UNIDAD TEMÁTICA 4: Algoritmos No Lineales

Trabajo de Aplicación 1

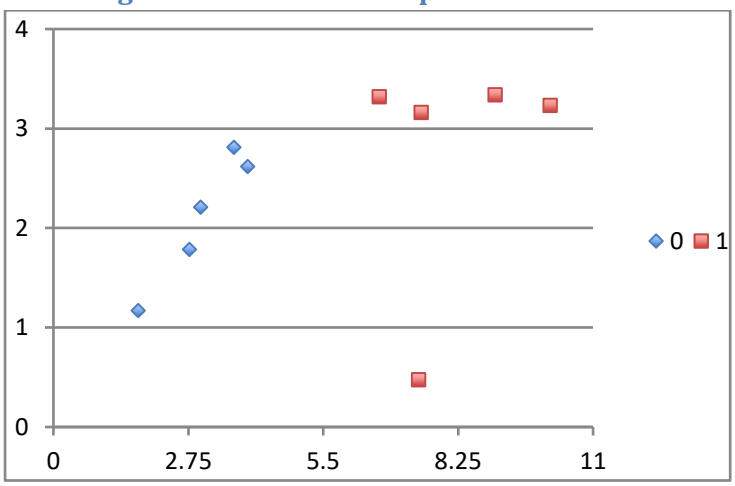
Ejercicio CART básico

- El dataset está disponible para descargar de la webasignatura: “CART-dataset.csv”
- Trabajaremos sobre un problema de clasificación binaria simple, usando CART.
- Tenemos solamente dos variables de entrada (X1 yX2) y una sóla variable de salida (Y).
- El ejemplo está diseñado para que el algoritmo encuentre al menos dos puntos de división para clasificar el conjunto de entrenamiento.

Los datos de entrada son:

X1	X2	Y
2.771245	1.784784	0
1.728571	1.169761	0
3.67832	2.812814	0
3.961043	2.61995	0
2.999209	2.209014	0
7.497546	3.162954	1
9.002203	3.339047	1
7.444542	0.476683	1
10.12494	3.234551	1
6.642287	3.319984	1

Realizar un gráfico bidimensional para los datos.



Aprendizaje del modelo CART

El modelo se aprende observando los puntos de división en los datos. Un punto de división es un valor, de un atributo (ej, el primer valor de X1 es 2.771244718).

- Dividir los datos en un punto de division involucre separar todos los datos en ese nodo, en dos grupos, a la izquierda o derecha del punto de división.
- Si estamos trabajando en el primer punto de split del árbol, entonces todo el dataset estará involucrado.
- Si estamos trabajando en un punto de división, por ejemplo, en un nivel más de profundidad, entonces solo los datos que han pasado hacia abajo en el árbol desde los nodos ancestros, y están disponibles en ese nodo, serán afectados por el punto de división.

- No nos interesa la clase a que pertenece el punto de división, sino solamente la composición de los datos asignados al subárbol izquierdo y derecho.
- Usamos una función de costo para evaluar la combinación de clases de la información de entrenamiento asignada a cada lado de la división.
- En problemas de clasificación, se utiliza la función de costo "índice GINI".

Indice Gini

- Calculamos el indice Gini para un nodo hijo:

$$G = 1 - ({p_1}^2 + {p_2}^2)$$

- Donde p1 es la proporcion de instancias en el nodo con clase 1 y p2 para la clase 2.
- La proporción se calcula así: si el grupo IZQ. Tiene 3 instancias con clase 0 y 4 con clase 1, entonces la proporción de instancias con clase 0 será 3/7 y la proporción con clase 1 será 4/7

- Abrir el archivo "gini.xlsx" para observar varios escenarios con distintas proporciones de instancias de dos clases y los correspondientes valores del indice Gini.

