Tecnológico de Monterrey

Aprendizaje automático

Árboles de decisión

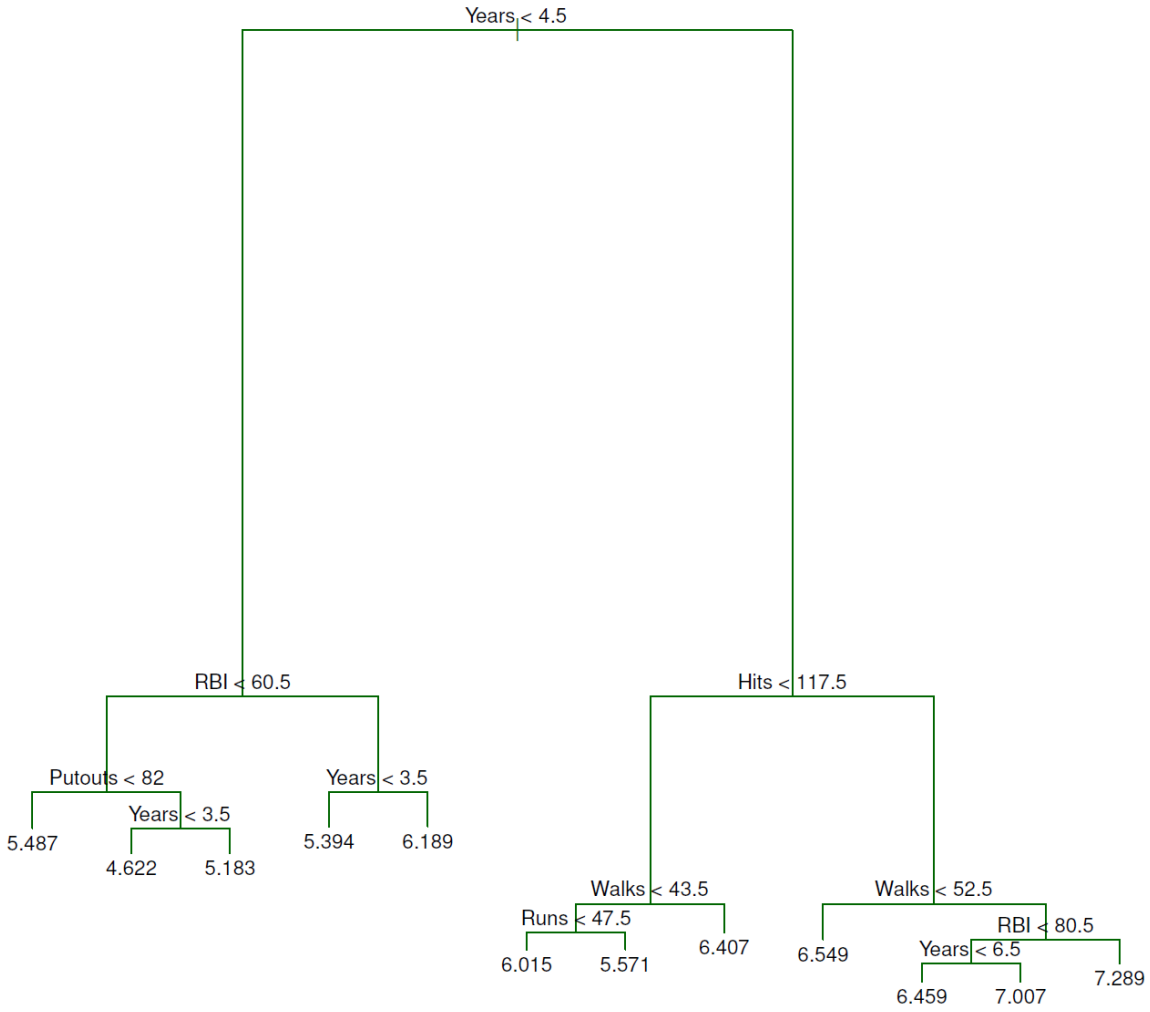
Dr. Omar Mendoza Montoya

07/08/2024

¿Qué es un árbol de decisión?

