Minería de datos. 2023-2. 3CV15

|  |  |
| --- | --- |
| Nombre: | Mauro Sampayo Hernández |

**GUÍA DE ESTUDIO**

Tan, Pang-Ning, Steinbach, Michael & Kumar, Vipin. (2014). Introduction to data mining. Pearson.

**Capítulo 4. Clasificación: conceptos básicos, árboles de decisión y evaluación de modelos**

1. Explique qué es el Proceso de Clasificación

*Es el proceso de aprendizaje de una* ***función objetivo f*** *(conocida informalmente como* ***modelo de clasificación****), que asigna cada conjunto de atributos* ***x*** *a una de las etiquetas de clases predefinidas* ***y****.*

1. Describa los siguientes modelos
   1. Modelo descriptivo

*Sirve como una herramienta explicativa para distinguir entre objetos de diferentes clases. Por ejemplo, puede ser útil para tener un modelo descriptivo que explique qué características definen a un “vertebrado” como mamífero, reptil, ave, pescado o anfibio.*

* 1. Modelo predictivo

*Se utiliza para predecir la etiqueta de clase de registros desconocidos.*

1. Describa el enfoque general para resolver problemas de clasificación (Sección 4.2)

*Primero, se debe proporcionar un* ***training set*** *(conjunto de entrenamiento)**cuyas etiquetas de clases sean conocidas. El* ***training set*** *se utiliza para construir un* ***modelo de clasificación****, el cual es posteriormente aplicado sobre un* ***test set*** *(conjunto de prueba), que consta de registros con etiquetas de clase desconocidas.*

Diagram

Description automatically generated

1. Describa la forma en que se resuelve un problema de clasificación (Sección 4.2.1)

*Primero, se debe proporcionar un* ***training set*** *(conjunto de entrenamiento)**cuyas etiquetas de clases sean conocidas. El* ***training set*** *se utiliza para construir un* ***modelo de clasificación****, el cual es posteriormente aplicado sobre un* ***test set*** *(conjunto de prueba), que consta de registros con etiquetas de clase desconocidas.*

*Posteriormente se debe realizar una evaluación del rendimiento del o de los* ***modelos de clasificación****. La evaluación del rendimiento de un* ***modelo de clasificación*** *se basa en contar cuantos registros de prueba fueron predichos de forma correcta e incorrecta por el modelo.*

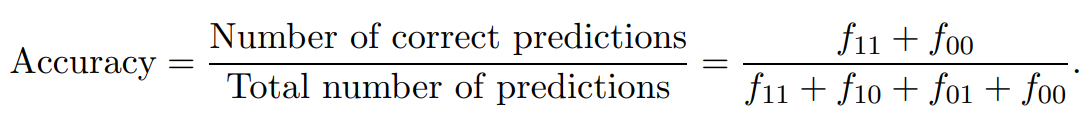
Table

Description automatically generated

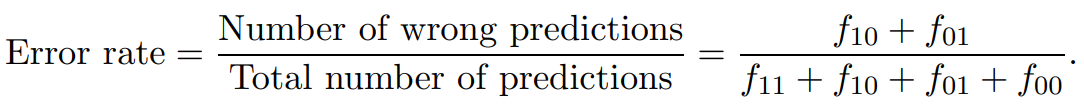
*Estas cuentas se tabulan en una tabla conocida como* ***matriz de confusión****, en donde cada entrada denota el número de registros de la clase predichos que serían de la clase . Por ejemplo, representa el número de registros de la clase 0 incorrectamente predichos como parte de la clase 1.*

*Con base en las entradas de la* ***matriz de confusión****, el número total de predicciones correctas realizadas por el modelo es y el número total de predicciones incorrectas es .*

Esta información se puede simplificar utilizando **métricas de rendimiento**, como lo son la **PRECISION:**



Y, la **TASA DE ERROR:**



*La mayoría de los* ***algoritmos de clasificación*** *buscan* ***modelos*** *que tengan la mayor* ***precisión****, o de manera equivalente, la* ***tasa de error*** *más baja.*

1. Identifique los criterios considerados en la construcción de la figura 4.4 (Sección 4.3.1)

Diagram

Description automatically generated

*El* ***árbol de decisión*** *de la* ***figura 4.4*** *tiene 3 tipos de nodos:*

* *Un* ***nodo raíz*** *que NO tiene ramas entrantes y cero o más ramas salientes.*
* ***Nodos internos****, cada uno de los cuales tiene exactamente UNA rama entrante y dos o más ramas salientes.*
* ***Nodos hoja*** *o* ***nodos terminales****, cada uno de los cuáles tiene exactamente UNA rama entrante y NINGUNA rama saliente.*

