Instituto Politécnico Nacional

Escuela Superior de Cómputo

Unidad de Aprendizaje: Data Mining
Práctica: Árboles
Grupo: 3CV11
Profesora: Ocampo Botello Fabiola
Fecha de entrega: 19 de abril de 2021

Integrantes:
Aguilar Martínez Oswaldo
Arévalo Andrade Miguel Ángel

1. ¿Cuál es la diferencia entre un árbol de clasificación y un árbol de regresión?

Los árboles de regresión y clasificación son métodos de aprendizaje automático para construir modelos de predicción a partir de conjuntos de datos específicos. Los datos se dividen en varios bloques de forma recursiva y el modelo de predicción se ajusta a cada una de dichas particiones del modelo de predicción. Ahora, cada partición representa los datos como un árbol de decisiones gráfico. **La principal diferencia entre los árboles de decisión de clasificación y de regresión es que los árboles de decisión de clasificación se construyen con valores desordenados con variables dependientes.** Los árboles de decisión de regresión toman valores ordenados con valores continuos. En el caso del árbol de decisión de clasificación, para el conjunto de datos entrenado td para m número de observaciones, para una variable de clase Cl para p y l variables predictoras de Z1…. Zn. El objetivo es construir un modelo predictivo para los valores de Cl a partir de nuevos valores de Z. La Z debe dividirse en varios bloques. El algoritmo inicial que se construyó en las primeras etapas de los árboles de decisión de clasificación es THAID. El algoritmo del árbol de decisión de clasificación tiene varias características, como poda, divisiones no sesgadas, ramas / divisiones, tipo de división, antecedentes especificados por el usuario, clasificación de variables, costos especificados por el usuario, valores perdidos y ensacado y conjuntos.

2. Considerando los algoritmos: ID3, C4.5, CART y Random Forest. Realice un cuadro comparativo que considere los siguientes aspectos: descripción del algoritmo, criterio de partición que utiliza, si utiliza poda o no, tipo de datos que utiliza, ventajas y desventajas.

Algoritmo	**Descripción**	**Criterio de Partición**	**Poda**	**Tipo de datos a usar**	**Ventajas**	**Desventajas**
ID3	Es un algoritmo de aprendizaje que pretende modelar los datos mediante un árbol, llamado árbol de decisión. En este árbol los nodos intermedios son atributos de los ejemplos presentados, las ramas representan valores de dichos atributos y los nodos finales son los valores de la clase, como ya vimos al hablar de los árboles de decisión binarios.	Ganancia de información	No	**1.-** Atributos categóricos	**1.-** Se obtienen reglas comprensibles de un conjunto de datos de entrenamiento. **2.-** Es un algoritmo muy rápido. **3.-** Construye un árbol pequeño. **4.-** Sólo necesita comprobar unos cuantos datos, hasta que todos estén clasificados. **5.-** Al encontrar nodos-hoja el algoritmo no continúa, por lo que se reduce el número de comprobaciones. **6.-** Se usa todo el conjunto de datos que se le den.	**1.-** Es fácil incurrir en un sobreentrenamiento o una sobreclasificación. **2.-** Sólo se comprueba un atributo en cada paso. **3.-** Clasificar datos continuos puede ser computacionalmente muy costoso, ya que deben crearse muchos árboles para ver dónde romper la continuidad.