Gini						
escenario	**Clase 0**	**Clase 1**	**cuenta**	**Clase 0 / cuent**	**Clase 1 / cuent**	**Gini**
1	10	10	20	0.5	0.5	0.5
2	19	1	20	0.95	0.05	0.095
3	1	19	20	0.05	0.95	0.095
4	15	5	20	0.75	0.25	0.375
5	5	15	20	0.25	0.75	0.375
6	11	9	20	0.55	0.45	0.495
7	20	0	20	1	0	0

- Cuando el grupo tiene una mezcla 50-50 el Gini es 0.5, peor scenario possible
- En el ultimo ejemplo, vemos que todas las instancias caen en una sóla clase: el Gini es 0, un ejemplo de división perfecta

El cálculo del índice Gini para un punto de división seleccionado incluye calcular el Gini para cada nodo hijo y ponderar los valores por la cantidad de instancias en el nodo padre:

$$G = ((1 - ({g1_1}^2 + {g1_2}^2)) \times \frac{n_{g1}}{n}) + ((1 - ({g2_1}^2 + {g2_2}^2)) \times \frac{n_{g2}}{n})$$

- G es el Gini del punto,
- g11 es la proporción de instancias en el grupo 1 para la clase 1, g12 para la clase 2
- g21 es la proporción de instancias en el grupo 2 y clase 1, g22 grupo 2 y clase 2,
- ng1 y ng2 son los números totales de instancias en los grupos 1 y 2
- n es el numero total que queremos agrupar, del nodo padre

El objetivo al seleccionar un punto de división es evaluar el Gini de todos los posibles puntos de división y seleccionar, en forma ávida, el punto de division con el menor costo.

Primer punto de división candidato

X1 = 2.7712

- Si X1 < 2.7712 entonces IZQ
- Si X1 >= 2.7712 entonces DER

Copiar los valores de X1 e Y de entrada en otra parte de la planilla, y poner una columna a la derecha que indique "Grupo", y aplicar estas reglas para determiner el grupo correspondiente.

División #1		
2.771244718		
X1	**Y**	**grupo**
2.771244718	0	DER
1.728571309	0	IZQ
3.678319846	0	DER
3.961043357	0	DER
2.999208922	0	DER
7.497545867	1	DER
9.00220326	1	DER
7.444542326	1	DER
10.12493903	1	DER
6.642287351	1	DER

Para el grupo IZQ, las proporciones son:

- Y=0 : 1/1 = 1
- Y=1 : 0/1 = 0

Para el grupo DER, las proporciones son:

- Y=0 : 4/9 = 0.4444
- Y=1 : 5/9 = 0.5555

El total de ejemplos, del padre, es 10

CALCULAR EL ÍNDICE GINI:

$$Gini(X1 = 2.7712) = ((1 - (\frac{1^2}{1} + \frac{0^2}{1})) \times \frac{1}{10}) + ((1 - (\frac{4^2}{9} + \frac{5^2}{9})) \times \frac{9}{10})$$

$$Gini(X1 = 2.7712) = 0.4444$$

Cuentas de Clases					
	IZQ	DER	PADRE		
Y=0	1	4	5		
Y=1	0	5	5		
CUENTAS	1	9	10		
Gini					
	IZQ	DER	Gini IZQ	Gini DER	Gini izq
Y=0	1	0.444444444	0	0.44444444	0.44444444
Y=1	0	0.555555556			
Peso	0.1	0.9			

Mejor punto de división candidato

Podemos evaluar todos los puntos de división candidatos siguiendo el proceso anterior.
Solamente para ilustrar: vemos en el gráfico que los puntos podrían ser separados por una línea vertical. Esto se traduciría en un punto de división para X1 con un Gini cercano a 0.5
Un valor possible de X1 sería entonces el del ultimo ejemplo, X1 = 6.6422.

- Si X1 < 6.6422 entonces IZQ
- Si X1 >= 6.6422 entonces DER

Aplica el procedimiento anterior para agrupar las instancias y calcular el índice Gini correspondiente.

¿cuánto es el valor?

¿cómo quedaría al árbol de decisión correspondiente?