Machine Learning para Inteligencia Artificial

Clasificación con Árboles de Decisión

Universidad ORT Uruguay

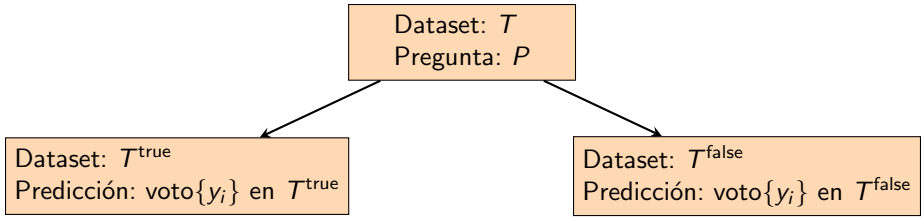
23 de Abril, 2025

Ejemplo: Jugar al tennis

Day	Outlook	Temp	Humidity	Wind	Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Para el día D15, sin información adicional, ¿qué pronosticaría?

Nos permitimos hacer una pregunta



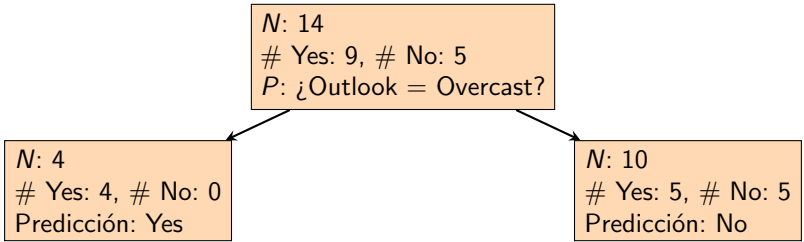
- Una pregunta P divide al dataset T en dos datasets

$$T^{\text{true}} = \{(x_i, y_i) \in T : P(x_i) = \text{true}\} \quad T^{\text{false}} = \{(x_i, y_i) \in T : P(x_i) = \text{false}\}$$

- ¿Cuál es la mejor pregunta P que uno puede hacer?

Nos permitimos hacer una pregunta

Supongamos que elegimos la pregunta **¿Outlook = Overcast?**

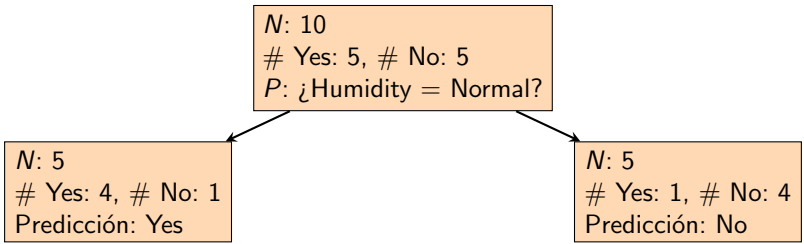


¿Cuánto es la probabilidad de error en T? Respuesta: $\frac{10}{14} \cdot \frac{5}{10} = 0.36$

Recursive Binary Splitting

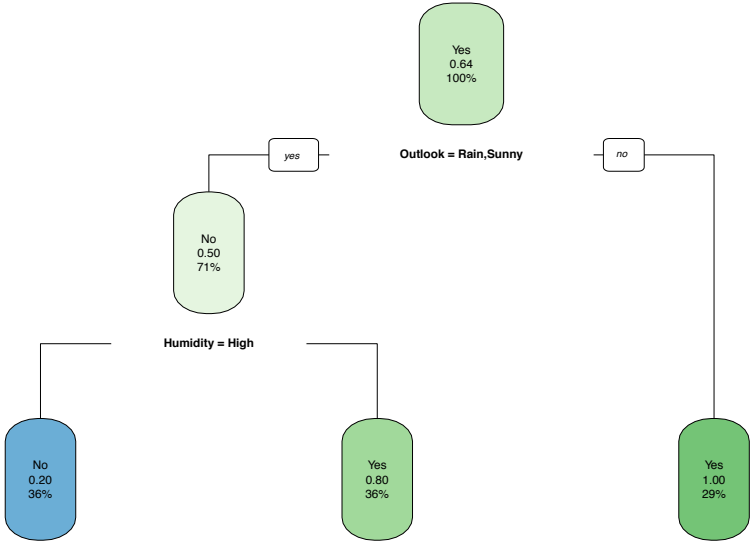
¿Y si podemos seguir preguntando?

Podemos hacer la pregunta **¿Humidity = Normal?** en el nodo derecho.

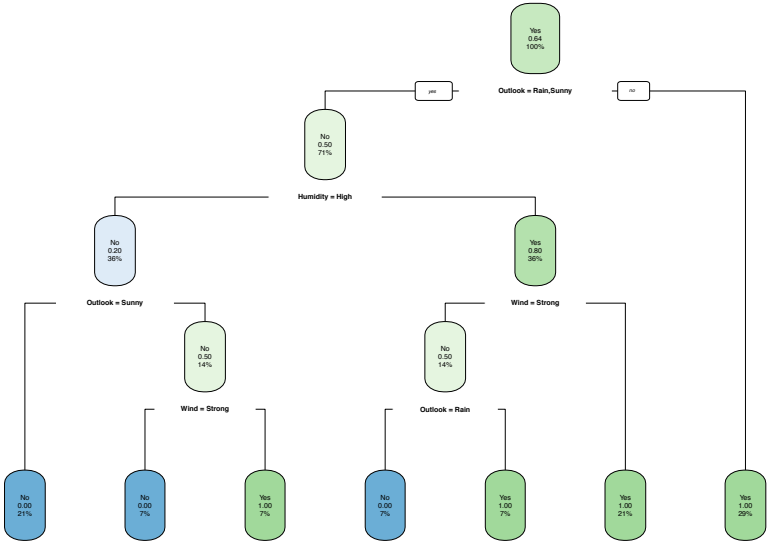


¿Cuánto es la probabilidad de error en T? Respuesta: $\frac{5}{14} \cdot \frac{1}{5} + \frac{5}{14} \cdot \frac{1}{5} = 0.14$

