Árboles de Decisión

César Olivares

Pontificia Universidad Católica del Perú
Maestría en Informática
INF648 - Aprendizaje Automático: Teoría y Aplicaciones

2017

Árboles de decisión / Introducción

- Modelo ampliamente utilizado para **clasificación** y para **regresión**.
- El modelo aprende una jerarquía de preguntas bajo la modalidad «divide y vencerás».
- Los árboles de decisión producen modelos expresivos y fáciles de entender.

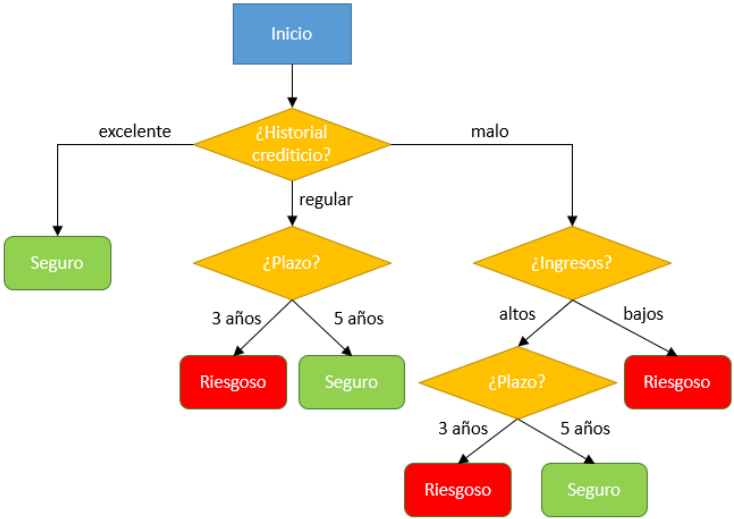
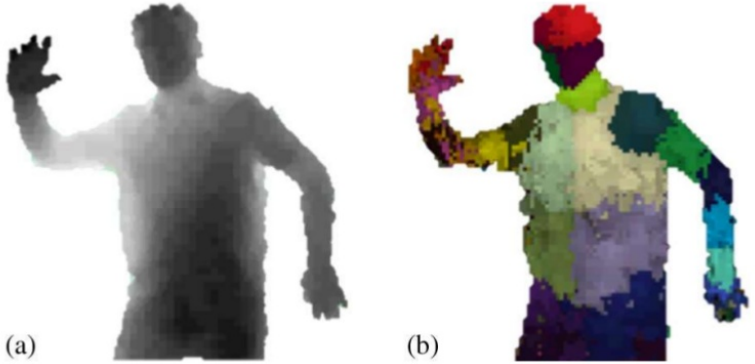


Figura 1: Árbol de decisión para evaluación de préstamo

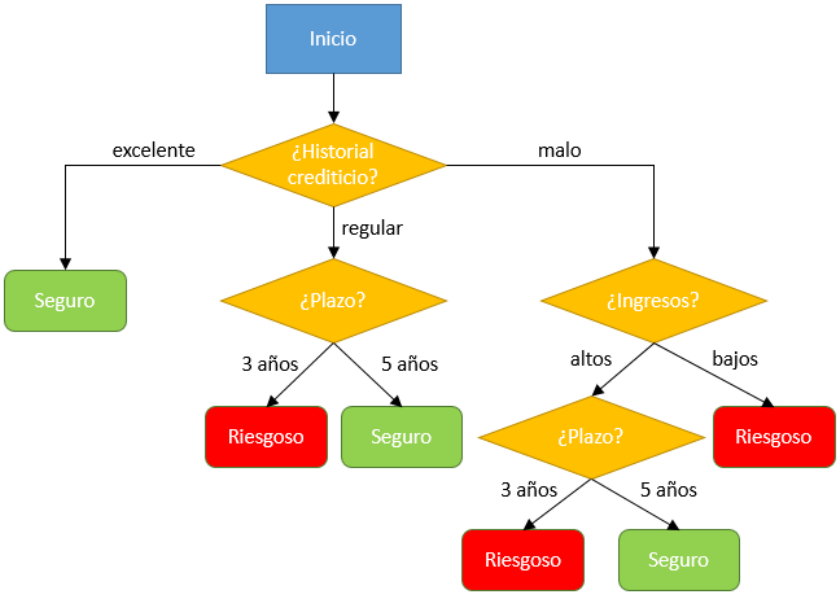
Gaming – Kinect for Xbox 360



Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., ... & Moore, R. (2013). Real-time human pose recognition in parts from single depth images. Communications of the ACM, 56(1), 116-124.

