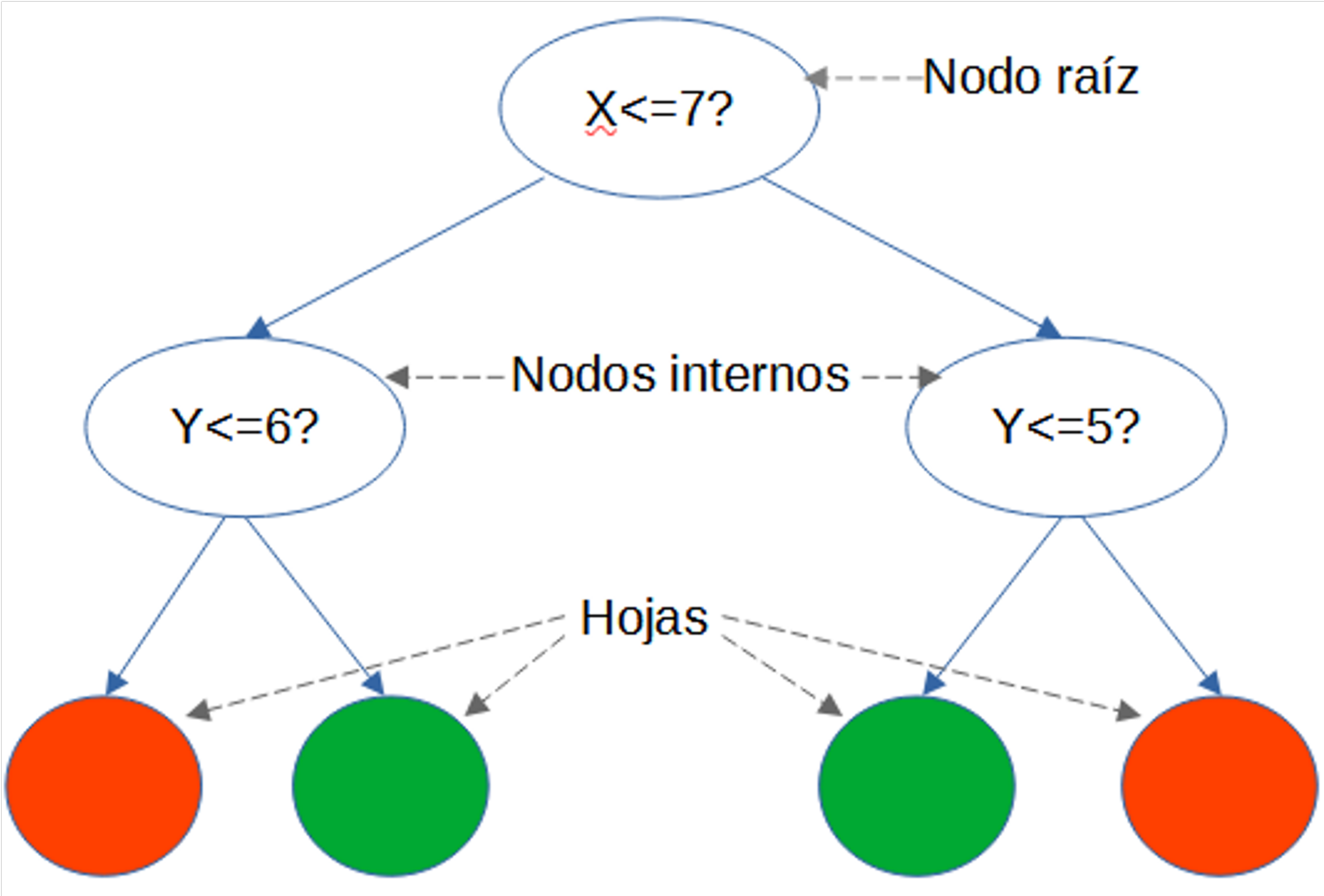


Ejemplo del problema





Ejemplo del problema



Características del Árbol de Decisión

- Los árboles de decisión son fáciles de interpretar.
- Permiten dividir los datos en subconjuntos homogéneos.
- A la hora de trazar las divisiones se busca maximizar la homogeneidad de los datos. Para ello se utilizan distintas medidas de homogeneidad: Índice de Gini, Entropía, etc.
- En caso de árboles de regresión se trata de minimizar una vez más la suma de cuadrados de los errores: $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

Índice de Gini

El índice de Gini mide la homogeneidad de un nodo en un árbol de decisión.