C4.5	Es una máquina de aprendizaje para predicciones con una variable dependiente puede llegar al objetivo deseado, basándose en los atributos de los datos disponibles. Los nodos internos son los diferentes atributos, las ramas son los posibles valores y los nodos finales (hojas) son ya la clasificación. Este método al ser iterativo va colocando los posibles valores de las características (información Ganada). Cuando todas caigan en una clasificación y ya no exista ambigüedad entonces se asigna una raíz o un nodo.	Ganancia de información	Si	**1.-** Atributos categóricos **2.-** Numéricos	**1.-** Se manejan atributos continuos. **2.-** Se mejora la eficiencia computacional. **3.-** Se lleva un control de qué tan profundo va a ser el tamaño del árbol de decisión construido. **4.-** Se evita el sobreajuste (overfitting) de datos, esto es que se aprendan a clasificar demasiado bien los datos de prueba, entonces al momento de mostrar ejemplos desconocidos, este no los clasifique de la misma forma que clasificó los ejemplos de prueba. **5.-** Manejo de atributos con diferentes valores. **6.-** Manejo de datos de entrenamiento con valores desconocidos.	**1.-** C4.5 construye ramas vacías; es el paso más crucial para la generación de reglas en C4.5. Se pueden encontrar muchos nodos con valores cero o valores cercanos a cero. Estos valores no contribuyen a generar reglas ni ayudan a construir ninguna clase para la tarea de clasificación. Más bien, hace que el árbol sea más grande y complejo. **2.-** El algoritmo C4.5 construye árboles y hace crecer sus ramas "lo suficientemente profundas para clasificar perfectamente los ejemplos de entrenamiento". Esta estrategia funciona bien con datos sin ruido. Pero la mayoría de las veces este enfoque se ajusta a los ejemplos de entrenamiento con datos ruidosos. Actualmente, hay dos enfoques que se utilizan ampliamente para evitar este ajuste excesivo en el aprendizaje del árbol de decisiones. **3.-** Susceptible al ruido
CART	Algoritmo basado en árbol que funciona examinando muchas diversas maneras de particionar o dividir localmente los datos en	Índice de Gini	Si	**1.-**Nominales. **2.-** Continuos.	**1.-** Los árboles de decisión no son paramétricos, por lo que están condicionados por el hecho de que los datos	**1.-** Sobreajuste. **2.-** Pérdida de información al categorizar variables continuas.

	segmentos más pequeños con base en diferentes valores y combinaciones de predictores. CART selecciona las divisiones de mejor rendimiento y luego repite este proceso de forma recursiva hasta encontrar el conjunto óptimo. El resultado es un árbol de decisión representado por una serie de divisiones binarias que conducen a nodos terminales que pueden ser descritos por un conjunto de reglas específicas.				de entrada tengan un tipo de distribución específica. **2.-** No requieren una especial preparación de los datos de entrada, ya que no se hacen distinciones en el tipo de distribución. **3.-** Son muy fáciles de interpretar y de entender. **4.-** Permiten que falten valores en alguna de las variables. **5.-** Son capaces de tratar con grandes bases de datos formadas por un gran número de variables.	**3.-** Inestabilidad: un pequeño cambio en los datos puede modificar ampliamente la estructura del árbol. Por lo tanto la interpretación no es tan directa como parece.
Random Forest	Suma las predicciones hechas con cada árbol de CART para determinar la predicción general del bosque, al tiempo que garantiza que los árboles de decisión no se vean afectados entre sí.	Índice de Gini	No	**1.-** Nominales. **2.-** Continuos. **3.-** Atributos categóricos	**1.-** Ser uno de los algoritmos de aprendizaje más certeros que hay disponible. Para un set de datos lo suficientemente grande produce un clasificador muy certero. **2.-** Correr eficientemente en bases de datos grandes. **3.-** Manejar cientos de variables de entrada sin excluir ninguna. **4.-** Dar estimaciones de qué variables son importantes en la clasificación. **5.-** Tener un método eficaz para estimar datos perdidos y mantener la exactitud cuando una gran proporción de los datos está perdida. **6.-** Computar los prototipos que dan información sobre la relación entre las variables y la clasificación.	**1.-** Se ha observado que **Random forests** sobreajusta en ciertos grupos de datos con tareas de clasificación/regresión ruidosas. **2.-** A diferencia de los árboles de decisión, la clasificación hecha por random forests es difícil de interpretar. **3.-** Para los datos que incluyen variables categóricas con diferente número de niveles, el random forests se parcializa a favor de esos atributos con más niveles. Por consiguiente, la posición que marca la variable no es fiable para este tipo de datos. Métodos como las permutaciones parciales se han usado para resolver el problema. **4.-** Si los datos

					7.- Computar las proximidades entre los pares de casos que pueden usarse en los grupos, localizando valores atípicos, o (ascendiendo) dando vistas interesantes de los datos. **8.-** Ofrecer un método experimental para detectar las interacciones de las variables.	contienen grupos de atributos correlacionados con similar relevancia para el rendimiento, entonces los grupos más pequeños están favorecidos sobre los grupos más grandes.