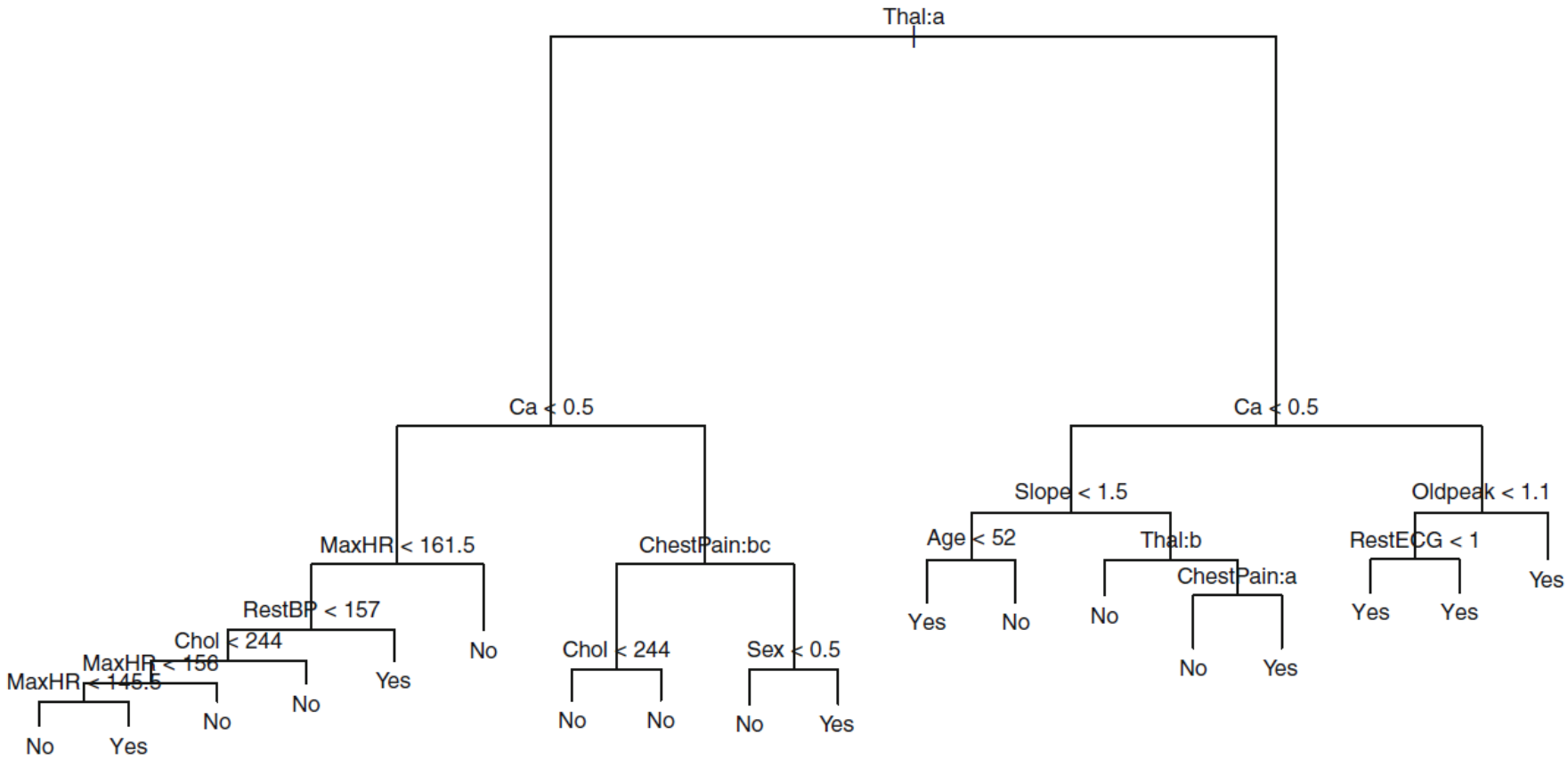
Árboles de decisión

- Los árboles de decisión son una estructura jerárquica en la cual cada nodo representa una prueba sobre un atributo o característica, cada rama representa el resultado de la prueba, y cada nodo hoja representa una etiqueta o resultado.