*Cada* ***nodo hoja*** *o* ***nodo terminal*** *tiene asignado una* ***etiqueta de clase.*** *Mientras que, cada uno de los* ***nodos NO terminales****, lo que incluye el* ***nodo raíz*** *y los* ***nodos internos****, contiene condiciones de prueba de atributos cuya función es el de separar registros que tienen diferentes características.*

*En el* ***árbol de decisión*** *de la* ***figura 4.4****, el* ***nodo raíz*** *hace uso del atributo “Body Temperature” para separar a los vertebrados de sangre caliente de los de sangre fría. Dado a que todos los vertebrados de sangre fría NO son mamíferos, se crear un* ***nodo hoja*** *etiquetado como “Non-mammals” como nodo hijo derecho del* ***nodo raíz****. Por otro lado, si el vertebrado es de sangre caliente, un atributo subsecuente, “Gives Birth”, es utilizado para distinguir a los mamíferos de otras criaturas de sangre caliente, que en su mayoría son aves.*

1. Describa los dos pasos del Algoritmo de Hunt para la construcción de árboles de decisión (Sección 4.3.2)

*Sea un* ***conjunto de registros de entrenamiento (training set)*** *que están asociados con un nodo e las* ***etiquetas de clase****,*

* ***Paso 1:*** *Si todos los registros en pertenecen a la misma clase , entonces es un* ***nodo hoja*** *etiquetado como .*
* ***Paso 2:*** *Si contiene registros que pertenecen a más de una* ***clase****, se selecciona una* ***condición de prueba de atributo*** *para dividir los registros en subconjuntos más pequeños. Se crea un nodo hijo para cada uno de los resultados de la* ***condición de prueba*** *y los registros en son distribuidos a los nodos hijo en función de los resultados. El algoritmo es entonces aplicado recursivamente sobre cada uno de los nodos hijo.*

1. Explique los dos casos que requieren condiciones adicionales en la construcción de un árbol de decisión considerando la combinación de valores y etiquetas asociadas en los elementos (Página 154).

*Considerando el problema de predecir si un solicitante de préstamo pagará sus obligaciones o si, de lo contrario, incumplirá con sus pagos, se tiene el siguiente* ***training set****.*

Table

Description automatically generated

*En él se enlistan registros que contienen la información personal de prestatarios junto a una* ***etiqueta de clase*** *que indica si el prestatario ha cumplido con sus pagos de préstamos o no.*

*El* ***árbol inicial*** *para el problema de clasificación contiene un único nodo con la* ***etiqueta de clase*** *“Defaulted = No”, lo que quiere decir que la mayoría de los prestatarios pagaron exitosamente sus préstamos.*

A picture containing text

Description automatically generated

*Sin embargo, este árbol necesita ser refinado, ya que el* ***nodo raíz*** *contiene registros de ambas clases. Para esto, se aplica una* ***condición de prueba*** *sobre el atributo “Home Owner”, y los registros se dividen en subconjuntos más pequeños basados en los resultados de dicha* ***prueba.*** *De esta manera, se obtiene el siguiente árbol:*

Diagram

Description automatically generated

*Ahora, se aplica de forma recursiva el algoritmo de Hunt sobre cada uno de los nodos hijos del* ***nodo raíz****.*

*Con base en el* ***training set*** *dado para este problema, se puede notar que, todos los prestatarios que son PROPIETARIOS DE UNA CASA pagaron exitosamente sus préstamos. El hijo izquierdo de la* ***raíz*** *es, por lo tanto, un* ***nodo hoja*** *etiquetado como “Defaulted = No”.*

*Sin embargo, el hijo derecho aun contiene registros de ambas clases, por lo que se debe continuar aplicando el algoritmo de Hunt sobre este nodo. De esta forma, se aplica otra* ***condición de prueba*** *sobre este nodo, ahora usando el atributo “Marital Status”, para así dividir los registros de este* ***nodo*** *en subconjuntos más pequeños. De esta forma, se obtiene el siguiente árbol:*

Diagram

Description automatically generated

*Nuevamente analizamos el árbol con base al* ***training set****. Podemos notar así, que todos los prestatarios que están CASADOS incumplieron con el pago de sus préstamos. Podemos concluir entonces, que el hijo derecho de este* ***nodo*** *es un* ***nodo hoja*** *etiquetado como “Defaulted = No”*

