SISTEMAS INTELIGENTES

Lecture 06:Regresión y Clasificación con Arboles de Decisión

Dr. Edwin Valencia Castillo
Departamento de Sistemas
Facultad de Ingeniería
Universidad Nacional de Cajamarca
2024

Introducción

- ✓ **Utilizamos mentalmente estructuras de árbol de decisión constantemente en nuestra vida diaria sin darnos cuenta:**
  - ➢ ¿Llueve? => lleva paraguas. ¿Soleado? => lleva gafas de sol. ¿estoy cansado? => toma café.
  - ➢ (decisiones del tipo IF THIS THEN THAT)
- ✓ **Objetivos de la clase:**
  - ➢ Los estudiantes comprenderán qué es un árbol de decisión y cómo se utiliza en el aprendizaje supervisado.
  - ➢ Aprenderán a construir y evaluar árboles de decisión.

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

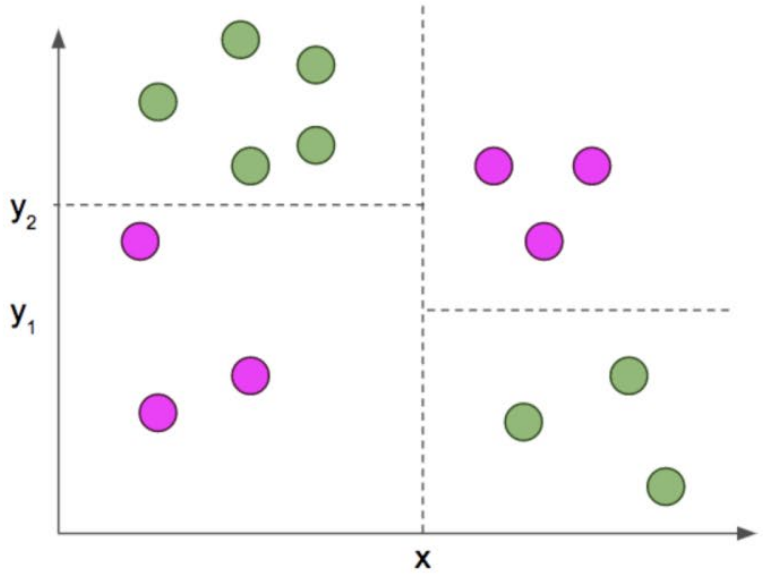
Árboles de decisión

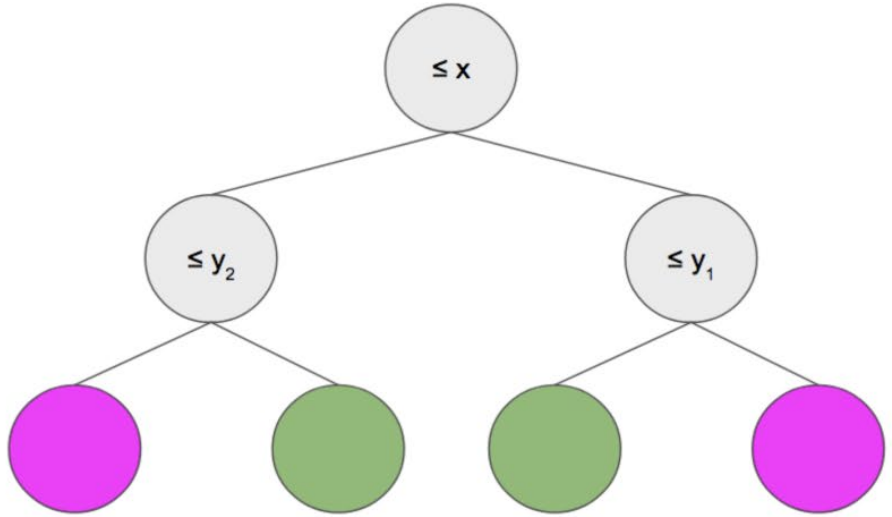
- ✓Los árboles de decisión son una técnica popular y poderosa en el campo del aprendizaje supervisado, utilizada tanto para tareas de clasificación como de regresión (CART), fue desarrollado por Breiman et al. (1984).
- ✓Estos modelos dividen o segmentan el espacio de las variables predictoras en una serie de regiones. Una vez creado el árbol de decisión es utilizado para predecir observaciones futuras.
- ✓Para este propósito se utiliza la moda en el caso de que la variable a predecir sea categórica o bien la media en el caso de que sea numérica