3. Considerando los siguientes criterios de selección de atributo para particionamiento: Entropy (Information Gain), Gain Ratio and Gini Index. Realice una descripción con sus propias palabras de cada uno de ellos.

Definición : La entropía es la medida de impureza , desorden o incertidumbre en un montón de ejemplos.

¿Qué hace básicamente una entropía?

La entropía controla cómo un árbol de decisión decide dividir los datos. De hecho, afecta la forma en que un árbol de decisión traza sus límites.

La ecuación de la entropía:

$$Entropy = -\sum p(X)\log p(X)$$

Gain ratio: Una modificación de la ganancia de información que reduce su sesgo en los atributos de la rama alta. La relación de ganancia tiene en cuenta el número y el tamaño de las ramas al elegir un atributo. Corrige la ganancia de información teniendo en cuenta la información intrínseca de una división. También se llama relación de división.

Gini Index:

El índice de Gini o la impureza de Gini mide el grado o la probabilidad de que una variable particular se clasifique incorrectamente cuando se elige al azar. Pero, ¿qué se entiende realmente por "impureza"? Si todos los elementos pertenecen a una sola clase, entonces se puede llamar puro. El grado del índice de Gini varía entre 0 y 1, donde 0 indica que todos los elementos pertenecen a una determinada clase o si existe solo una clase, y 1 indica que los elementos están distribuidos aleatoriamente en varias clases. Un índice de Gini de 0,5 denota elementos igualmente distribuidos en algunas clases.

Fórmula para el índice de Gini

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2$$

donde pi es la probabilidad de que un objeto se clasifique en una clase particular.

Al construir el árbol de decisiones, preferimos elegir el atributo / característica con el menor índice de Gini como nodo raíz.

4. Considerando el ejercicio de Juego de Golf, aplique el proceso de cálculo de medidas presentado en la sección 4 del artículo al ejercicio de juego de Golf.

JuegaGolf	Panorama	Temperatura	Humedad	Viento
No	Lluvioso	Caliente	Alta	FALSO
No	Lluvioso	Caliente	Alta	VERDADERO
Si	Nublado	Caliente	Alta	FALSO
Si	Soleado	Templado	Alta	FALSO
Si	Soleado	Frio	Normal	FALSO
No	Soleado	Frio	Normal	VERDADERO
Si	Nublado	Frio	Normal	VERDADERO
No	Lluvioso	Templado	Alta	FALSO
Si	Lluvioso	Frio	Normal	FALSO
Si	Soleado	Templado	Normal	FALSO
Si	Lluvioso	Templado	Normal	VERDADERO
Si	Nublado	Templado	Alta	VERDADERO
Si	Nublado	Caliente	Normal	FALSO
No	Soleado	Templado	Alta	VERDADERO

Figura 1.

En este ejercicio, hemos utilizado un conjunto de datos proporcionado por la maestra en el ejercicio de la clase pasada. para ver si una persona debería ir a jugar golf o no. La figura 1 muestra tabla con tuplas de entrenamiento del conjunto de datos. Cada atributo tomado es de un valor discreto. El atributo con etiqueta de clase JuegaGolf, tiene dos valores distintos (sí, No). Por lo tanto, hay dos clases distintas y el valor de m es igual a 2.

Asumimos:
Clase P: JuegaGolf= "sí"
Clase N: JuegaGolf= "no"

Como hay 9 sí y 5 no en el atributo JuegaGolf, por lo tanto, 9 tuplas pertenecen a la clase P y 5 tuplas pertenecen a clase N.

