UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

Árboles de Decisión

Alan Reyes-Figueroa

Aprendizaje Estadístico

(Aula 25) 13.mayo.2024

Árboles de Decisión

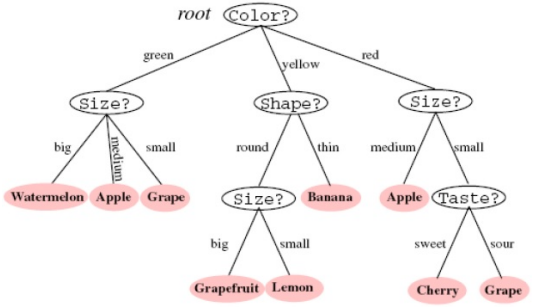
Origen: Fisher (1936). 1a generación: AID (Morgan y Sonquist, 1963), THAID (Messenger y Mandell, 1972), CHAID (Kass, 1980). 2a generación: CART (Leo Breiman *et al.*, 1984), ID3, (John Quinlan, 1986), y otros modelos más.

Un clasificador es una función:

$$\widehat{y} : \mathbf{x} \to \widehat{y}(\mathbf{x}).$$

- Nos limitamos a funciones con forma particular $\widehat{y} \in \mathcal{F}$.
- Construimos una función de costo $C(\widehat{y})$ para evaluar la calidad de una $\widehat{y}$ usando $\{(\mathbf{x}_i, y_i)\}$.
- Optimizamos C sobre $\mathcal{F}$
- Evaluamos e interpretamos la solución encontrada.

Árboles de Decisión



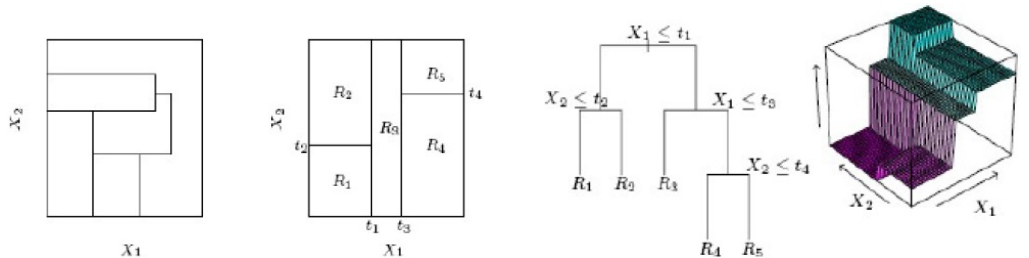
Un árbol de decisión define cierto tipo de función aditiva basada en una partición del espacio:

$$\widehat{y}(\mathbf{x}) = \sum_i c_i \mathbf{1}(\mathbf{x} \in R_i),$$

Árboles de Decisión

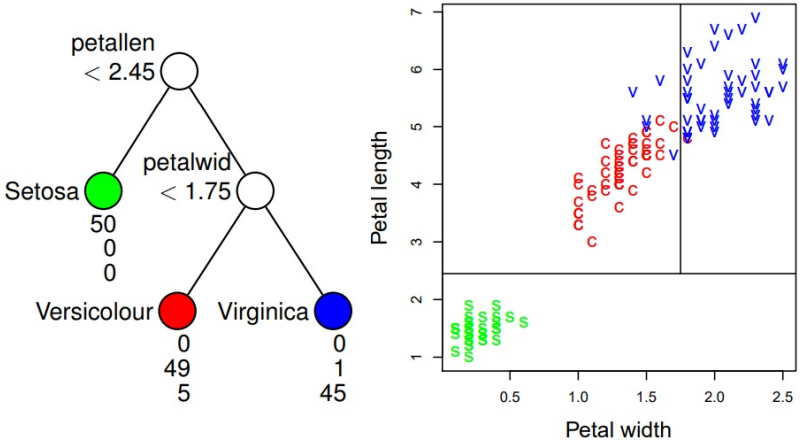
donde $c_i \in \mathbb{R}$ son ciertas constantes, y $R_i \subseteq \mathbb{R}^d$ son regiones que conforman una partición del espacio de los datos.