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Árboles de decisión

- ✓Como el conjunto de las reglas para separar las variables predictoras se pueden resumir en forma de árbol, a estos métodos se les conoce popularmente con el nombre de árboles de decisión.
- ✓Su estructura se asemeja a un árbol, donde cada nodo interno representa una "prueba" o "condición" sobre un atributo, cada rama representa el resultado de la prueba, y cada nodo hoja representa una clase o un valor de destino.





© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Árboles de decisión

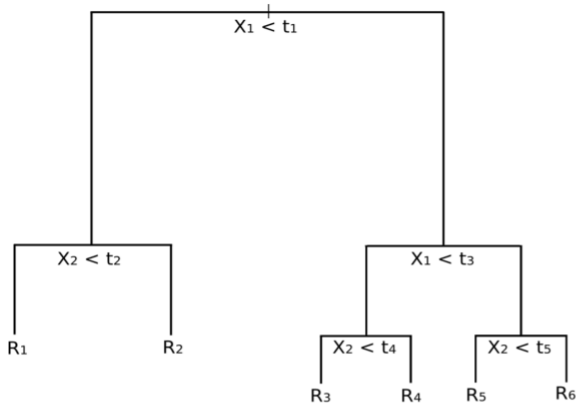
- ✓Los árboles de decisión tienen un primer nodo llamado raíz (root) y luego se descomponen el resto de atributos de entrada en dos ramas (podrían ser más, pero no nos meteremos en eso ahora) planteando una condición que puede ser cierta o falsa. Se bifurca cada nodo en 2 y vuelven a subdividirse hasta llegar a las hojas que son los nodos finales y que equivalen a respuestas a la solución: Si/No, Comprar/Vender, o lo que sea que estemos clasificando.

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

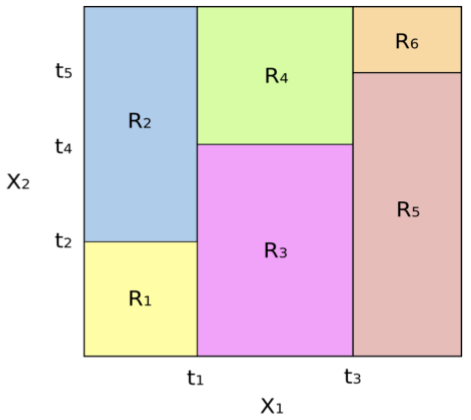
Componentes de un Árbol de Decisión

- ✓**Nodo Raíz: El nodo superior del árbol que representa el primer punto de decisión basado en la característica más importante.**
- ✓**Nodos Internos: Nodos que representan decisiones basadas en una característica particular. Cada nodo divide el conjunto de datos en subconjuntos más pequeños.**
- ✓**Nodos Hoja: Nodos terminales que representan la clasificación o el valor de destino final después de realizar todas las decisiones necesarias.**

Los **árboles de decisión** dividen o segmentan el espacio de las variables predictoras en una serie de regiones. En el caso de los árboles utilizados para modelos de regresión se utiliza la **media** para estimar los valores que se encuentran en una determinada región. En el caso de los modelos de clasificación se utiliza la **moda** de la clase.



A Decision Tree with six separate regions



The resulting partition of the subset of $\mathbb{R}^2$ into six regional "blocks"

$$\text{RSS} = \sum_{m=1}^{M} \sum_{i \in R_m} (y_i - \hat{y_{R_m}})^2$$

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Construcción de un Árbol de Decisión

Selección del mejor atributo basado en un criterio (ganancia de información, índice Gini, etc.).

División recursiva del conjunto de datos en subconjuntos más pequeños.

Condiciones de parada, como profundidad máxima del árbol o número mínimo de muestras en un nodo hoja, para evitar sobreajuste.

