CART's

Walter Sosa-Escudero

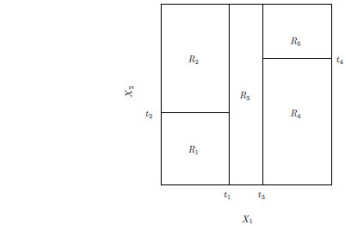
Universisad de San Andres y CONICET

Motivacion

- Modelo flexible e interpretable para la relacion entre Y y X.
- Arboles: partir el espacio de atributos en 'rectangulos', y ajustar un modelo simple para Y dentro de cada uno de ellos.
- Para que? No-linealidades, clasificacion, altas interacciones, reduccion de la dimensionalidad, prediccion.
- CART: Classification and Regression Tree

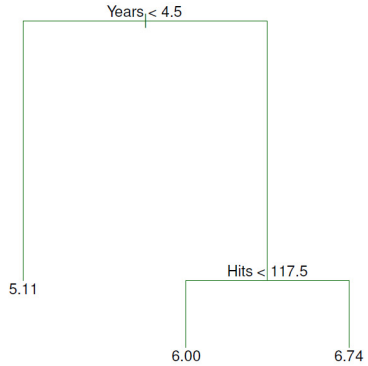
- Y es la respuesta y los inputs son X_1 y X_2.
- Partimos el espacio (X_1, X_2) en dos regiones, en base a una sola variable (particion horizontal o vertical).
- Dentro de cada region proponemos como prediccion la media muestral de Y en cada region.
- Punto: elegir la varaible y el punto de particion de manera optima (mejor ajuste global).
- Continuar partiendo las regiones, con el mismo criterio.

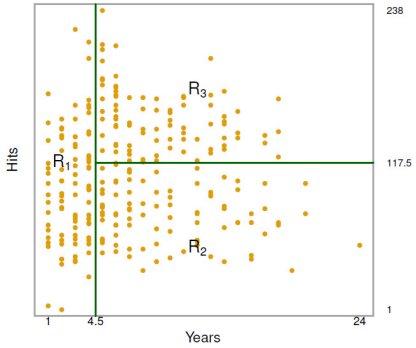
Esto implica una *particion recursiva binaria* del espacio de atributos



Fuente: James, Witten, Hastie and Tibshirani (2013)

Ejemplo: Hitters





Fuente: James, Witten, Hastie and Tibshirani (2013)

Regression Trees

Y , X, un vector de p variables de n observaciones.

Algoritmo: cual variable usar para la particion y que punto de esa variable usar para la particion.

- j es la variable de particion y el punto de particion es s.
- Definamos los siguientes 'semiplanos:

$$R_1(j,s) = \{X|X_j \leq s\} \text{ y } R_2(j,s) = \{X|Xj > s\}$$

- Problema: buscar la variable de particion X_j y el punto de particion s que resuelvan

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_1} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right]$$

- Para cada varible de particion y punto de particion, la minimizacion interna se corresponde con las medias dentro de cada region.
- El proceso se repite dentro de las regiones.

Si el arbol final tiene M regiones, el predictor es

$$\hat{f}(x) = \sum_{m=1}^{M} \hat{c}_m \, I(x \in R_m)$$

con $\hat{c}_m$ es el promedio de los y_i para todas las observaciones en la region R_m.

Complejidad y overfit en CART

Como parar?

- Un arbol demasiado extenso sobreajusta los datos (es como poner una dummy para cada observacion).
- Prunning: ajustar un arbol grande y luego podarlo (prune) usando un criterio de cost-complexity.

Tree prunning

- Indizamos los arboles con T.
- Un subarbol $T \in T_0$ es un arbol que se obtiene colapsando los nodos terminales de otro arbol (cortando ramas).
- $[T]$ = numero de nodos terminales del arbol T

Cost-complexity del arbol T:

$$C_\alpha(T) = \sum_{m=1}^{[T]} n_m \, Q_m(T) + \alpha \, [T]$$

con $Q_m(T) = 1/n_m \ \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2$ y n_m la cantidad de observaciones en cada particion.

- $Q_m(T)$ penaliza la heterogenidad (impureza) dentro de cada region, y el segundo termino la cantidad de regiones.
- Objetivo: para un α dado, encontrar la poda optima' que minimiza $C\alpha(T)$

Weakest link pruning

Mecanismo de busqueda de T_α (poda optima dado α).