El caso más popular: partición binaria basada en preguntas dicotómicas ¿$x_j = c$? ó ¿$x_j < s$?



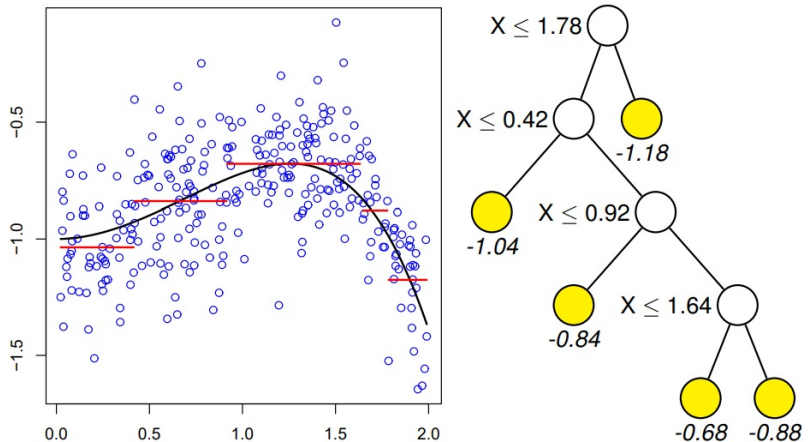
Terminología propia: raíz, hojas, nodos, profundidad.

Árboles de Decisión



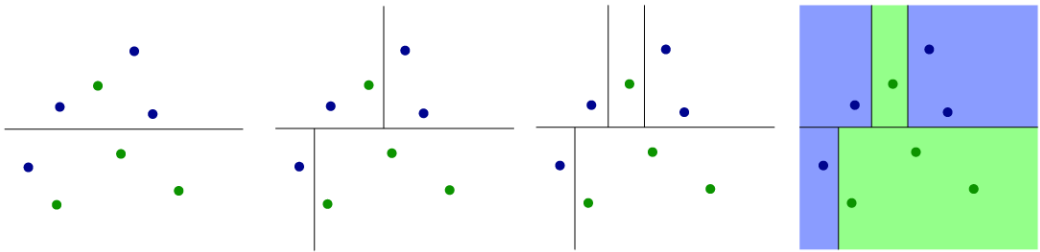
Ejemplo de árbol de decisión para clasificación de datos IRIS (Fisher, 1936).

Árboles de Decisión



Ejemplo de árbol de decisión para regresión de datos unidimensionales.

Árboles de Decisión



Mecanismo de construcción de las regiones de decisión en un árbol en $\mathbb{R}^2$.

Árboles de Decisión

¿Cuáles ventajas o desventajas ven?

Desventajas:

- función no tan simple (en contraste, *e.g.* regresión lineal),
- frontera de clasificación difícil,
- no es único (árboles diferentes pueden producir igual clasificador),
- no es robusto (sensible a cambios pequeños).

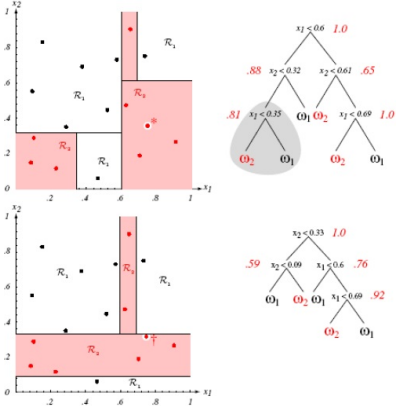
Ventajas:

- interpretable (sabemos qué hace el clasificador),
- sabemos cuáles variables son importantes,
- funciona bien en cualquier dimensión $\mathbb{R}^d$,
- efectivo al comunicar resultados mediante un grafo.

Árboles de Decisión

Obs!

- Dos árboles diferentes pueden definir la misma partición.
- Un pequeño cambio en datos puede modificar drásticamente el árbol.



Árboles de Decisión

Construcción: Punto de partida: una medida de impureza.

Supongamos que a las observaciones en una región $R \subseteq \mathbb{R}^d$ les asignamos el valor c.

