Tema 9

Arboles de regresión (tipo inductivo)

Tiene la ventaja que consume muy poca memoria y son muy rápidos
Son interpretables, puedes entender porque llegan a esa decisión

Estas se llaman "instancias"

Vamos a tener Nodos, que van a tomas una decisión binaria de si o no

Vamos a tener el nodo raiz y los nodos terminales, los del medio no los vemos

---

los arboles se pueden transformar en distintas reglas (ED)

En un arbol, le entra un imput y muy rápido te da un output

algoritmo

necesitamos un algoritmo ya que por fuerza bruta es imposible

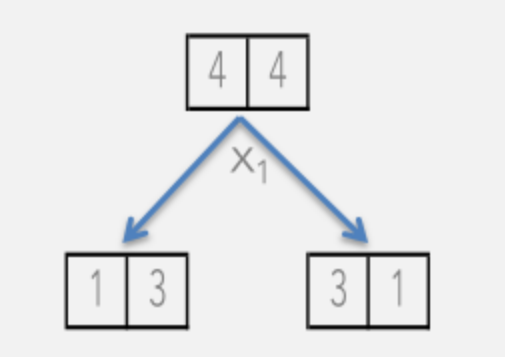
x1	x2	x3	y
0	0	0	1
0	0	1	0
0	1	0	1
0	1	1	1

si no tomo ni 0 ni 1

si tomox1 =0——-si tomox1 =1

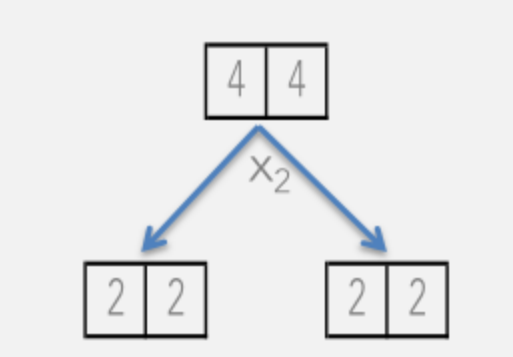
1error y 3 aciertos———

Tomo el más frecuente

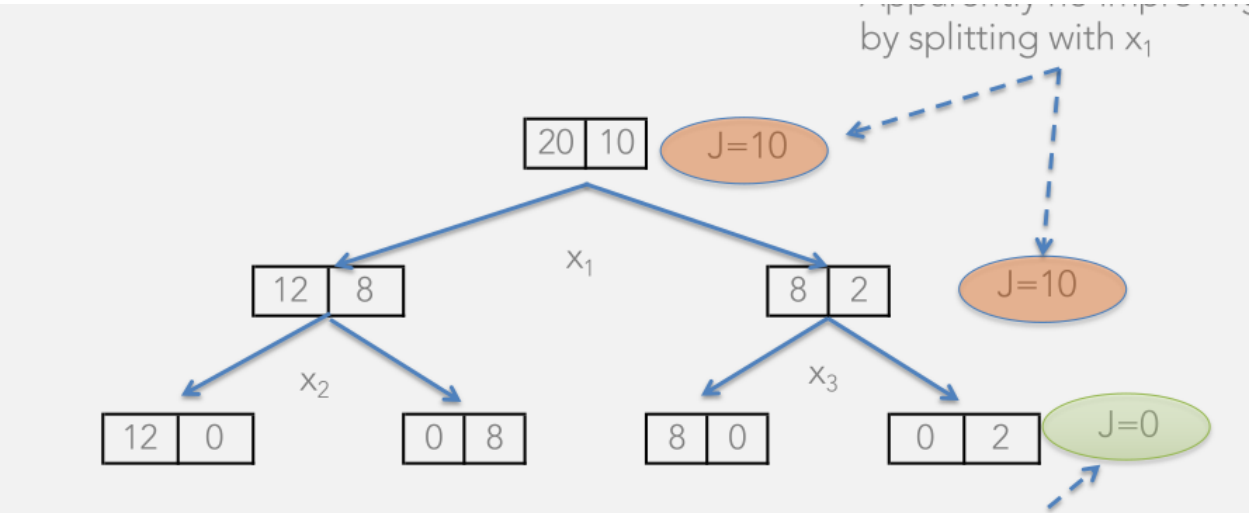


Errores totales J=2

En caso de que los 2 cuadros de abajo tengan el mismo valor, elijo el que me de la gana



Errores totales J=4



A corto plazo tiene una cantidad de error muy alto pero a largo es muy bueno (en este caso)

Tengo que mirar no solo al siguiente, sino más adelante

ID3 algorithm

sorpresa de una observacion :

$$= - \log_2 (P\ (V=v)).$$

log en base de 2 de la probabilidad de sacar un 1 (una caja entera de 0)

Si saco un 0, me llevo una sorpresa ya que tiene una probabilidad de 0 de que salga y la formula me da -00