*Sin embargo, el hijo izquierdo aun contiene registros de ambas clases, por lo que se debe continuar aplicando el algoritmo de Hunt sobre este nodo. Aplicamos de esta manera, una nueva* ***condición de prueba*** *sobre el nodo, usando ahora el atributo “Annual Income” para dividir los registros en subconjuntos más pequeños. Así, se tiene:*

Diagram

Description automatically generated

*Analizando este nuevo árbol con base al* ***training set****, podemos observar que todos los prestatarios con un ingreso anual MAYOR a los 80,000 dólares cumplieron con el pago de sus préstamos. Por el otro lado, podemos observar también, que todos aquellos prestatarios con un ingreso MENOR a la cifra mencionada anteriormente no cumplieron con el pago de sus préstamos. De esta forma, se concluye que ambos hijos son* ***nodos hoja*** *etiquetados como “Defaulted = No” y “Defaulted = Yes” como se muestra en la imagen.*

*Al ya no contar con ningún* ***nodo*** *que contenga registros de mas de una clase, se puede dar por finalizado el proceso de construcción de este* ***árbol de decisión.***

1. Responda las siguientes preguntas:
   1. ¿Cómo se deben dividir los registros del conjunto de entrenamiento?

*Por cada paso recursivo del* ***proceso de crecimiento de árboles****, se debe seleccionar una* ***prueba de atributo*** *para dividir los registros en subconjuntos más pequeños. Para ello, el algoritmo debe proporcionar un método para especificar* ***condiciones de prueba*** *para diferentes tipos de atributos, así como también una medida objetiva para evaluar cada una de estas.*

* 1. ¿Cómo debe terminar el procedimiento de división?

*Una posible estrategia es continuar expandiendo un nodo hasta que todos los registros pertenezcan a una misma clase, o todos los registros posean* ***valores de atributo*** *idénticos.*

1. Explique los métodos para expresar la condición de particionamiento aplicable a cada uno de los siguientes tipos de atributo:
   1. Binario

*Las* ***condiciones de prueba*** *para atributos BINARIOS generan únicamente dos resultados.*

Diagram

Description automatically generated

* 1. Nominales

*Dado que un atributo NOMINAL puede tener muchos valores, su* ***condición de prueba*** *puede ser expresada de dos formas:*

* ***División multidireccional (Multiway Split):*** *Para una DIVISIÓN MULTIDIRECCIONAL, el numero de resultados depende del número de valores distintos que pueda tener el atributo correspondiente. Por ejemplo, si un atributo como el “Estado Civil” tiene tres valores distintos (Soltero, Casado o Divorciado), su* ***condición de prueba*** *producirá una división en tres partes.*

Diagram

Description automatically generated

* ***División Binaria (Binary Split):*** *Algunos* ***algoritmos de******árboles de decisión****, como CART, producen solamente DIVISIONES BINARIAS considerando todas las formas de crear una partición binaria de* ***valores de atributo****. Retomando el ejemplo anterior, se muestran las 3 diferentes formas en las que se pueden agrupar los valores del atributo “Estado Civil” en dos subconjuntos.*

Diagram

Description automatically generated

* 1. Ordinales

*Los atributos ORDINALES pueden producir divisiones BINARIAS o MULTIDIRECCIONALES; y también pueden ser agrupados siempre y cuando la agrupación no viole la propiedad del orden de los* ***valores de los atributos.***

*A continuación, se muestran dos formas de agrupar registros en función de un atributo que representa la talla de una camiseta. Nótese que en ambos casos se preserva el orden existente entre los* ***valores de los atributos****.*

Diagram

Description automatically generated

* 1. Continuos

*Para los atributos CONTINUOS, la* ***condición de prueba*** *puede expresarse como una prueba de comparación o con resultados binarios, o por medio de una consulta de rangos con resultados de la forma para .*

*Para el primer caso, el* ***algoritmo de árbol de decisión*** *debe considerar todas las posibles divisiones y seleccionar la que produce la mejor partición.*

Diagram

Description automatically generated

*Para el segundo caso, el algoritmo debe considerar todos los rangos de valores continuos posibles. Los intervalos adyacentes también pueden ser incluidos en rangos más amplios siempre que se conserve la propiedad del orden de los* ***valores de los atributos****.*

Diagram, schematic

Description automatically generated

1. Explique el criterio que se utilice en las medidas para seleccionar la mejor forma de dividir los registros (Sección 4.3.4)

*Las medidas para seleccionar la mejor forma de dividir los registros, a menudo se basan en el* ***CRITERIO DE IMPUREZA*** *de los* ***nodos hijo****.*

1. Justifique por qué los autores mencionan que el mejor atributo de división es el Tipo de Automóvil (Car Type) en el árbol de la figura 4.12 (Sección 4.3.4)

*Debido a que, al realizar la división por dicho atributo, se obtienen particiones con un grado MENOR de* ***impurezas*** *que si se usaran otros atributos. Esto significa, que en este atributo se halla la mejor posición para la división de los registros.*

1. Explique en qué consiste el criterio de impureza de nodos tomado en cuenta para la elección de atributos de particionamiento (Sección 4.3.4).