Paso 1:
La entropía se calcula como:

$$Entropía(D) = -\frac{9}{14}log_2(\frac{9}{14}) - \frac{5}{14}log_2(\frac{5}{14}) = 0.94$$

Paso 2: Dividir el conjunto de datos en diversos atributos.

Atributo objetivo: JugarGolf	
Atributo	**Dominio**
Panorama	Lluvioso Nublado Soleado
Temperatura	Caliente Frío Templado
Humedad	Normal Alta
Viento	Falso Verdadero

Paso 3: Se calcula la entropía en cada rama y se suman proporcionalmente para calcular la entropía total:

$$E(T,X) = \sum_{c \in x} p(c)E(c)$$

				Count
Panorama	Lluvioso	JuegaGolf	No	3
			Sí	2
	Nublado	JuegaGolf	Sí	4
	Soleado	JuegaGolf	No	2
			Sí	3

$E(Soleado) = E(No, Sí) = E(2,3) = (-\frac{2}{14}log_2(\frac{2}{14})) + (-\frac{3}{14}log_2(\frac{3}{14}))$

$= (-0.14log_2(0.14)) + (-0.21log_2(0.21))$

$= 0.40 + 0.47 = 0.97 \rightarrow Entropía\ Soleado$

Para panorama:

$E(JugarGolf, Panorama) = P(Lluvioso) * E(3,2) + P(Nublado) * E(4,0) + P(Soleado) * E(2,3)$

- $Lluvioso: P(5/14) = 0.36\ E(3,2) = 0.44 + 0.53 = 0.971$
- $Nublado: P(4/14) = 0.29\ E(4,0) = 0$
- $Soleado: P(5/14) = 0.36\ E(3,2) = 0.971$

$E(JugarGolf,\ Panorama) = 0.36 * 0.971 + 0.29 * 0 + 0.36 * 0.971 = 0.35 + 0 + 0.35 = 0.7$

Paso 4: Ganancia de información

$Gain = 0.94 - 0.7 = 0.247$

Para temperatura:

				Count
Temperatura	Caliente	JuegaGolf	No	2
			Sí	2
	Frío	JuegaGolf	No	1
			Si	3
	Templado	JuegaGolf	No	2
			Sí	4

$E(JugarGolf, Temperatura) = P(Caliente) * E(2,2) + P(Frío) * E(1,3) + P(Templado) * E(2,4)$

$- Caliente: P(\frac{4}{14}) = 0.29\ E(2,2) = 0.5 + 0.5 = 1$

$- Frío: P(\frac{4}{14}) = 0.29\ E(1,3) = 0.5 + 0.31 = 0.81$

$- Templado: P(\frac{6}{14}) = 0.43\ E(2,4) = 0.53 + 0.39 = 0.92$

$E(JugarGolf, Temperatura) = 0.29 * 1 + 0.29 * 0.81 + 0.43 * 0.92 = 0.29 + 0.2349 + 0.39 =$

$Gain = 0.94 - 0.91 = 0.03$

Para humedad:

				Count
Humedad	Alta	JuegaGolf	No	4
			Sí	3
	Normal	JuegaGolf	No	1
			Si	6

$E(JugarGolf, Humedad) = P(Normal) * E(1,6) + P(Alta) * E(4,3)$

$- Normal: P(\frac{7}{14}) = 0.5\ E(1,6) = 0.4 + 0.19 = 0.59$

$- Alta: P(\frac{7}{14}) = 0.5\ E(4,3) = 0.46 + 0.52 = 0.98$

$E(JugarGolf, Humedad) = 0.5 * 0.59 + 0.5 * 0.98 = 0.295 + 0.49 = 0.785$
$Gain = 0.94 - 0.785 = 0.155$

Gain Ratio

$Gain(JugarGolf,\ Día) = (-\frac{5}{14}log_2(\frac{5}{14}) - \frac{9}{14}log_2(\frac{9}{14})) = 0.94028$