Árboles de decisión / Predicción

- Ejemplo de predicción: $x_i = (Historial = malo, Ingresos = altos, Plazo = 5\ años)$



Árboles de decisión / Modelo

Un **árbol de decisión** es un árbol tal que:

- Cada *nodo interno* (los nodos que no son *hojas*) es etiquetado con una característica.
- Cada arista que parte de un nodo interno es etiquetada con un *literal* (una afirmación sobre la característica correspondiente).
- El conjunto de literales de un nodo se denomina partición *(split)*.
- Cada hoja del árbol representa una expresión lógica, que es la conjunción de los literales encontrados en la ruta desde la raíz del árbol hasta dicha hoja.
- La extensión de dicha conjunción (el conjunto de instancias que cubre) es el *segmento del espacio de instancias* asociado a la hoja.

Árboles de decisión / Modelo

- Los árboles de decisión particionan los datos de manera **recursiva** y definen un modelo local en cada región resultante.
- Las variables de entrada y salida pueden ser tanto discretas como continuas.
- Modelo no paramétrico. (La complejidad del modelo crece con el número de datos y características. No significa que no tenga parámetros. Los modelos paramétricos, en cambio, tienen un número fijo y reducido de parámetros)
- Los nodos raíz e intermedios quedan definidos por los siguientes parámetros:
  - Variable de decisión
  - Valor umbral de decisión *(threshold)*
  - Nodos hijos
- Cada nodo hoja o terminal *(leaf node)* representa la conjunción de condiciones desde el nodo raíz *(root node)* y tiene como parámetro un valor de respuesta (la distribución sobre clases en el caso de clasificación o la respuesta media en el caso de regresión).

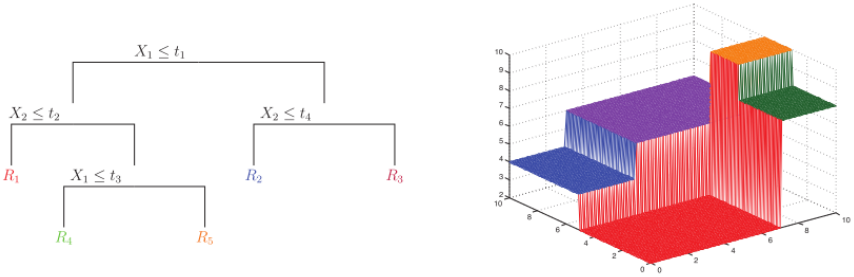


Figura 2: Árbol de regresión con entradas continuas (Murphy 2012)

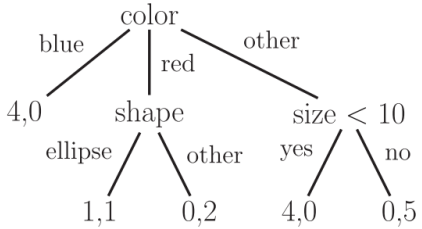


Figura 3: Árbol de clasificación binaria con entradas discretas y continuas (Murphy 2012)

Árboles de decisión / Capacidad del modelo

- Los árboles de decisión tienen expresividad máxima: son capaces de separar a la perfección cualquier conjunto de datos, con la única excepción de datos que no hayan sido etiquetados consistentemente (donde una misma instancia aparezca más de una vez con diferentes etiquetas de clase).
- La única manera de evitar el sobreajuste *(overfitting)* es poner restricciones a la complejidad del modelo.
- Encontrar la partición óptima para un conjunto de datos es un problema de complejidad exponencial NP-completo (Hyafil & Rivest 1976)
- Los árboles de decisión se entrenan usando un algoritmo «voraz» *(greedy)*.

