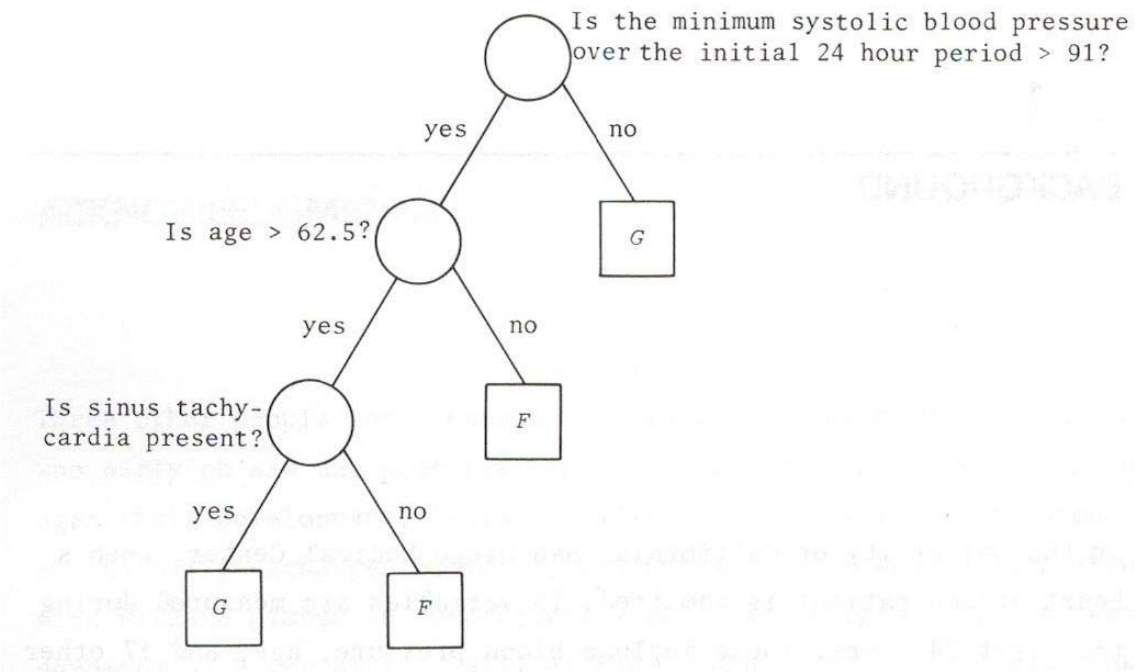
ARBOLES DE CLASIFICACIÓN Y REGRESIÓN

Planteamiento del problema

Los árboles de clasificación y regresión son un procedimiento no paramétrico de clasificación de una variable dependiente a partir de un conjunto de variables predictoras o explicativas.

La respuesta puede ser categórica (árboles de clasificación) o continua (árboles de regresión).

EJEMPLO: Se podrá establecer un método para identificar pacientes de alto nivel de riesgo con infartos de miorcardio a partir de la medición de 19 variables durante las primeras 24 horas del ingreso hospitalario. Se establece una regla de clasificación cuyas posibles asignaciones a los pacientes que ingresan son F (no alto riesgo) o G (alto riesgo). Esta regla consistió en responder a tres cuestiones como máximo.



Fuente: Breiman, Friedman, Olshen & Stone: *Classification and Regression Trees*

Objetivo

El objetivo principal será identificar qué combinaciones de variables explicativas predicen mejor la asignación de cada individuo a una

determinada categoría o valor. Para ello se parte de una muestra (homogénea respecto de la variable respuesta) que sirve para construir el árbol. A cada nodo terminal del árbol se le asigna una clase en el caso de que la variable respuesta sea categórica o un valor en el caso en que la variable respuesta sea continua.

Planteamiento formal del problema

Definimos un vector de variables:

$$\mathbf{x}=\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ \vdots \\ x_p \end{bmatrix}$$

donde x_i $i=1, 2, \ldots, p$ son variables que pueden ser de naturaleza cuantitativa o cualitativa, continua o discreta, ordinal o nominal. Suponemos que los casos o individuos particulares que denotaremos por **x,** pertenecen o se incluyen en una y solo una de la J clases de la variable respuesta. Sea C el conjunto de clases $C=\{1, 2, \ldots, J\}$.

Una regla de clasificación es una función d(x) definida sobre X (conjunto de posibles muestras o casos) que toma valores en C:

$$d: X \rightarrow C$$

$$\mathbf{x} \rightarrow d(\mathbf{x})=j \qquad j=1, 2, \ldots, J$$

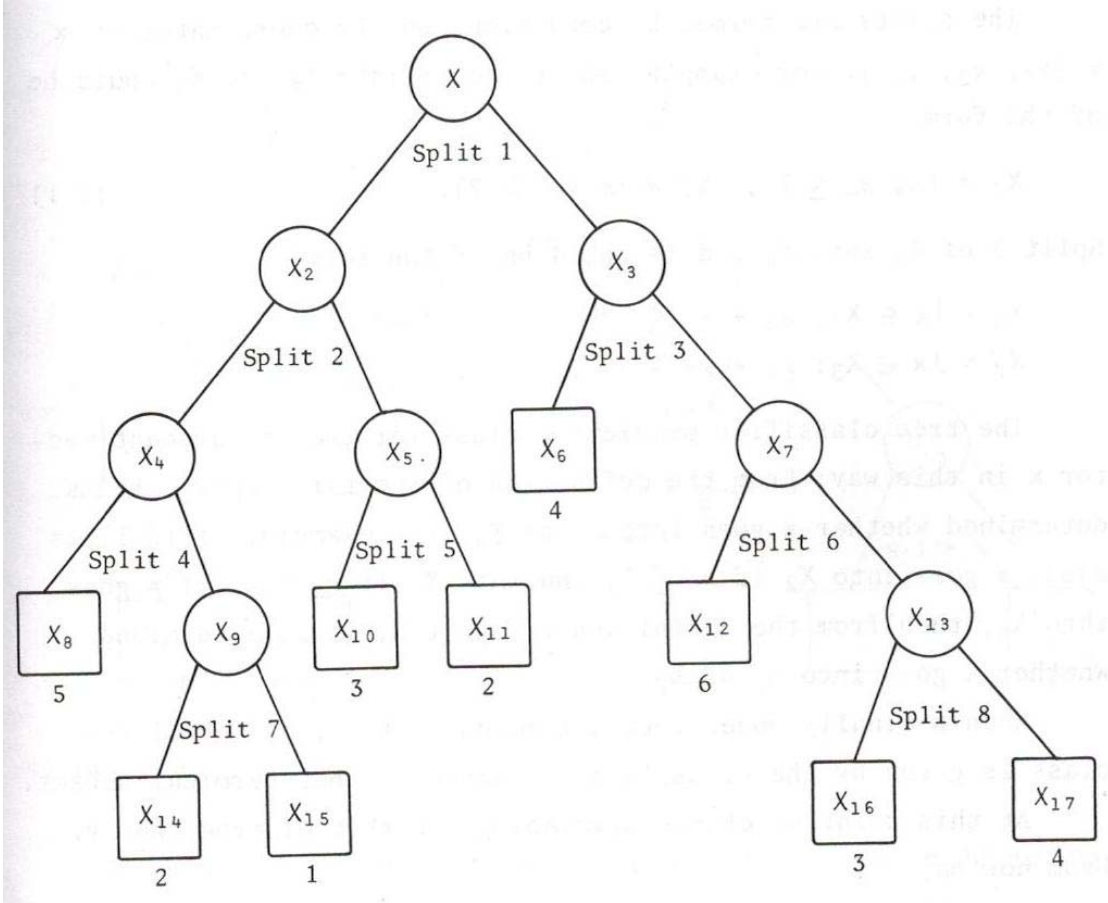
Para la construcción de la regla de clasificación utilizaremos una muestra de aprendizaje, que equivale a lo que coloquialmente denominamos experiencia, que es la que nos permite en la vida cotidiana "clasificar" situaciones, objetos, personas, acciones, etc. a partir de la experiencia acumulada.

Formalmente, una muestra de aprendizaje será una colección de casos **x** de X para las que conocemos su asignación a alguna de las clases de C. En definitiva, una muestra de aprendizaje será una colección de N pares $(\mathbf{x_i}, j)$ $i=1, 2, \ldots, N$; $j=1, 2, \ldots, J$. Para poder valorar la regla de clasificación d, podemos definir una medida de precisión R*(d). Para ello, resulta conveniente utilizar una muestra para validar la regla d.

Desde un punto de vista teórico R*(d) se definirá como la probabilidad de que la regla de clasificación d asigne correctamente una nueva muestra seleccionada del conjunto (X, C).

Estructura del árbol de clasificación

Desde un punto de vista gráfico, podemos representar un árbol de clasificación, tal y como se refleja en la siguiente imagen:



Fuente: Breiman, Friedman, Olshen & Stone: *Classification and Regression Trees*

Existen dos tipos de nodos: terminales que tienen forma cuadrada y no terminales que tienen forma circular. El primer nodo se divide en dos ($X = X_2 \cup X_3$) cuyas subdivisiones sucesivas terminan en nodos terminales. A cada uno de estos nodos terminales les asignaremos un valor j=1, 2, …, J.

Si n es el número total de divisiones resultarán n+1 nodos terminales. En la construcción de un árbol se deben considerar los siguientes aspectos:

1. Los criterios de división de los nodos
2. La decisión de declarar un nodo terminal o seguir dividiendo
3. La asignación de cada nodo terminal a una clase.