$Split(JugarGolf,\ Día) = 14(-\frac{1}{14}log_2(\frac{1}{14})$=3.8073 Gain Ratio($JugarGolf$, Día)= $\frac{0.94028}{3.8073}$ =0.2469

Split($JugarGolf$, Panorama) =$-\frac{5}{14}log_2(\frac{5}{14}) * 2 + (-\frac{4}{14}log_2(\frac{4}{14})) = 1.5774$ Gain Ratio($JugarGolf$,Panorama)= $\frac{0.247}{1.5774} = 0.1565$

Gain Ratio($JugarGolf$, Humedad)= $\frac{0.155}{1}$=0.155

Gain Ratio($JugarGolf$, Temperatura)= $\frac{0.03}{1.362}$=0.0220

Gain Ratio($JugarGolf$, Viento)= $\frac{0.048}{0.985} = 0.0487$

Gini Index

$Gini(JugarGolf, Panorama)$

$- Lluvioso: 1 - (\frac{3}{5})^2 - (\frac{2}{5})^2 = 0.48$

$- Nublado: 1 - (\frac{0}{4})^2 - (\frac{4}{4})^2 = 0$

$- Soleado: 1 - (\frac{2}{5})^2 - (\frac{3}{5})^2 = 0.48$

$Gini(JugarGolf, Temperatura)$

$- Caliente: 1 - (\frac{2}{4})^2 - (\frac{2}{4})^2 = 0.5$

$- Frío: 1 - (\frac{1}{4})^2 - (\frac{3}{4})^2 = 0.375$

$- Templado: 1 - (\frac{2}{6})^2 - (\frac{4}{6})^2 = 0.444$

$Gini(JugarGolf, Humedad)$

$- Normal: 1 - (\frac{4}{7})^2 - (\frac{3}{7})^2 = 0.489$

$- Alta: 1 - (\frac{1}{7})^2 - (\frac{6}{7})^2 = 0.244$

5. Plantee un conjunto de datos, con 15 registros y 5 atributos, cuyo atributo objetivo sea dicotómico y aplique las actividades que realizó en el ejercicio número 4.

JuegaTenis

Día	Panorama	Temperatura	Humedad	Viento	Juega Tenis
1	Soleado	Caliente	Alta	Ligero	No

2	Soleado	Caliente	Alta	Fuerte	No
3	Nublado	Caliente	Alta	Ligero	Sí
4	Lluvioso	Templado	Alta	Ligero	Sí
5	Lluvioso	Frío	Normal	Ligero	Sí
6	Lluvioso	Frío	Normal	Fuerte	No
7	Nublado	Frío	Normal	Fuerte	Sí
8	Soleado	Templado	Alta	Ligero	No
9	Soleado	Frío	Normal	Ligero	Sí
10	Lluvioso	Templado	Normal	Ligero	Sí
11	Soleado	Templado	Normal	Fuerte	Sí
12	Nublado	Templado	Alta	Fuerte	Sí
13	Nublado	Caliente	Normal	Ligero	Sí
14	Lluvioso	Templado	Alta	Fuerte	No
15	Soleado	Caliente	Alta	Ligero	No

Asumimos:
Clase P: JuegaTenis= "sí"
Clase N: JuegaTenis= "no"

Como hay 9 sí y 6 no en el atributo JuegaTenis, por lo tanto, 9 tuplas pertenecen a la clase P y 6 tuplas pertenecen a clase N.

$$Entropía(D) = -\frac{9}{15}log_2(\frac{9}{15}) - \frac{6}{15}log_2(\frac{6}{15}) = 0.9709$$

Paso 2: Dividir el conjunto de datos en diversos atributos.