*Suponiendo que denota la fracción de los registros que pertenecen a la clase en un nodo dado (Esto también puede ser expresado omitiendo la referencia al nodo de la forma ).*

*En un problema de dos clases, la* ***distribución de clases*** *en cualquier nodo se puede escribir como , donde . Para ilustrar lo anterior, considere las siguientes* ***condiciones de prueba****.*

Diagram

Description automatically generated

*La* ***distribución de las clases*** *antes de realizar la división es (0.5, 0.5), debido a que hay un número de registros igual en cada clase. Si dividimos los datos utilizando el* ***atributo*** *“Gender”, entonces la* ***distribución de las clases*** *de los* ***nodos hijo*** *serán (0.6, 0.4) y (0.4, 0.6) respectivamente. A pesar de que las clases ya no están uniformemente distribuidas, los* ***nodos hijo*** *aún contienen registros de ambas clases.*

*Realizando la división utilizando el segundo* ***atributo*** *“Car Type”, podremos obtener particiones más puras.*

Diagram

Description automatically generated

*A menudo, las medidas desarrolladas para seleccionar la mejor forma de dividir los registros se basan en el* ***grado de impureza*** *de los* ***nodos hijos****. ENTRE MENOS SEA EL* ***GRADO DE IMPUREZA****, MÁS SESGADA SERA LA* ***DIVISIÓN DE LAS CLASES****. Por ejemplo, un nodo con una* ***distribución de clases*** *(0, 1) tienen CERO impurezas, mientras que un nodo con una* ***distribución de clases uniforme*** *(0.5, 0.5) tiene la MAYOR IMPUREZA posible.*

*Algunos ejemplos de* ***medidas de impureza*** *son:*

Text

Description automatically generated

*Donde c es el* ***número de clases*** *y en el cálculo de la* ***Entropía.***

*En la siguiente gráfica se comparan los valores de las* ***medidas de impurezas*** *para problemas de clasificación binaria. En el eje horizontal, p se refiere a la fracción de registros que pertenecen a una de las dos clases.*

Chart

Description automatically generated with medium confidence

*Podemos observar que las tres medidas alcanzan su valor máximo cuando la* ***distribución de las clases*** *es uniforme (es decir, cuando p = 0.5). Mientras que los valores mínimos para las medidas se obtienen cuando todos los registros pertenecen a la misma clase (es decir, p = 0 o 1). A continuación, se muestran algunos ejemplos del cálculo de las tres* ***medidas de impureza.***

Text, letter

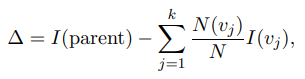
Description automatically generated

*Basándonos en los cálculos enlistados anteriormente y la gráfica de comparación de valores de las* ***medidas de impurezas****, podemos concluir que el nodo tiene el MENOR* ***grado de impureza,*** *seguido por los nodos y .*

1. Explique el significado de Δ (Sección 4.3.4).

*La* ***ganancia* Δ**, es un criterio que se utiliza para determinar que tan bien se desempeña una **condición de prueba** (bondad de la división de registros).

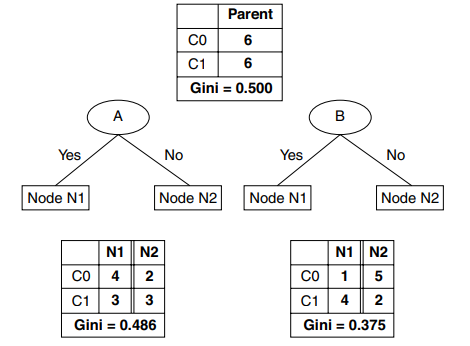
Para ello, se debe comparar el **grado de impureza** del **nodo padre** (antes de realizar la división de los registros) con el **grado de impureza** de los **nodos hijo** (después de haberse realizado la división de los registros). CUANTO MAYOR SEA SU DIFERENCIA, MEJOR SERA LA **CONDICIÓN DE PRUEBA**.



