ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Modelos de árboles de clasificación

Andrés G. Abad, Ph.D.

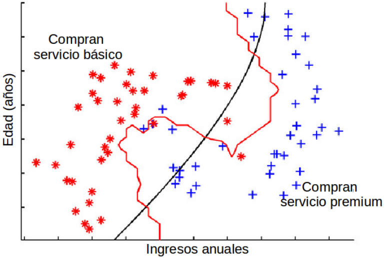
Definición del problema de clasificación I

- Un objeto $\mathbf{x} = [x_1, \ldots, x_p]$, con características x_i, pertenece exactamente a una clases $y \in \{1, 2, \ldots, C\}$.
- Asumimos que tenemos un conjunto de datos

$$\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(n)}, y^{(n)})\}$$

- Buscamos una función $\hat{f}$ que asigne $\mathbf{x}^{(i)}$ a $c^{(i)}$ lo mejor posible:

$$\hat{f} = \arg\min_{f} \mathbb{P}_{(\mathbf{x},y)}[\mathbb{1}(f(\mathbf{x}) \neq y)]$$



- Objeto $\mathbf{x}$ pertenece a una de dos clases: $\{Basico, Premium\}$
- Objeto $\mathbf{x}$ medidos en dos características: x_1 ingresos anuales, y x_2 edad en años
- Dos clasificadores $\hat{f}$'s: convexo-cuadrático (linea negra) y no-convexo (linea roja)

Métodos basados en árboles

- Los modelos basados en árboles dividen el espacio de características en rectágulos
  - Luego ajustan un model muy simple en cada rectágulo.
- Funciona para y discreta y contínua, i.e. para clasificación y regresión
- Los rectágulos son construidos con divisiones sucesivas del tipo

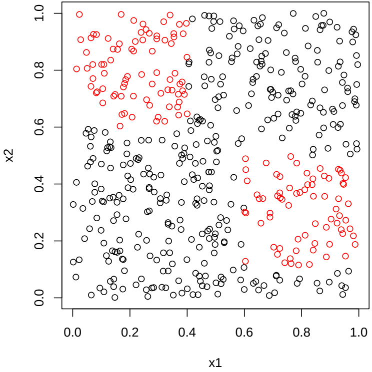
$$X_j \leq \theta \quad \text{y} \quad X_j > \theta$$

Mitad pura

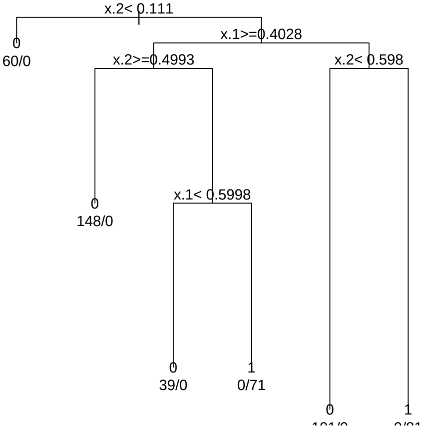
Decimos que una mitad es "pura" si contiene principalmente observaciones de una clase, en cuyo caso no continuamos con las divisiones; de lo contrario, continuamos diviendo.

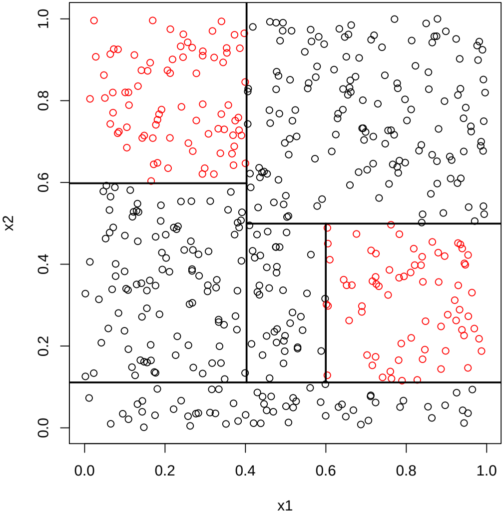
Ejemplo: árbol simple de clasificación

Ejemplo: $n = 500$ puntos en $p = 2$ dimensiones, en dos clases 0 y 1, marcadas con colores