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Algoritmos de Construcción

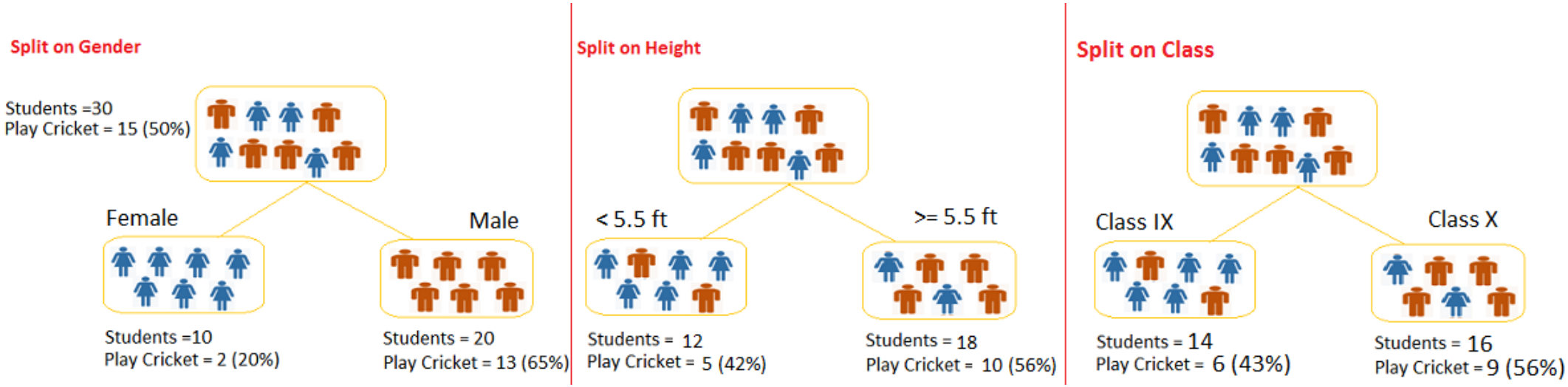
- ✓ID3 (Iterative Dichotomiser 3): Utiliza la entropía y la ganancia de información para construir el árbol. Selecciona el atributo que reduce más la incertidumbre.
- ✓C4.5: Una extensión de ID3 que maneja datos continuos y atributos faltantes.
- ✓C.5.0 desarrollado por Ross Quinlan (1990) es una evolución del C4.5
- ✓CART (Classification and Regression Trees): Utiliza el índice Gini para clasificación y la suma de los errores cuadrados para regresión.

De todos ellos el algoritmo *C.5.0* está considerado como el estándar en la industria.

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Veamos el siguiente ejemplo:

- ✓ 30 estudiantes
- ✓ 3 variables: Género (hombre/mujer), Clase (IX/X) y Altura (5 a 6 pies).
- ✓ 15 estudiantes juegan cricket en su tiempo libre
- ✓ Crear un modelo para predecir quien jugará cricket
- ✓ Segregar estudiantes basados en todos los valores de las 3 variables e identificar aquella variable que crea los conjuntos más homogéneos de estudiantes y que a su vez son heterogéneos entre ellos.



¿Qué variable es la más significativa para la mejor división de la población?

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

¿Cómo decide un árbol donde ramificarse?

- ✓La decisión de hacer divisiones estratégicas afecta altamente la precisión del árbol.
- ✓Los criterios de decisión son diferentes para árboles de clasificación y regresión.
- ✓Existen varios algoritmos para decidir la ramificación.
- ✓La creación de subnodos incrementa la homogeneidad de los subnodos resultantes. Es decir, la pureza del nodo se incrementa respecto a la variable objetivo.
- ✓Se prueba la división con todas las variables y se escoge la que produce subnodos más homogéneos.