*donde* *es la* ***medida de la impureza*** *de un nodo dado,* *es el número total de registros del* ***nodo padre****,* *es el número de* ***valores del atributo*** *y* *es el número de registros asociados con el* ***nodo hijo*** *.*

*Los* ***algoritmos de inducción de árboles de decisión*** *a menudo suelen elegir una* ***condición de prueba*** *que maximice la* ***ganancia* Δ**. Dado que tiene el mismo valor para todas las condiciones de prueba, maximizar la **ganancia** es equivalente a minimizar la ponderación del promedio de las **medidas de impureza** de los **nodos hijo**.

1. Considerando la figura 4.14, explique la conveniencia de la elección del atributo B para realizar el particionamiento (Sección 4.3.4).



*Antes de realizar la división, el índice Gini de los datos es de 0.5, ya que existe un número de registros igual en ambas clases.*

*Si se elige el atributo A para realizar la división de los datos, el índice Gini para el nodo* ***N1*** *sería de* ***0.4898****, y para el nodo* ***N2*** *sería de* ***0.480****. De esta forma, el promedio ponderado del índice Gini de ambos nodos sería de 0.486.*

*De manera similar, si se elige el atributo B, podemos calcular que el promedio ponderado del índice Gini de sus nodos sería 0.375.*

*De esta forma, y dado a que los subconjuntos obtenidos a partir del atributo B tienen un índice Gini más pequeño que los obtenidos a partir del atributo A, es que* ***el atributo B se prefiere sobre el A.***

1. Analice el efecto que tienen las divisiones binarias con respecto a la división múltiple de la figura 4.15 (Sección 4.3.4).

A picture containing diagram

Description automatically generated

*Para la primera agrupación binaria del atributo “Car Type”, el índice Gini de* ***{Sports, Luxury}*** *es de 0.4922, y el índice Gini de* ***{Family}*** *es de 0.3750. Así, el promedio ponderado del índice Gini para esta agrupación es igual a 0.468.*

*La segunda agrupación por otro lado, cuenta con subconjuntos mucho mas puros que la primera agrupación, por lo que su índice Gini es MENOR.*

*Por último, para la* ***división múltiple****, el índice Gini debe se calcula para cada* ***valor de los atributos.*** *Como* ***Gini({Family}) = 0.375****,* ***Gini({Sports}) = 0*** *y* ***Gini({Luxury}) = 0.219****, el índice Gini general para esta* ***división múltiple*** *es igual a 0.163.*

*Como se puede observar, la* ***división múltiple*** *tiene un índice de Gini más pequeño a comparación de las dos agrupaciones binarias. Esto sucede debido a una agrupación binaria fusiona algunos de los resultados de una* ***división múltiple****, lo que genera subconjuntos menos puros.*

1. Explique cómo se realiza la división de atributos continuos (Sección 4.3.4).

A screenshot of a computer

Description automatically generated with medium confidence

*Considerando la siguiente figura, en donde la* ***condición de prueba*** *se utiliza para dividir los registros de un* ***training set*** *para el problema de clasificación de incumplimiento de pagos de préstamos.*

*Se puede aplicar un método de* ***fuerza bruta*** *para encontrar considerando cada valor que el atributo pueda tomar en los registros* ***N*** *como una* ***posición de división candidata.*** *Para cada candidato , se escanea el conjunto de datos una sola vez para así contar el numero de registros con un ingreso anual menor o mayor que . Posteriormente calculamos el índice Gini para cada candidato y se elije aquel que tenga el valor más bajo. Sin embargo, la complejidad computacional de este enfoque resulta ser muy costosa, ya que requiere de un tiempo .*

*Para reducir esta complejidad, se pueden clasificar los registros del* ***training set*** *en función a los ingresos anuales, un cálculo que requiere un tiempo de* ***.*** *De esta manera las* ***posiciones de división candidatas*** *son identificadas tomando los puntos medios existentes entre dos valores ordenados adyacentes (55, 65, 72,…) para después evaluar el índice Gini de cada uno de ellos. Con esta nueva aproximación, ya no se tienen que examinar todos los registros* ***N****, al ser que solo se tendrían que evaluar las* ***posiciones de división candidatas*** *que sean definidas.*