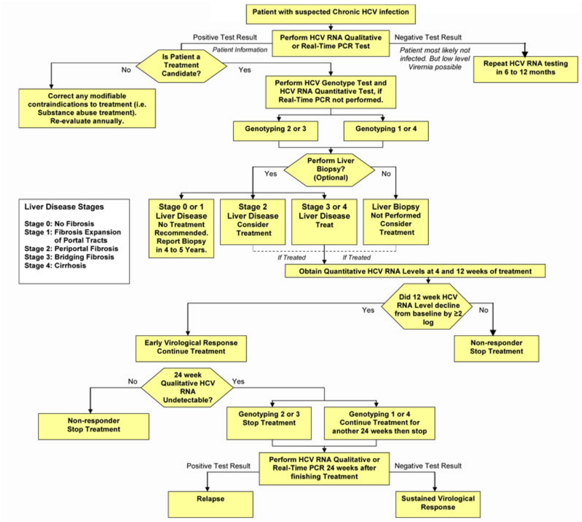


¿Dividir el espacio de características en rectágulos funcionaría aquí?


x.2< 0.111
0
60/0
x.1>=0.4028
x.2>=0.4993
x.2< 0.598
0
148/0
x.1< 0.5998
0
39/0
1
0/71
0
1


x2
x1
0.0
0.2
0.4
0.6
0.8
1.0

Ejemplo: diagrama de flujo del tratamiento de HCV



(Tomado de `http://hcv.org.nz/wordpress/?tag=treatment-flow-chart`)

Árboles de clasificación

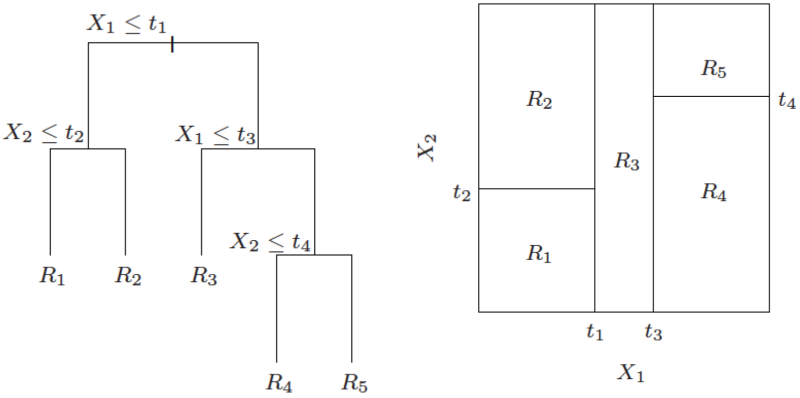
Un árbol de clasificación define m regiones (rectágulos) $R_1, \ldots R_m$, cada uno correspondiendo a una hoja del árbol.

Asignamos a cada R_j una etiqueta de clase $c_j \in \{1, \ldots K\}$.

Luego clasificamos un nuevo punto $x \in \mathbb{R}^p$ mediante

$$\hat{f}^{\text{tree}}(x) = \sum_{j=1}^{m} c_j \cdot \mathbf{1}\{x \in R_j\} = c_j \text{ siempre que } x \in R_j.$$

Ejemplo: regiones definidas por un árbol



Predicción de probabilidades de clases

- Cada región R_j contiene un subconjunto de datos de entrenamiento (x_i, y_i), $i = 1, \ldots n_j$
- La clase predicha c_j es la clase más común entre estos puntos.
- Definimos la probabilidad $\mathrm{P}(C = k | X \in R_j)$ por $\hat{p}_k(R_j)$, como

$$\hat{p}_k(R_j) = \frac{1}{n_j} \sum_{x_i \in R_j} 1\{y_i = k\},$$

  i.e., la proporción de puntos en la región que son de la clase k.
- Podemos expresar la clase predicha como

$$c_j = \underset{k=1,\ldots K}{\operatorname{argmax}} \ \hat{p}_k(R_j)$$

¿Cómo construir un árbol?

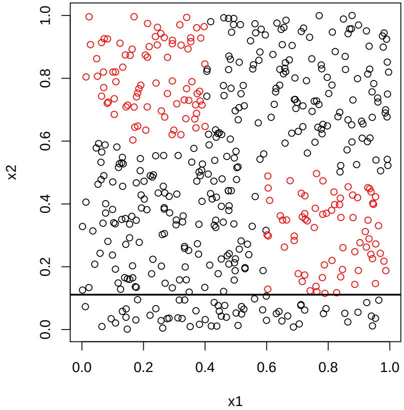
Hay dos problemas principales a considerar:

1. ¿Cómo escoger las divisiones?
2. ¿Qué tan grande construir el árbol?