Árboles de decisión / Algoritmo de aprendizaje

- Los diferentes algoritmos de aprendizaje comparten la siguiente estructura general:
  1. Empezar con un árbol vacío.
  2. Seleccionar la característica y literales que representen la **Mejor partición**
  3. Para cada partición resultante:
     - Si los datos son **Homogéneos** o no hay más por hacer, retornar una **Predicción**.
     - De otro modo, ir al Paso 2 y continuar (recursivamente).
- Los diferentes algoritmos se diferencian en:
  - Definición de **Homogeneidad**
  - Procedimiento para la **Predicción**
  - **Medida de calidad** para determinar la mejor partición.
  - Condiciones para detener la recursión.
  - Procesamiento posterior como p.ej. poda del árbol *pruning*

- En los casos de **clasificación**, las medidas de calidad determinan la impureza de cada nodo. A continuación las principales medidas empleadas, con sus fórmulas para clasificación binaria, donde $p^{\oplus}$ es la proporción de ejemplos positivos y $p^{\ominus}$ la proporción de ejemplos negativos.
  - **Error de clasificación.** $min(p^{\oplus}, p^{\ominus})$ - En el caso de la clasificación binaria, corresponde a la proporción de la clase minoritaria y toma valores del rango $[0, 0{,}5]$. Mientras más puro sea el conjunto de ejemplos, se cometerá menos errores de clasificación.
  - **Entropía.** $-p^{\oplus} log_2 p^{\oplus} - p^{\ominus} log_2 p^{\ominus}$ - Representa la cantidad de información, en bits, requerida para comunicar la clase de un ejemplo tomado aleatoriamente. Mientras más puro sea el conjunto de ejemplos, el mensaje es más predecible y se requiere menor información. Toma valores del rango $[0, 1]$.
  - **Índice de Gini.** $2p^{\oplus}p^{\ominus}$ - Representa el error esperado si se realiza una clasificación aleatoria: $p^{\oplus}p^{\ominus}$ es la probabilidad de un falso positivo, y $p^{\ominus}p^{\oplus}$ es la probabilidad de un falso negativo. Toma valores del rango $[0, 0{,}5]$
  - **Raíz cuadrada de Gini** $\sqrt{p^{\oplus}p^{\ominus}}$ - Transformación del índice de Gini, con la propiedad de ser invariable ante cambios en la distribución relativa de las clases (Datos no balanceados o costos FP/FN no simétricos). Toma valores del rango $[0, 0{,}5]$.
- En los casos de **regresión** se emplea el error cuadrático medio (ECM)

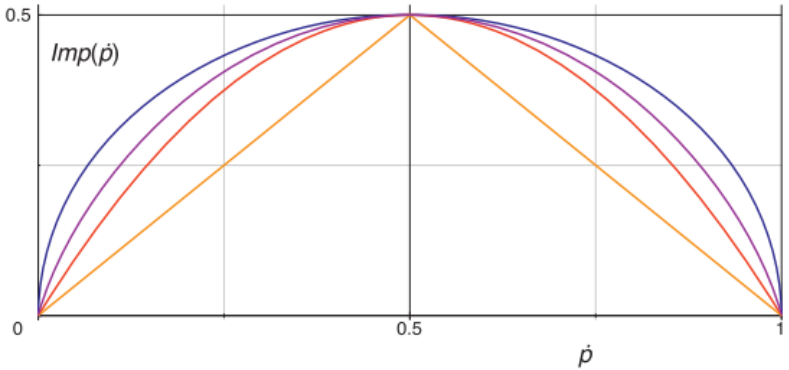


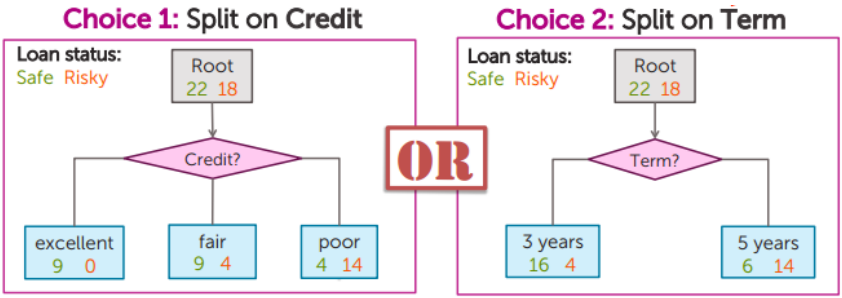
Figura 4: Funciones de impureza en relación a la probabilidad empírica de la clase positiva (Flach 2012). De abajo hacia arriba: el tamaño relativo de la clase minoritaria $min(p^{\oplus}, p^{\ominus})$ (anaranjado) ; el índice de Gini $2p^{\oplus}p^{\ominus}$ (rojo); entropía $-p^{\oplus}log_2 p^{\oplus} - p^{\ominus}log_2 p^{\ominus}$ (morado), dividida entre 2 para que coincida el punto máximo en 0,5; la raíz cuadrada (escalada) del índice de Gini $\sqrt{p^{\oplus}p^{\ominus}}$ (violeta).