- Resultado: para cada α hay un unico subarbol T_α que minimiza $C_\alpha(T)$.
- Weakest link: eliminar sucesivamente las ramas que producen el minimo incremento en $\sum_{m=1}^{[T]} n_m \, Q_m(T)$
- Idea: sacar ramas es colapsar, esto aumenta la varianza, ergo, colapsamos la particion menos necesaria.
- Esto eventualmente colapsa en el nodo inicial, pero pasa por una sucesion de arboles, desde el mas grande, hasta el mas chico, por el proceso de weakest link pruning.
- Breiman et al. (1984): T_α pertenece a esta sucesion.
- Reducir la busqueda a esta sucesion de subarboles.
- Eleccion de α: cross validation.

Classification Trees

- Y ahora toma un conjunto de valores fijos $(1, 2, ..., K)$ denotando pertenencia a alguna clasificacion (no necesariamente ordinal).
- $\hat{p}_{mk} \equiv 1/n_m \sum_{x_i \in R_m} I(y_i = k)$, proporcion de observaciones en la clase k en la region m.
- Idea: clasificar todas las observaciones de la region R_m a la clase $k(m) = \text{argmax}_k \hat{p}_{mk}$ o sea, clasificar todas las observaciones en la region m de acuerdo a la clase correspondiente a la mayoria.

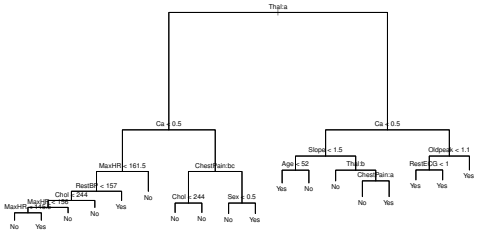
A partir de esta definicion, hay varias alternativas para la funcion de impureza en la region ($Q_m(T)$):

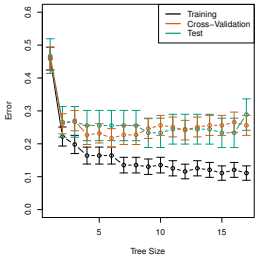
Misclassification error:	$\frac{1}{N_m}\sum_{i\in R_m} I(y_i \neq k(m)) = 1 - \hat{p}_{mk(m)}$.
Gini index:	$\sum_{k\neq k'} \hat{p}_{mk}\hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1-\hat{p}_{mk})$.
Cross-entropy or deviance:	$-\sum_{k=1}^{K} \hat{p}_{mk}\log\hat{p}_{mk}$.

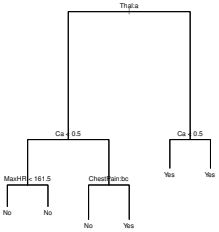
Luego el procedimiento es identico al anterior.

Heart data set

- 303 pacientes con dolor en el pecho
- $Y = 1$: problema cardiaco.
- 13 predictores: edad, genero, colesterol, etc.
- Thallium stress test: 'A thallium stress test is a nuclear imaging test that shows how well blood flows into the heart' (normal, defective)

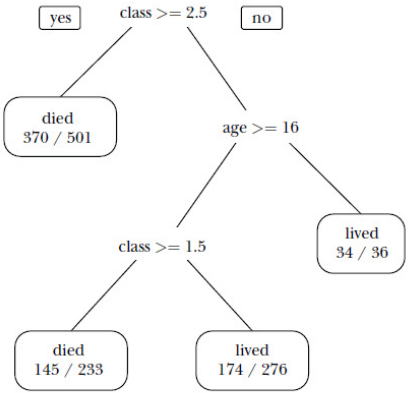

Thal:a
Ca < 0.5
Ca < 0.5
MaxHR < 161.5
ChestPain:bc
Slope < 1.5
Oldpeak < 1.1
Age < 52
Thal:b
RestECG < 1
ChestPain:a
RestBP < 157
Chol < 244
Chol < 244
Sex < 0.5
MaxHR < 156
MaxHR < 145
No
Yes


Training
Cross−Validation
Test
Error
Tree Size
0.0
0.1
0.2
0.3
0.4
0.5
0.6
5
10
15


Thal:a
Ca < 0.5
Ca < 0.5
MaxHR < 161.5
ChestPain:bc
Yes
No

Titanic

Figure 1
A Classification Tree for Survivors of the *Titanic*



Note: See text for interpretation.

Fuente: Varian(2013)

Comentarios

- Forma inteligente de representar no linealidades.
- Arriba: variables mas relevantes.
- Muy facil de comunicar. Reproduce proceso decisorio humano.
- Si la estructura es lineal, CART no anda bien.
- Lineal: las variables importan siempre (no dentro de un nodo)
- Poco robusto

Bagging, random forests y boosting al rescate