Entropia: valor esperado de la sorpresa (1 = despipote, 50 50) (0 todo es igual, todos los elementos son de una clase)

$$H(v) = -\sum_{v} p(V = v) \times log_2(p(V = v))$$

ejm 9 positivos y 5 negativos

ese conjunto está bastante ordenado

$$= -(9/14)\log(9/14) - (5/14)\log(5/14) = 0.940$$

diapo 20, si particiono por outlook, overcast tendria (4,0)

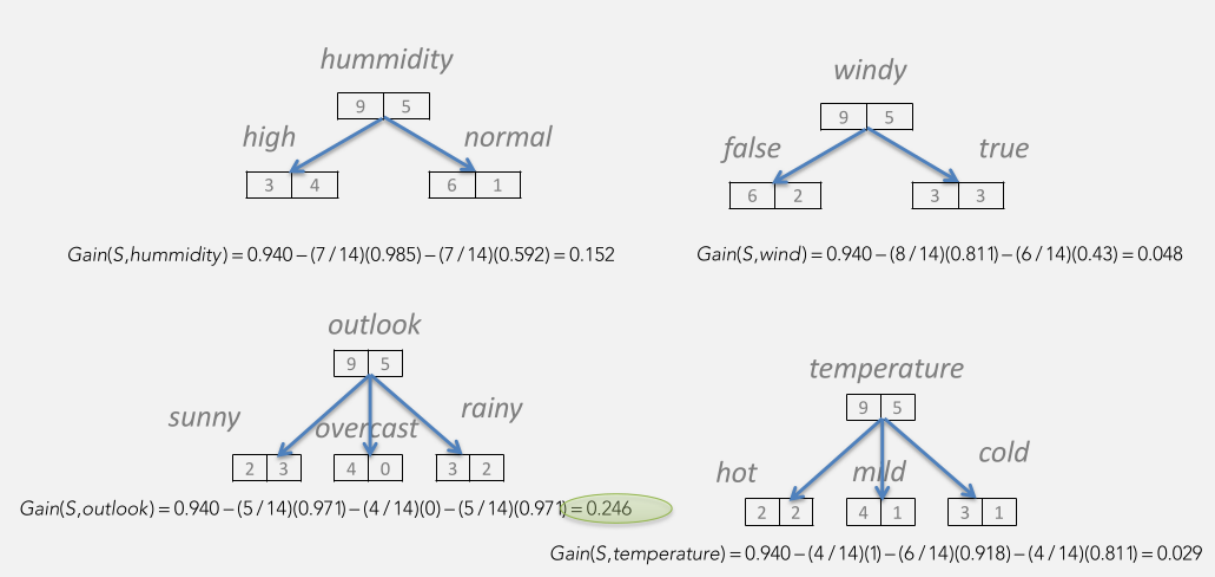
$$E(outlook=overcast) = -4/4\log(4/4) - 0/4\log(0/4) = 0$$

pondero cada una de las entropias por la cantidad de ejemplos de cada una

$$(5/14)(0.971) + (4/14)(0) + (5/14)(0.971) = 0.346 + 0 + 0.346 = 0.693$$

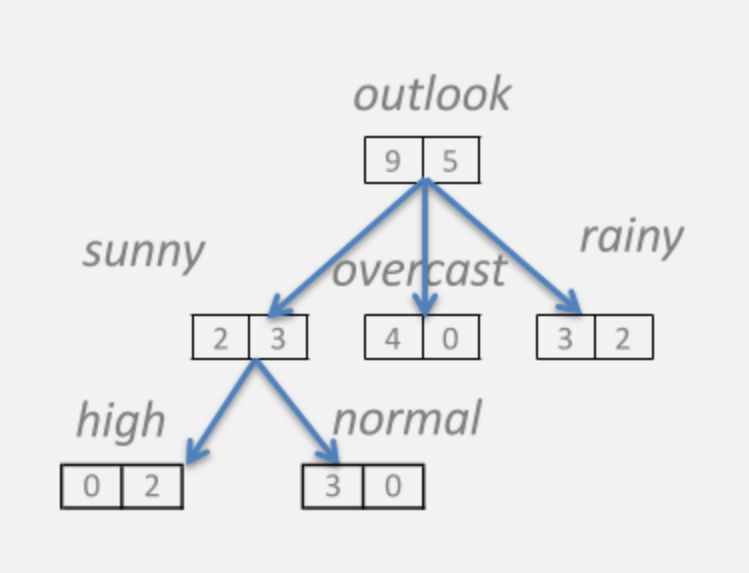
¿Por que particiono por outlook y no por otro?

miro el resultado de particionar por ese (hacer las cuentas ya que no está bien)



Calculo las entropias de cada características y luego miro la ganancia de cada uno

a continuación sigo mirando los de abajo



Cuando tenemos clasificado todos los ejemplos, podemos saber porque se ha llegado a esa conclusión

Contras:

- sobreaprendizaje

Se debe utilizar un criterio distinto para podar el arbol al que se ha empleado para expandir el arbol (calcular mse)

fin diapositiva 31, el resto nada

en vez de ganancia de entropia, podriamos usar gini index