James, Witten, Hastie, & Tibshirani, 2021

Árboles de decisión



James, Witten, Hastie, & Tibshirani, 2021

Árboles de decisión

- Estos modelos aplican tanto a problemas de regresión como de clasificación.

- Por lo regular, por si solos no brindan una mejora respecto a otros modelos de aprendizaje automático. Sin embargo, tienen algunas propiedades que los hacen interesantes:
  - Permiten manejar de manera natural predictores que sean variables categóricas.
  - Se pueden combinar varios árboles de decisión fácilmente en un ensamble para generar clasificadores de mayor poder predictivo.

Formulación lógica

- Un árbol de decisión se puede considerar como un ejemplo de modelo lógico por la forma en que llega a una respuesta.

- En lógica de primer orden, podemos escribir que, un resultado objetivo $Goal(x)$ es verdadero si y sólo sí se cumple una serie de condiciones:

$$\forall_x Goal(x) \Leftrightarrow C_j(x)$$

donde $C_j(x)$ es una expresión candidato que involucra a los descriptores.

Formulación lógica

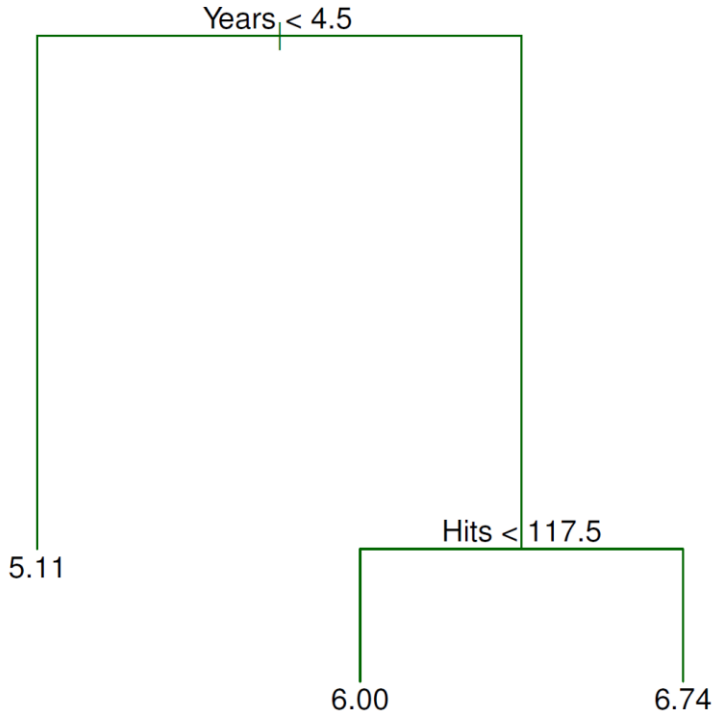
- Por ejemplo, para el caso de estar enfermo de diabetes, un árbol nos indica las condiciones para las cuales la sentencia $IsSick(x)$ es verdadera:

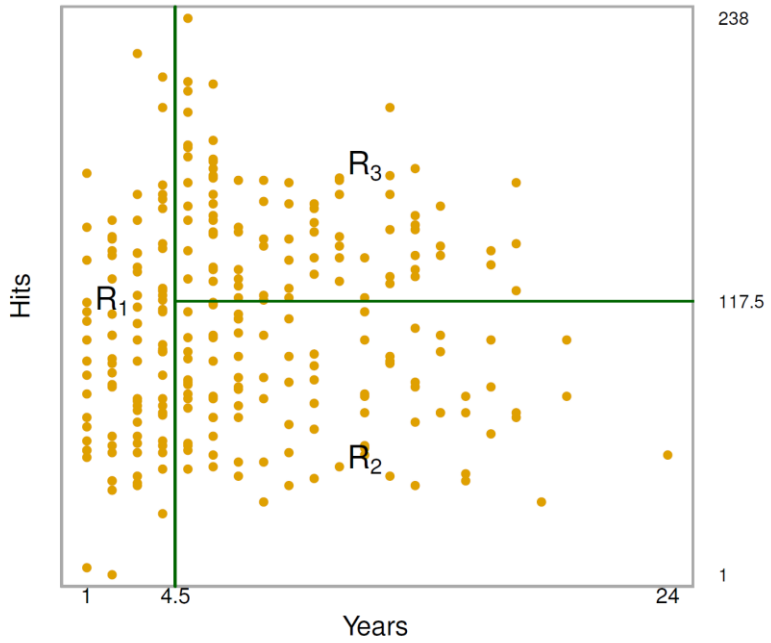
$$\begin{aligned} \forall_x IsSick(x) \quad \Leftrightarrow \quad & \big((Thal = a) \wedge (Ca < 0.5) \wedge (MaxHR < 161.5) \wedge (RestBP > 157)\big) \\ \vee \quad & \big((Thal \neq a) \wedge (Slope < 1.5) \wedge (Age < 52)\big) \\ \vee \quad & \big((Thal \neq a) \wedge (Ca > 0.5) \wedge (Oldpeak > 1.1)\big) \\ \vee \quad & \qquad\qquad\qquad\qquad\qquad\qquad \ldots \end{aligned}$$

- Esta forma de describir un modelo de aprendizaje permite estudiar dicho modelo desde el punto de vista de los sistemas basados en conocimiento.

Interpretación geométrica

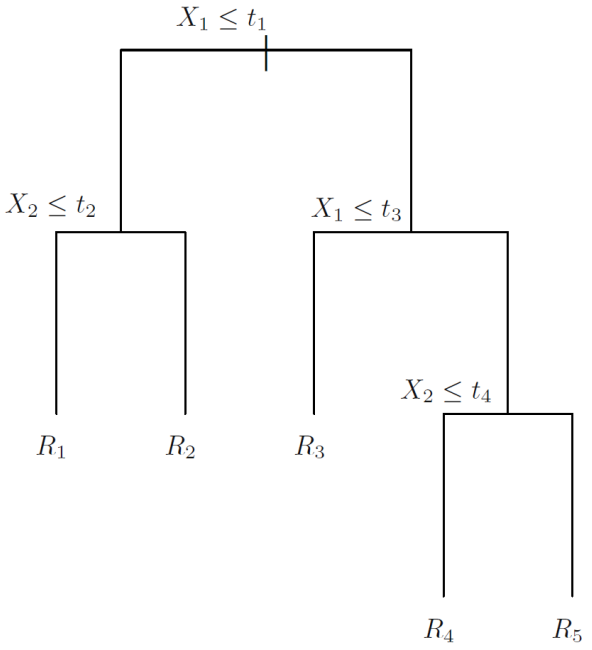
- Otra forma de entender a un árbol de decisión es considerar que el conjunto de preguntas divide el espacio en regiones, a cada una de las cuales se le asigna un valor de respuesta.