- Se calcula como:

$$GINI = 1 - \sum_{i=1}^{n} P_i^2$$

  donde P_i es la proporción de elementos de la clase i en el nodo.
- Un nodo completamente homogéneo (solo una clase) tiene GINI $= 0$.
- Un nodo heterogéneo (clases equilibradas) tiene GINI cercano a 0.5.

Otras Medidas de Homogeneidad

- Suma de Residuos Cuadrados (SSR): útil en árboles de regresión.
- Entropía: mide la incertidumbre de una partición.

$$H(S) = -\sum_{i=1}^{n} P_i \log_b(P_i)$$

donde P_i es la proporción de elementos de la clase i, y b es la base del logaritmo (usualmente 2).

Variables Numéricas

- Los árboles de decisión trabajan con variables categóricas.
- Las variables numéricas deben ser categorizadas.
- Métodos:
  - Uso de la mediana como umbral de partición.
  - Probar con una secuencia de valores separados unos de otros mediante un valor fijo (paso).
  - Probar umbrales mediante un método de bisección.

Sobreajuste (Overfitting)

El sobreajuste ocurre cuando un modelo se ajusta demasiado a los datos de entrenamiento, clasificando incluso el ruido.

- Consecuencias: El modelo pierde capacidad de generalización.
- Soluciones:
  - Limitar la **profundidad máxima** del árbol.
  - Definir a priori el **número mínimo de individuos** por nodo.
  - Definir a priori el **la profundidad máxima** del árbol.
  - Usar técnicas de **pruning** (poda) para eliminar ramas que no aportan mejora al rendimiento.
  - Aplicar **validación cruzada** para determinar cuándo detener el crecimiento del árbol.

Algoritmo CART (Classification and Regression Tree)

- Método forward: selecciona en cada iteración la variable que menor índice de Gini proporcione.
- Se detiene cuando se agotan las variables o se detecta sobreajuste.

Desventajas de CART

Una importante ventaja de los árboles de decisión es su facilidad de interpretar. Sin embargo, también presentan desventajas:

- Los resultados pueden variar dependiendo de la muestra de entrenamiento.
- Son susceptibles al sobreajuste.
- Algoritmo greedy (avaricioso): en cada iteración elige la mejor de las posibilidades. Ello puede tener un coste computacional alto (tardar mucho) puesto que en cada partición ha de evaluar todas las posibles opciones.

Ejercicio para realizar a mano

Individuo	Atributo 1	Atributo 2	Clase
1	Sí	125	No
2	No	100	No
3	No	70	No
4	Sí	120	No
5	No	95	Sí
6	No	60	No
7	Sí	220	No
8	No	85	Sí
9	No	75	No
10	No	90	Sí

- Obtén mediante CART manualmente el árbol de decisión de los datos proporcionados.
- Para el atributo 2, utiliza en todos los nodos la media como umbral de partición.
- Obtén la tabla de confusión sobre los propios datos y calcula su exactitud.

Otro ejercicio (largo) para realizar a mano

Individuo	Atributo 1	Atributo 2	Atributo 3	Clase
1	Sí	125	Grande	No
2	No	100	Medio	No
3	No	70	Pequeño	No
4	Sí	120	Medio	No
5	No	95	Grande	Sí
6	No	60	Medio	No
7	Sí	220	Grande	No
8	No	85	Pequeño	Sí
9	No	75	Medio	No
10	No	90	Pequeño	Sí

- Introduce dos variables dummies para el atributo 3.
- Obtén mediante CART manualmente el árbol de decisión de los datos proporcionados.
- Para el atributo 2, utiliza en todos los nodos la media como umbral de partición.
- Obtén la tabla de confusión sobre los propios datos y calcula su exactitud.


UNIVERSITAS
Miguel Hernández