Obtenemos así un árbol



Otro árbol posible

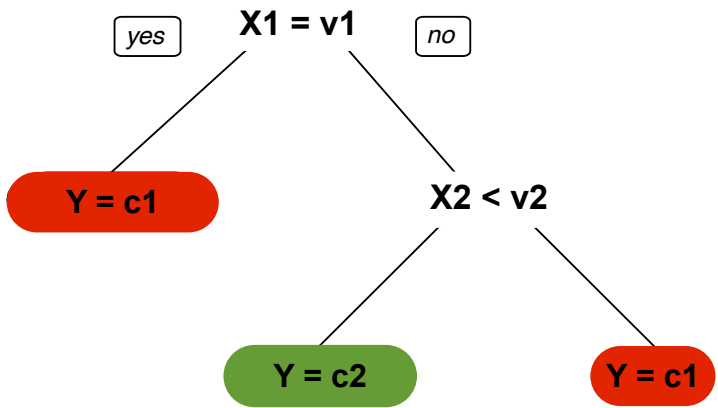


¿Qué es un árbol de decisión?

Nodos = preguntas

Arcos = respuestas

Hojas = predicción



¿Cómo se construye?

Se comienza por la raíz igual al conjunto T de datos de entrenamiento

Se repite el siguiente proceso para cada hoja H

- Si las condiciones de parada no son satisfechas:

  se elige una pregunta P para dividir H

- Se construye un arco por cada respuesta $a \in \{\text{True}, \text{False}\}$ a P

- Se desciende por cada arco a con el conjunto:

$$H_a = \{s \in H \mid P(s) = a\} \quad a \in \{\text{True}, \text{False}\}$$

La predicción en una hoja del árbol es el voto mayoritario.

Función de pérdida: el criterio de impureza de Gini

- Etiquetas $\boldsymbol{y} \in \{1, \ldots, m\}$ y distribución $p_c = \text{Prob}\left[\boldsymbol{y} = c\right]$.
- Sean $\boldsymbol{y}_1$ e $\boldsymbol{y}_2$ dos etiquetas muestreadas de la distribución (independientes).
- Pregunta: ¿Cuál es la probabilidad de que $\boldsymbol{y}_1 \neq \boldsymbol{y}_2$?
- Respuesta:

$$\text{Prob}\left[\boldsymbol{y}_1 \neq \boldsymbol{y}_2\right] = 1 - \sum_{c=1}^{m} p_c^2$$

- Esta probabilidad es máxima cuando $p_c = 1/m$ para todo $c = 1, \ldots, m$.

Costo: la impureza de Gini del árbol

Llamemos h a un árbol (hipótesis) generado por el procedimiento anterior.

La impureza de Gini de h en T es: $J_T(h) = \sum_{H \in \text{hojas}(h)} \frac{|H|}{|T|} G(y; H)$

La impureza de Gini de la etiqueta y en una hoja del árbol H es

$$G(y; H) = 1 - \sum_{c=1}^{m} \left(\frac{\#\{i \in H : y_i = c\}}{|H|} \right)^2$$

El algoritmo greedy de optimización local intenta minimizar la impureza.

Puede pasar que una buena pregunta sea descartada por una opt. local anterior.

Función de pérdida: ¿Cómo se elige la pregunta P?

- Para una pregunta P: ¿$A = v$? (o ¿$A \leq v$?) la impureza esperada después de dividir H usando P es:

$$G(P; H) = \frac{|H_{\text{True}}|}{|H|} \cdot G(y; H_{\text{True}}) + \frac{|H_{\text{False}}|}{|H|} \cdot G(y; H_{\text{False}})$$

- Se elige P correspondiente al par (A, v) que minimiza la impureza de Gini:

¿Cuándo se termina?

Condiciones estándares para no dividir una hoja del árbol

- No hay preguntas que reduzcan la impureza.
- Se alcanza una profundidad máxima para el árbol.
- Se alcanza un mínimo de observaciones en el nodo.
- Si al dividir una hoja se alcanza un mínimo de observaciones en sus hijos.

La profundidad y los mínimos de observaciones caracterizan el sesgo inductivo.

Otra forma de medir el costo: entropía de un árbol

La entropía de h en T es: $J_T(h) = \sum_{H \in \text{hojas}(h)} \frac{|H|}{|T|} S(y; H)$

La impureza de Gini de la etiqueta y en una hoja del árbol H es

$$S(y; H) = -\sum_{c=1}^{m} \frac{\#\{i \in H : y_i = c\}}{|H|} \log_2 \left(\frac{\#\{i \in H : y_i = c\}}{|H|} \right)$$

Bibliografía

- An introduction to statistical learning with applications in Python. Cap 8.
- Machine Learning - A First Course for Engineers and Scientists. Capítulo 2.3.
- Machine Learning Refined: Foundations, Algorithms, and Applications. Cap 14.