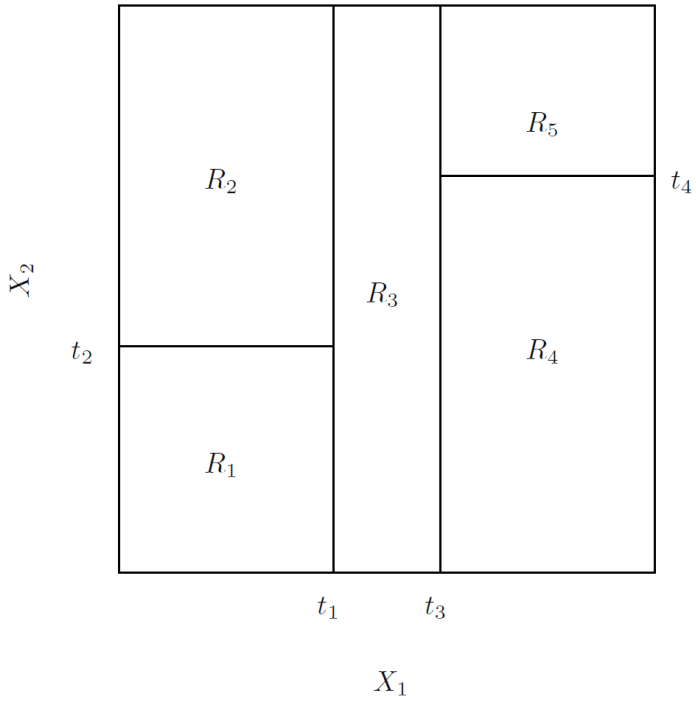


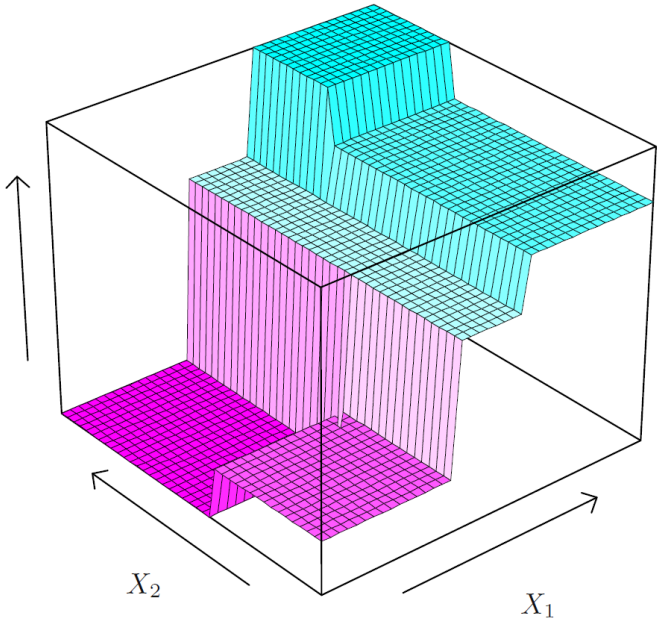


James, Witten, Hastie, & Tibshirani, 2021

Interpretación geométrica







James, Witten, Hastie, & Tibshirani, 2021

A cada región se le asigna el valor promedio de las observaciones del conjunto de entrenamiento que han caído en dicha región.

Ajuste de árboles de decisión

Ajuste del modelo para regresión

- Dado un conjunto de entrenamiento D, idealmente, si queremos dividir el espacio en J regiones $R_1, R_2, R_3, \ldots R_J$, queremos minimizar una medida de error tal como:

$$MSE(D) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) = \frac{1}{n}\sum_{j=1}^{J}\sum_{i \in R_j}\left(y_i - \hat{y}_{R_j}\right)^2$$

donde $\hat{y}_{R_j}$ es la respuesta promedio para la region R_j.

- La pregunta es, ¿qué regiones $R_1, R_2, R_3, \ldots R_J$ minimizan el error de predicción?

Para un número J de regiones, encontrar qué regiones reducirían los más posible el error se vuelve un problema difícil de tratar en términos computacionales, sobre todo si hay un número considerable de variables categórica o hay variables continuas.