Atributo objetivo: JuegaTenis	
Atributo	**Dominio**

Panorama	Lluvioso Nublado Soleado
Temperatura	Caliente Frío Templado
Humedad	Normal Alta
Viento	Ligero Fuerte

$$E(T,X) = \sum_{c \in x} p(c)E(c)$$

				Count
Panorama	Lluvioso	JuegaTenis	No	2
			Sí	3
	Nublado	JuegaTenis	Sí	4
	Soleado	JuegaTenis	No	4
			Sí	2

$E(Soleado) = E(No, Sí) = E(4,2) = (-\frac{4}{15}log_2(\frac{4}{15})) + (-\frac{2}{15}log_2(\frac{2}{15}))$

$= (-0.266log_2(0.14)) + (-0.133log_2(0.21))$

$= 0.5085 + 0.3875 = 0.8960 \rightarrow Entropía\ Soleado$

Para panorama:

$E(JuegaTenis, Panorama) = P(Lluvioso) * E(2,3) + P(Nublado) * E(4,0) + P(Soleado) * E(4,2)$

$-\ Lluvioso: P(5/15) = 0.33 \quad E(2,3) = 0.38758 + 0.4643 = 0.85197$

$-\ Nublado: P(4/15) = 0.26 \quad E(4,0) = 0$

$-\ Soleado: P(6/15) = 0.4 \quad E(4,2) = 0.8960$

$E(JuegaTenis,\ Panorama) = 0.33 * 0.85197 + 0.26 * 0 + 0.4 * 0.896 = 0.2811501 + 0 + 0.3584$

Paso 4: Ganancia de información

$Gain = 0.9709 - 0.6395501 = 0.3313499$

Para temperatura:

				Count
Temperatura	Caliente	JuegaTenis	No	3
			Sí	2

	Frío	JuegaTenis	No	1
			Si	3
	Templado	JuegaTenis	No	2
			Sí	4

$E(JuegaTenis, Temperatura) = P(Caliente) * E(3,2) + P(Frío) * E(1,3) + P(Templado) * E(2,4)$

$- Caliente: P(\frac{5}{15}) = 0.33 \quad E(3,2) = 0.46438 + 0.38758 = 0.85197$

$- Frío: P(\frac{4}{15}) = 0.266 \quad E(1,3) = 0.26045 + 0.46438 = 0.72484$

$- Templado: P(\frac{6}{15}) = 0.4 \quad E(2,4) = 0.38758 + 0.5085 = 0.89608$

$E(JuegaTenis, Temperatura) = 0.33 * 0.85197 + 0.266 * 0.72484 + 0.4 * 0.89608 = 0.2811501 + 0 = 0.8323821$

$Gain = 0.9709 - 0.8323821 = 0.138579$

Para humedad:

				Count
Humedad	Alta	JuegaTenis	No	5
			Sí	3
	Normal	JuegaTenis	No	1
			Si	6

$E(JuegaTenis, Humedad) = P(Normal) * E(1,6) + P(Alta) * E(5,3)$

$- Normal: P(\frac{8}{15}) = 0.533 \quad E(1,6) = 0.26045 + 0.52877 = 0.78923$

$- Alta: P(\frac{7}{15}) = 0.466 \quad E(4,3) = 0.5085 + 0.46438 = 0.9728$

$E(JuegaTenis, Humedad) = 0.533 * 0.78923 + 0.466 * 0.9728 = 0.42065 + 0.4533 = 0.87397$

$Gain = 0.9709 - 0.87397 = 0.0969252$

Gain Ratio

$Split(S, Día) = 15(-\frac{1}{15}log_2(\frac{1}{15})$=3.9068 Gain Ratio(S, Día)= $\frac{0.9709}{3.9068}$=0.2485

Split(S, Panorama) =$-\frac{5}{15}log_2(\frac{5}{15}) * 2 + (-\frac{4}{15}log_2(\frac{4}{15})) = 1.47929$ Gain Ratio(S,Panorama)= $\frac{0.3313499}{1.47929} = 0.22399$

Gain Ratio(S, Humedad)= $\frac{0.152}{1}$=0.152

Gain Ratio(S, Temperatura)= $\frac{0.138579}{1.362}$=0.1011746

Gain Ratio(S, Viento)= $\frac{0.048}{0.985} = 0.049$

Gini Index

$Gini(D) = 1 - (\frac{9}{15})^2 - (\frac{6}{15})^2 = 0.48$