1. En regresión:
$$I(R) = \frac{1}{|R|} \sum_{i:\, \mathbf{x}_i \in R} (y_i - c)^2.$$

2. En clasificación binaria:
$$\begin{aligned} I(R) &= \frac{1}{|R|} \sum_{i:\, \mathbf{x}_i \in R} \mathbf{1}(y_i \neq c). \quad \text{(error de clasificación).} \\ I(R) &= -p \log p - (1-p) \log(1-p), \;\; p = \tfrac{|\{i:\, y_i = 1\}|}{|R|}. \quad \text{(entropía)} \\ I(R) &= 1 - p^2 - (1-p)^2 = 2p(1-p). \quad \text{(impureza de Gini)} \end{aligned}$$

Árboles de Decisión

La idea aquí es aprovechar la recursividad de un árbol: un árbol es un tronco y dos sub-árboles.

Si dividimos R en R_1 y R_2, maximizamos

$$I(R) - \big(\beta_1 I(R_1) + \beta_2 I(R2)\big), \quad \text{con } \beta_i = \% \text{ de datos de } R \text{ en } R_i.$$

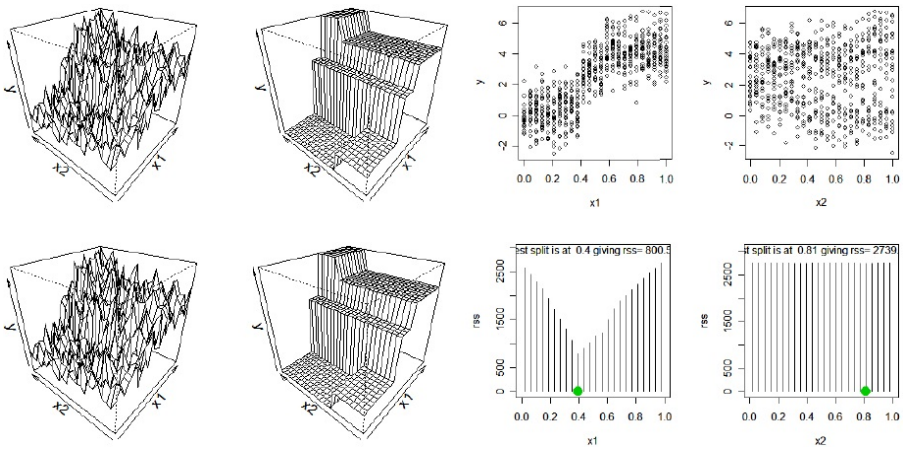
En regresión, si la partición es en base de ¿$x_j < s$?, lo anterior se reduce a:

$$\text{argmin}_{j,s} \left(\text{argmin}_{c_1} \sum_{i:\, x_i \in R_1} (y_i - c_1)^2 + \text{argmin}_{c_2} \sum_{i:\, x_i \in R_2} (y_i - c_2)^2 \right).$$

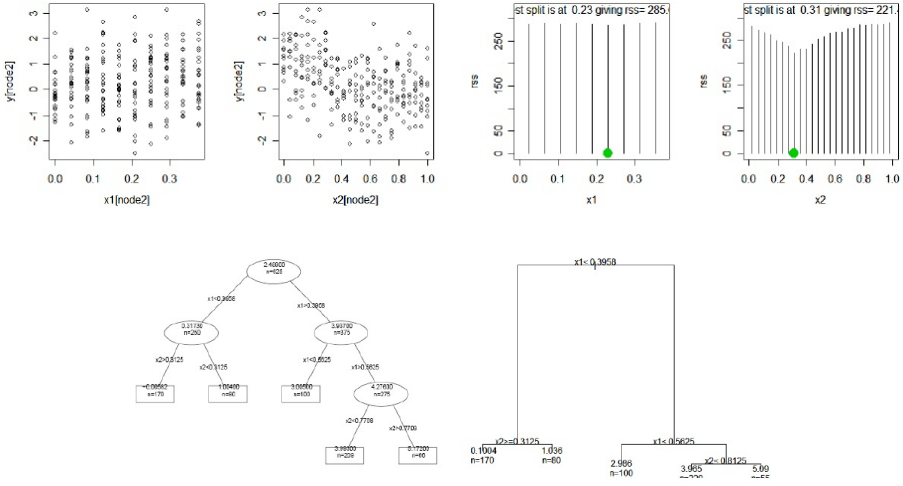
Obs! Vemos que c_i debe ser el promedio de los y_i en R_i.