La idea fundamental es dividir cada nodo en dos subconjuntos de manera que estos sean lo más "puros" posible. La partición en cada nodo se concretará en función de un criterio de optimización del poder predictivo del árbol. Habitualmente se utilizan medidas de impurezas. Para una variable respuesta con J categorías se suelen utilizar

$$I(g) = \sum_{j=1}^{J} p_j(1 - p_j)$$

Donde p_j : proporción de individuos de la categoría j en el nodo g

I(g) es mínimo cuando todos los individuos pertenecen a una categoría, en cuyo caso I(g)=0

I(g) es máximo cuando el nodo contiene el mismo número de individuos en cada una de las categorías, en cuyo caso I(g)= Jp(1-p)

Sea un nodo t, definamos p(j/t) la proporción de casos **x** pertenecientes a la clase j, lo que implica que p(1/t)+ p(2/t)+…+ p(J/t)=1. Otra posible medida de impureza podría ser:

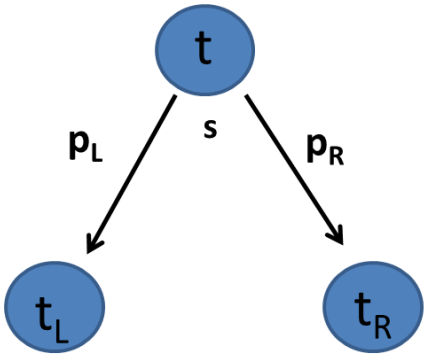
$$i(t) = -\sum_{j=1}^{J} p(j/t)\, log\{p(j/t)\}$$

- Si en el nodo sólo existieran casos de una misma clase i(t)=0 [mínimo]
- Si en el nodo existieran casos de todas las clases en la misma proporción, entonces p(1/t)=p(2/t)=…=p(J/t), lo que implicaría que i(t)=-J p(j/t) log{p(j/t)} [máximo]

La función de partición se define en términos de reducción de la impureza al pasar del nodo padre a los nodos hijos como diferencia entre la impureza del nodo padre y la suma de las impurezas de los nodos hijos [algorítmicamente se prueban todas las posibles particiones y se calcula la reducción de impureza en cada una de las particiones, seleccionándose aquella que más reducción represente].

Partiendo de un nodo t, para una división s la reducción de la impureza vendría dada por

$$\Delta_i(s,t) = i(t) - p_L i(t_L) - p_R i(t_{RL})$$



siendo p_L y p_R las proporciones de casos que caen en t_L y t_R, respectivamente a partir de la división s de t.

De todas las posibles divisiones seleccionamos aquella en la que la reducción de la medida de impureza sea máxima

$$\Delta_i(s^*, t) = \max_{s \in S} \Delta_i(s, t)$$

Un nodo será terminal cuando no es posible una reducción de la medida de impureza. Por otra parte, concretado un nodo terminal, se asignará a la clase j_k cuando $p(j_k, t) = \max_j p(j/t)$.

El conjunto S de posibles divisiones dependerá de las variables clasificadoras x_1, x_2, …, x_p y de su naturaleza (cuantitativa o cualitativa, continua o discreta, etc). Este será un problema numérico resuelto en el software que utilizaremos.

Surge en este momento una pregunta: ¿Cuándo parar en la construcción del árbol?. A mayor complejidad mayor ajuste pero se debe tener en cuenta que esto recogerá las características de la "muestra de aprendizaje" del modelo. Una medida adecuada para evaluar el poder predictivo o de clasificación es utilizar una medida del error de clasificación a partir de una "muestra de validación" del árbol [Desde un punto de vista algorítmico se suele construir un modelo lo más complejo posible y se realiza una poda hacia atrás, fijándose un "parámetro de complejidad", podándose los nodos hijos si la reducción del error de clasificación es inferior a una determinada magnitud de la complejidad del árbol].

Dado que pudiera darse el caso en que tuviéramos nodos terminales con un solo caso, se podrían optar por dos criterios de parada en la división de nodos:

1. Establecer un valor mínimo en la reducción de la medida de impureza
2. Fijar criterios de podado del árbol a posteriori, es decir una vez desarrollado el número máximo de nodos posible.

Respecto de la alternativa 1 puede ocurrir que tengamos soluciones poco satisfactorias. Si el umbral establecido en la reducción de la impureza es demasiado pequeño el árbol podría llegar a ser demasiado grande. La alternativa 2 aporta mejores resultados y es la que habitualmente emplearemos.

Las principales ventajas: 1. Carácter no paramétrico, que no requiere ninguna hipótesis sobre la distribución de la variable respuesta y variables predictoras; 2. Simple de interpretar; 3. Es posible construir árboles de decisión incluso con datos faltantes.

ARBOLES BASADOS EN INFERENCIA

Los árboles basados en la inferencia (conditional inference tree) constituyen una variante importante de los árboles de decisión tradicional. Los árboles basados en la inferencia son similares a los tradicionales pero las variables y divisiones se basan en la significatividad de algunos contrastes más que en las medidas de puridad u homogeneidad. En este caso, se siguen los siguientes pasos:

1. Se calculan los p valores para cada una de las relaciones entre cada predictor y el resultado.
2. Se selecciona el predictor con el menor p-valor.
3. Se contemplan todas posibles divisiones basadas en el predictor elegido y la variable dependiente, seleccionando la división más significativa
4. Separación de los datos en estos dos grupos y se continua el proceso para cada uno de los subgrupos resultantes hasta que las divisiones son ya significativas o el tamaño mínimo del nodo se alcance.
5. Se utiliza la función ctree() del paquete party.