*Siguiendo el procedimiento anterior, para el primer candidato, , ninguno de los registros tiene ingresos anuales menores a los $55K, lo que significa que el índice de Gini para el nodo descendiente es igual a cero. Por otra parte, el numero de registros que tienen un ingreso anual mayor o igual a los $55k es de 3 para la clase* ***Si****, y 7 para la clase* ***No****; lo que significa que el índice Gini para este nodo es de 0.420. De esta forma, el índice Gini general para esta* ***posición de división candidata*** *es igual a 0.420.*

*Para el segundo candidato, , podemos determinar su* ***distribución de clases*** *actualizando la distribución del candidato anterior. Esto se hace examinando la* ***etiqueta de clase*** *del registro con el ingreso anual más bajo (para este caso* ***$60k****). Dado a que la**etiqueta de la clase para este registro es* ***No****, el conteo para la clase* ***No*** *se incrementa de 1 a 0 para y se reduce de 7 a 6 para . El nuevo índice promedio ponderado Gini para para esta* ***posición de división candidata*** *es igual a 0.400.*

*El mismo procedimiento se repite hasta que todos los valores del índice Gini de todos los candidatos hayan sido calculados. Finalmente se elige la posición que haya producido el índice Gini más pequeño (para el caso del ejemplo sería el candidato ).*

*El procedimiento descrito anteriormente puede optimizarse aún más si se consideran únicamente aquellas* ***posiciones de división candidatas*** *que estén ubicadas entre dos registros adyacentes con DIFERENTES* ***etiquetas de clase****.*

*Para el caso de la tabla, los primeros registros ordenados (aquellos con ingresos anuales de $60k, $70k y $75k) tienen* ***etiquetas de clase*** *IDÉNTICAS, por lo cual no puede existir una* ***posición de división candidata*** *entre alguno de estos registros. Sin embargo, entre el tercer y cuarto registro ($75k y $85k respectivamente) sí que puede existir una* ***posición de división candidata*** *al ser que ambos tienen* ***etiquetas de clase*** *DIFERENTES. Este enfoque nos permite así, reducir el número de* ***posiciones de división candidata*** *de 11 a 2.*

1. Explique la forma en que se podrían tratar atributos de clave principal aplicando criterios de división (Gain Ratio Sección 4.3.4).

*Los atributos de una clave principal no pueden ser considerados como* ***atributos predictivos*** *ya que sus valores son únicos para cada registro.*

***Una condición de prueba que resulte en una gran cantidad de resultados no es deseable, debido a que la cantidad de registros asociados con cada partición es muy pequeño como para realizar predicciones fiables****.*

*Tomando en cuenta este problema, existen dos estrategias con las que se podría abordar este problema, y así tratar los atributos de una clave principal:*

* *Restringir las* ***condiciones de prueba*** *a únicamente DIVISIONES BINARIAS*
* ***Modificar el criterio de división*** *de tal manera que tenga en cuenta el numero de resultados producidos por la* ***condición de prueba del atributo****. El criterio de uso más común es el* ***Gain Ratio*** *(Relación de Ganancia), y se utiliza para determinar la bondad de una división.*

A picture containing text

Description automatically generated

*Donde,* *y es el número total de divisiones. Sin embargo, si cada* ***valor del atributo*** *tiene el mismo número de registros, entonces* ***.*** *Esto sugiere, que, si un atributo produce una gran cantidad de divisiones, la división de su información también será grande, lo que a su vez reduce su relación de ganancia (****Gain Ratio****)****.***

1. Explique por lo menos seis de las once características de la inducción del árbol de decisión (Sección 4.3.7)

* Pag 168

1. Explique lo siguiente (Sección 4.4):
   1. En qué consisten los errores de entrenamiento y los errores de generalización;
   2. Las características que tiene un buen modelo
   3. ¿Qué significa el sobreajuste?
   4. ¿Qué significa el subajuste?
2. Explique el comportamiento de las tasas de error de entrenamiento y prueba expresados en la figura 4.23 (Sección 4.4)
   1. Cuando el árbol es pequeño (pocos nodos
   2. Conforme aumenta la cantidad de nodos en el árbol
3. Describa de forma breve los siguientes casos de sobreajuste del modelo:
   1. Sobreajuste por presencia de ruido
   2. Sobreajuste debido a la falta de muestras representativas
   3. Sobreajuste debido a la falta de muestras representativas
4. Describa los siguientes procesos de evaluación del desempeño de un clasificador:
   1. Método de retención (holdout) (Sección 4.5.1)
   2. Submuestreo aleatorio (Sección 4.5.2)
   3. Validación cruzada (Sección 4.5.3)
   4. Bootstrap (Sección 4.5.4)