Árboles de Decisión

Ejemplo:



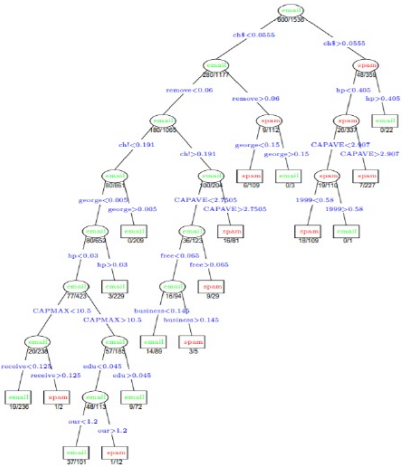
Árboles de Decisión



UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

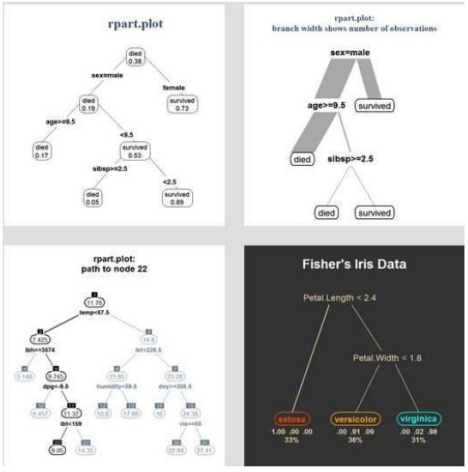
Árboles de Decisión

Ejemplo: Clasificador de correo *spam*.



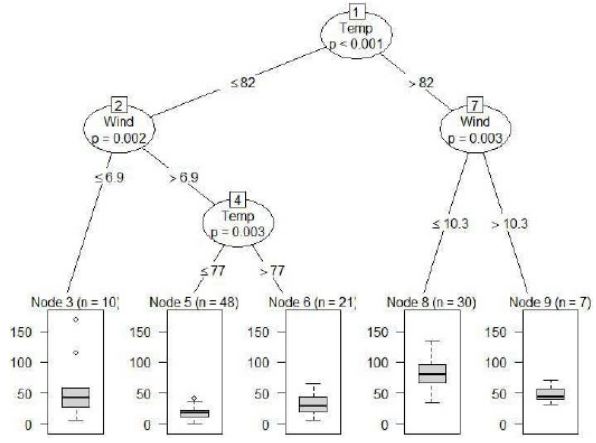
Árboles de Decisión

Ejemplos: Visualizaciones de árboles.



Árboles de Decisión

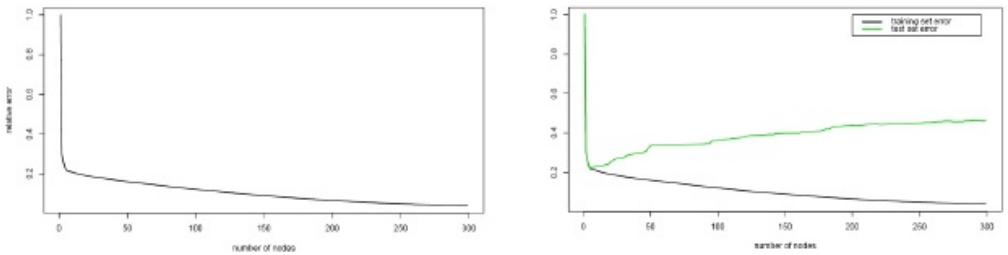
Ejemplos: Visualizaciones de árboles.



Árboles de Decisión

Pregunta: ¿Dónde detenemos la construcción del árbol?

- Entre mayor profundidad, mejor ajuste a los datos de entrenamiento...
- ...pero no necesariamente mejor ajuste al conjunto de prueba.



Típicamente se usa un parámetro de profundidad máxima *max_depth*. Además, se utiliza un valor de tolerancia $I(R_i) > \varepsilon$ para decidir si una región R_i se sigue subdividiendo.