UNIVERSIDAD AUSTRAL | INGENIERÍA

Introducción a Data Mining

Trabajo Práctico Nº 2 - Árboles de Decisión

Nota: Es importante que realice estos ejercicios sin emplear software de mineria de datos

1. Considere los ejemplos de entrenamiento de la tabla siguiente para un problema de clasificación binario

ID Cliente	Genero	Tipo de Auto	Medida de Camisa	Clase
1	M	Familiar	Small	0
2	M	Deportivo	Medium	0
3	M	Deportivo	Medium	0
4	M	Deportivo	Large	0
5	M	Deportivo	Xlarge	0
6	F	Deportivo	Xlarge	0
7	F	Deportivo	Small	0
8	F	Deportivo	Small	0
9	F	Deportivo	Large	0
10	F	4 x 4	Large	0
11	M	Familiar	Large	1
12	M	Familiar	Xlarge	1
13	M	Familiar	Medium	1
14	M	Familiar	Xlarge	1
15	M	4 x 4	Small	1
16	F	4 x 4	Small	1
17	F	4 x 4	Medium	1
18	F	4 x 4	Medium	1
19	F	4 x 4	Medium	1
20	F	4 x 4	Large	1

a) Calcule el índice Gini para todos los ejemplos de entrenamiento.

b) Calcule el índice Gini para el atributo **ID Cliente.**

c) Calcule el índice Gini para el atributo **Genero.**

d) Calcule el índice Gini para el atributo **Tipo de Auto** usando *split multiway*.

e) Calcule el índice Gini para el atributo **medida de camisa** usando *split multiway*.

f) ¿Cuál es el mejor atributo para utilizar en el *Split*.?

g) Explique por qué **ID Cliente** no debería ser usado como atributo para condición de testeo, aunque tenga el menor valor de Gini.

2. Considere los ejemplos de entrenamiento de la tabla siguiente para un problema de clasificación binaria.

Instancia	a_1	a_2	a_3	Clase
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	9.0	-
8	T	F	7.0	+
9	F	T	5.0	-

a) ¿Cuál es la entropía de este conjunto de ejemplos de entrenamiento con respecto al atributo de clase?

b) Para a_3 calcule la ganancia de información (para todo posible split).

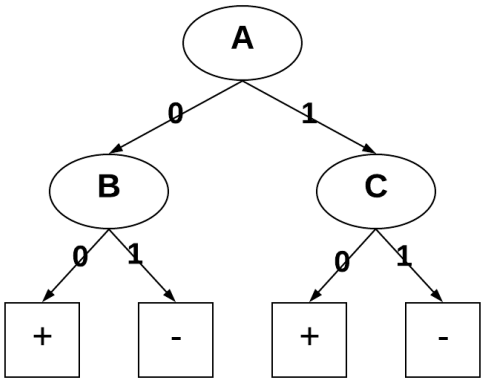
c) ¿Cuál es el atributo que produce el mejor split, de acuerdo con:

   i. ganancia de información

   ii. índice Gini?

3. Considere el árbol de decisión siguiente: **Training**

Instancia	A	B	C	Clase
1	0	0	0	+
2	0	0	1	+
3	0	1	0	+
4	0	1	1	-
5	1	0	0	+
6	1	0	1	+
7	1	1	0	-
8	1	0	1	-
9	1	1	0	+
10	1	1	1	-



Test

11	0	0	0	+
12	0	1	1	+
13	1	1	0	+
14	1	0	1	-
15	1	0	0	+

a) Calcule la tasa de error de generalización del árbol usando el método optimista.

b) Calcule la tasa de error de generalización del árbol usando el método pesimista.

c) Calcule el error de generalización del árbol usando el conjunto de testeo indicado (Error de poda reducido).

4. Un vendedor de autos está tratando de optimizar el esfuerzo que realiza con sus prospectos (clientes). Para esto ha decidido construir un árbol de decisión basándose en datos que ha recolectado sobre sus ofertas y los resultados obtenidos (compra o no compra).

   Usando la siguiente tabla y el índice Gini como criterio de impureza, se le pide:

Consulta	Tipo de Auto	Pintura	combustible	¿Compra?
C1	Deportivo	Mate	Diesel	No
C2	Deportivo	Metalizado	Nafta	Si
C3	Sedan	Mate	Diesel	No
C4	Sedan	Metalizado	Hibrido	No
C5	Familiar	Metalizado	Nafta	Si
C6	Famliar	Metalizado	Diesel	Si

   a) Identifique el mejor atributo para el nodo raíz.
   b) Construya el árbol completo.
   c) Explique la estructura del árbol.

5 Mientras que bootstrap es útil para obtener una estimación confiable de la precisión del modelo, tiene una limitación. Considere un problema de dos clases, donde hay un igual número de ejemplos positivos y negativos en los datos. Suponer también que los labels de clase para los ejemplos están generados aleatoriamente. El clasificador usado es un árbol no podado. Determinar la precisión del clasificador usando cada uno de los siguientes métodos[1]:

   a) Holdout, donde dos terceras partes de los datos son usados para training y la tercera parte restante para testing.
   b) Ten-fold cross validation.
   c) .632 bootstrap.

[1] Para la resolucion de este ejercicio es necesaria la lectura de (Kohavi, 1995)

d) De los resultados en a), b) y c) cuáles métodos proporcionan una evaluación más confiable de la precisión del clasificador.

6. Queremos seleccionar entre dos modelos predictivos, M1 y M2. Para cada modelo se han realizado diez corridas de 10-fold cross-validation, donde en la corrida *i* se uso el mismo particionamiento de datos para cada modelo.

   Las tasas de error obtenidas para M1 son 30.5, 32.2, 20.7, 20.6, 31.0, 41.0, 27.7, 26.0, 21.5, 26.0.

   Las tasas de error correspondientes a M2 son 22.4, 14.5, 22.4, 19.6, 20.7, 20.4, 22.1, 19.4, 16.2, 35.0.

   Decida si un modelo es significativamente mejor que el otro considerando un nivel de significación de 1%.