División recursiva binaria (búsqueda avara)

1. Se establecen todas las formas como podemos dividir nuestras variables.
   - En el caso de variables continuas, podemos dividir su rango de valores en 100 divisiones.
   - Para variables categóricas, las divisiones las determinan las categorías.
2. Comenzando en el nodo raíz, probamos cuál de todas las posibles divisiones establecidas en el paso $\mathbf{1}$ divide todo el espacio en dos regiones R_1 y R_2 tales que el error $MSE(D)$ sea lo menor posible.
3. Para cada subregión R_i.
   A. Si el número de observaciones que cayeron a R_i está por debajo de un valor deseado, o $MSE(D_i)$ está por debajo de un valor umbral, donde D_i son las observaciones que cayeron en la región R_i, no dividir la región.
   B. De lo contrario encontrar para los datos D_i cuál sería la mejor división posible de acuerdo a las opciones dadas en el paso 1.
4. Repetir el paso 3 para cada una de las regiones nuevas creadas.

Poda (prunning)

- Es posible que el árbol haya crecido considerablemente en su construcción, llegando al caso donde cada hoja es una sola observación.

- Para mejorar el árbol y evitar que tenga una profundidad considerable, se puede aplicar alguna técnica de podado (prunning).

- Una vez que se tenga un árbol, se puede evaluar la siguiente expresión que representa tanto el error como la complejidad del árbol en término de la cantidad de nodos terminales $|T|$:

$$\frac{1}{n}\sum_{j=1}^{|T|}\sum_{i\in R_j}\left(y_i-\hat{y}_{R_j}\right)^2+\alpha|T|$$

- En este caso, $|T|$ es mejor o igual a la complejidad total del árbol encontrado con división recursiva binaria. En otras palabras, es la complejidad de un subárbol.

El término de penalización α se puede encontrar con validación cruzada. Con ello, podemos determinar qué subárbol sería adecuado para no perder aumentar el error de clasificación.

Árboles para clasificación

- En el caso de árboles para problemas de clasificación, es necesario determinar una medida de error para dividir los nodos.

- En principio, podríamos utilizar el error de clasificación:

$$\frac{1}{n}\sum_{i=1}^{N} I\left(L_i \neq \hat{L}_i\right) = \frac{1}{n}\sum_{j=1}^{J}\sum_{i \in R_j} I\left(L_i \neq \hat{L}_i\right)$$

- Sin embargo, dicha medida ha demostrado ser poco efectiva en la práctica, por lo que se han propuesto otras medidas de error.

Medidas de error para clasificación

- Índice Gini

$$G = \sum_{k=1}^{K} p_k(1 - p_k)$$

donde K es el número de clases, y p_k representa la proporción de observaciones de la clase k clasificadas correctamente en la clase k.

- Entropía

$$D = -\sum_{k=1}^{K} p_k \log(p_k)$$

Ensambles de árboles

Una de las características más interesantes sobre árboles de decisión es su simplicidad para formar ensambles que mejoran significativamente el rendimiento de un solo árbol.

Ensambles de árboles

- Bagging de árboles
  - Árboles ajustados con diferentes conjuntos de entrenamiento generados aleatoriamente muestreando con remplazo un conjunto de datos.
- Random forest
  - A diferencia de Bagging, para cada árbol del ensamble sólo se permite trabajar con un subconjunto de las variables predictoras con la finalidad que todos los árboles trabajen con diferentes subconjuntos de variables.

Ensambles de árboles

- Boosting - AdaBoost
  - Se ajustan varios árboles de manera secuencial, de tal manera que los errores de uno se propagan al que sigue para que dicho árbol intente hacerlo mejor que el anterior para dichos datos. Esto se logra modificando la función de error dando pesos a las observaciones. Para evitar sobreajuste, se utilizan árboles pequeños (de una división).

En regresión, lo equivalente que hay en cuanto a la votación de modelos de clasificación en un ensamble es simplemente el promedio de las respuestas de los modelos.

Bibliografía

- James, G., Witten, D., Hastie, T. & Tibshirani, R. (2023). *An introduction to statistical learning: with applications in Python* (2da ed.). Springer.
  - Capítulo 8

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2da ed.). Springer.
  - Capítulo 9 y 10