- Para determinar la mejor partición de un nodo se toma en cuenta la impureza ponderada de las hojas resultantes de cada eventual partición.
- La impureza de un conjunto de hojas mutuamente exclusivas $D_1, ..., D_l$ se define como un promedio ponderado:

$$Imp(D_1, ..., D_l) = \sum_{j=1}^{l} \frac{|D_j|}{|D|} Imp(D_j)$$

- Al evaluar la calidad de una partición se acostumbra observar la ganancia de pureza $Imp(D) - Imp(D_1, ..., D_l)$ entre un nodo padre D y sus hojas $D_1, ..., D_l$.
- Cuando la medida de impureza es la entropía, la ganancia de pureza se denomina **ganancia de información**, que representa la medida de información adquirida sobre los datos luego de la partición.
- Dado que al seleccionar la mejor partición se compara siempre con el mismo nodo padre, no es necesario calcular la ganancia de pureza o información. Basta escoger la partición que resulte en una impureza ponderada más baja.
- La ganancia de pureza o de información puede ser usada para establecer una condición para detener anticipadamente la recursión, o también como criterio de poda del árbol.

- **Homogeneidad:** Las instancias son homogéneas cuando todas son de la misma clase.
- **Predicción:** Retornar la clase mayoritaria de las instancias del nodo.
- **Medida de calidad:** Índice de Gini $2p^{\oplus}p^{\ominus}$.
- **Impureza ponderada:** $Imp(D_1, ..., D_l) = \sum_{j=1}^{l} \frac{|D_j|}{|D|} Imp(D_j)$



Árboles de decisión / Condiciones para detener la recursión

1. Todos los datos pertenecen a la misma clase. (Homogeneidad perfecta)
2. Ya se particionó con todas las características. (Características discretas)
3. Limitar la profundidad del árbol.
4. Exigir una mejora mínima en la medida de calidad
5. Limitar el número de instancias por nodo.
6. Limitar el número de características a evaluar en cada partición.

Árboles de decisión / Poda

- Un algoritmo voraz sólo puede «ver» un nivel de profundidad al decidir una partición.
- Que no se reduzca suficientemente el error en una partición no significa que luego no pueda reducirse en mayor cantidad.
- La detención temprana del aprendizaje puede perder de vista «buenas» particiones que podrían darse después de otras «inútiles».
- La estrategia de **poda** consiste en dejar crecer por completo el árbol y simplificarlo después.
- Durante la poda se puede simplificar el árbol según diversos criterios que permitan limitar la complejidad del árbol, como por ejemplo el número de nodos hoja.
- De manera general se busca minimizar de manera equilibrada, tanto el error de clasificación como la complejidad del modelo:
  $Costo(T) = Error(T) + \lambda Complejidad(T)$

Árboles de decisión / Ventajas y desventajas

- Ventajas
  - Fáciles de entender y visualizar
  - Útiles para la exploración de los datos
  - Requieren poco pre-procesamiento de los datos
  - No se limitan a un tipo de dato (discretos o continuos)
- Desventajas
  - Riesgo de overfitting
  - No son lo más apropiado para variables continuas.

- CART (Breiman et al. 1984)
- ID3 (Quinlan 1986)
- C4.5 (Quinlan 1993)

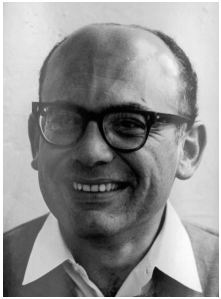


Figura 5: Leo Breiman

Figura 6: John Ross Quinlan

Árboles de decisión / Bibliografía

- Hyafil, L., & Rivest, R. L. (1976). Constructing optimal binary decision trees is NP-complete. Information Processing Letters, 5(1), 15–17.
- Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. CRC.
- Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81–106.
- Quinlan, J. R. (1993). C4.5. Programs for Machine Learning. Elsevier Inc.
- Shotton, J., Sharp, T., Kipman, A. et al. (2013). Real-time human pose recognition in parts from single depth images. Communications of the ACM, 56(1), 116–124.