El algoritmo de CART I

El algoritmo CART “Classification and Regression Trees” procede de arriba a abajo en el árbol

1. En cada etapa se selecciona la división que produce la mayor reducción en el error de clasificación (estrategia avara)
2. Se decide crecer un árbol grande y luego depurarlo al final



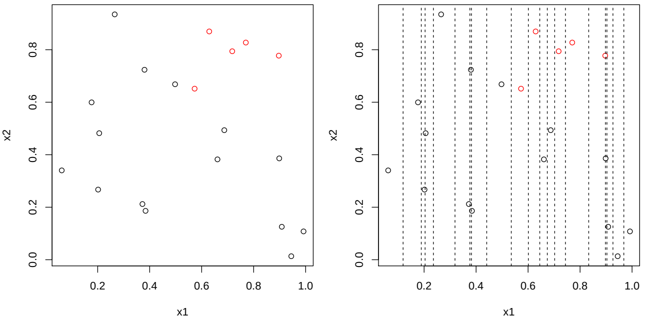
El algoritmo de CART II

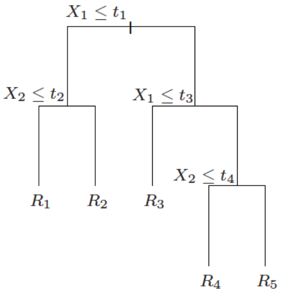
1. Empieza considerando las divisiones dadas por s en la variable j definiendo regiones:
$$R_1 = \{X : X_j \leq s\}, \text{ y } \; R_2 = \{X : X_j > s\}.$$
2. Escoja j y s de manera avara minimizando el error de clasificación
$$\underset{j,s}{\operatorname{argmin}} \left(\left[1 - \hat{p}_{c_1}(R_1) \right] + \left[1 - \hat{p}_{c_2}(R_2) \right] \right)$$
Aquí $c_1 = \operatorname{argmax}_{k=1,\ldots K} \hat{p}_k(R_1)$ es la clase más común en R_1, y $c_2 = \operatorname{argmax}_{k=1,\ldots K} \hat{p}_k(R_2)$ es la clase más común en R_2
3. Repetimos los pasos 1 y 2 recursivamente en cada nueva región R_1, R_2.

El algoritmo de CART III

¿Cómo definimos la mejor división s? ¿No hay infinitas posibilidades?

No, para dividir una región R_m en la variable j, realmente solo debemos considerar n_m divisiones posibles (o $n_m - 1$ divisiones)





- Continuando de esta manera, obtendremos un gran árbol T_0.
- Sus hojas definen regiones $R_1, \ldots, R_m$
- Podamos el árbol colapsando algunas de sus hojas en sus nodos padres

Hagamos que $|T|$ denote el número de hojas de un árbol

$$C_\alpha(T) = \sum_{j=1}^{|T|} \left[1 - \hat{p}_{c_j}(R_j)\right] + \alpha|T|$$

Buscamos un árbol $T \subseteq T_0$ que minimice $C_\alpha(T)$, podando las hojas.
Note que α es un hyper parámetro que puede ser ajustado utilizando validación cruzada

Otras medidas de impureza

Utilizamos el *error de clasificación* como medida de impureza de la región R_j,

$$1 - \hat{p}_{c_j}(R_j)$$

Pero hay otras medidas utiles también: el índice de Gini:

$$\sum_{k=1}^{K} \hat{p}_k(R_j)\left[1 - \hat{p}_k(R_j)\right],$$

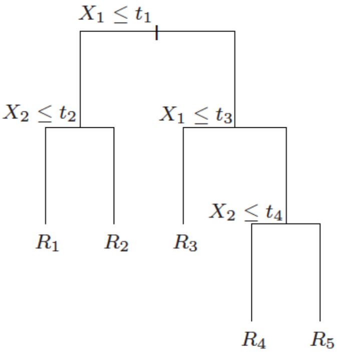
y la entropía cruzada o deviance:

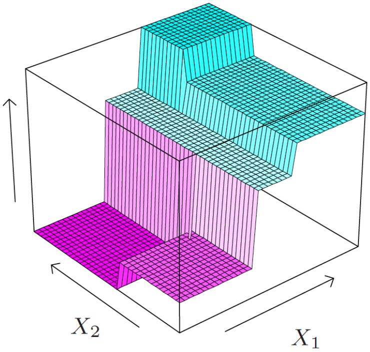
$$-\sum_{k=1}^{K} \hat{p}_k(R_j) \log\left\{\hat{p}_k(R_j)\right\}.$$

Algunas de estas medidas son más sensibles a cambios en la probabilidad de las clases. Pero, en general los resultados son similares.

Árboles de regresión

Suponga que queremos predecir una respuesta continua. Todo procede igual que antes, solo que ahora ajustamos una contante dentro de cada región.





La función de regresión estimada tiene la forma

$$\hat{f}^{\text{tree}}(x) \;=\; \sum_{j=1}^{m} c_j \cdot 1\{x \in R_j\} \;=\; c_j \;\text{ such that }\; x \in R_j,$$

donde

$$c_j = \frac{1}{n_j} \sum_{x_i \in R_j} y_i$$

Usamos ahora la función de pérdida cuadrática para decidir que región dividir.