Algunos algoritmos más comunes para la selección: Índice Gini, Chi Cuadrado, Ganancia de la información y entropía, y Reducción en la varianza (regresión)

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Índice Gini

- ✓ **El índice Gini es una métrica utilizada para evaluar la impureza o heterogeneidad de un conjunto de datos en el contexto de los árboles de decisión.**
- ✓ **Específicamente, mide la probabilidad de que un elemento seleccionado al azar del conjunto sea clasificado incorrectamente si se asigna a la clase de acuerdo con la distribución de la clase en el conjunto.**
- ✓ **Fórmula del Índice Gini:**
  - ➢ **Para un conjunto de datos con *c* clases, el índice Gini se define como:**

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

donde:

p_i es la proporción de elementos que pertenecen a la clase i en el conjunto de datos S.

- ✓ **Interpretación del Índice Gini:**
  - ➢ **Valor de 0: Indica pureza total, es decir, todos los elementos pertenecen a una sola clase.**
  - ➢ **Valor cercano a 1: Indica alta impureza o heterogeneidad, es decir, los elementos están distribuidos de manera uniforme entre todas las clases.**
- ✓ **Uso del Índice Gini en Árboles de Decisión**
  - ➢ **En el contexto de los árboles de decisión, el índice Gini se utiliza para decidir el mejor punto de división en un nodo. La reducción del índice Gini al dividir un nodo se calcula como la diferencia entre la impureza del nodo padre y la suma ponderada de las impurezas de los nodos hijos. El objetivo es maximizar esta reducción, lo que implica una división que crea nodos hijos más puros.**

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Índice Gini ponderado y ganancia de Gini

- ✓El índice Gini ponderado se utiliza para evaluar la calidad de una división en un árbol de decisión.
- ✓Se calcula al dividir el índice Gini total en subconjuntos y promediando los índices Gini de esos subconjuntos, ponderados por el tamaño relativo de cada subconjunto.
- ✓Fórmula del Índice Gini Ponderado
  - ➢Supongamos que tenemos un conjunto de datos S que se divide en dos subconjuntos S_1 y S_2 . El índice Gini ponderado para esta división se calcula como:

$$Gini_{ponderado} = \frac{|S_1|}{|S|} \times Gini(S_1) + \frac{|S_2|}{|S|} \times Gini(S_2)$$

donde:

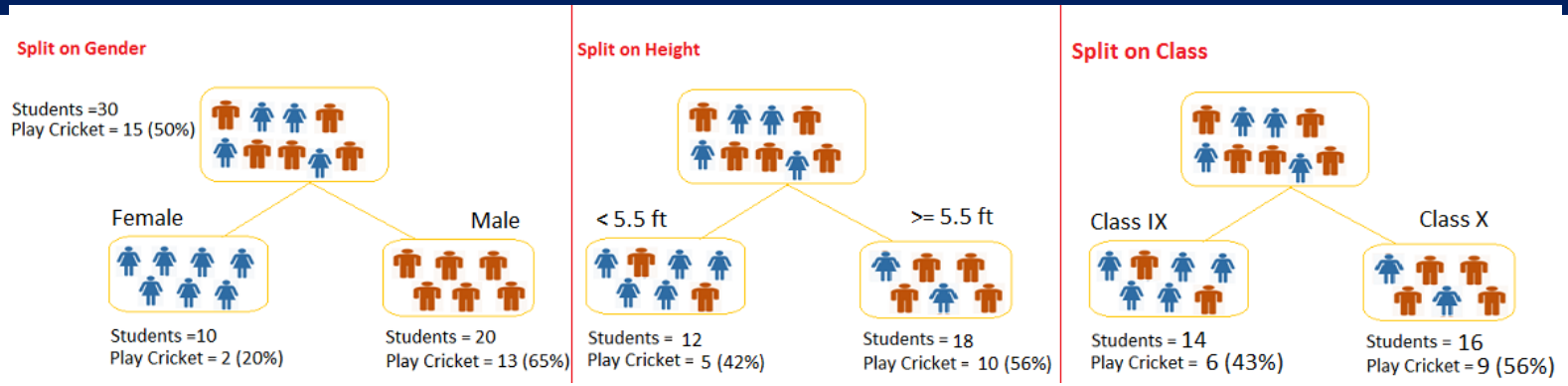
- $|S|$ es el tamaño del conjunto total de datos.
- $|S_1|$ y $|S_2|$ son los tamaños de los subconjuntos resultantes de la división.
- $Gini(S_1)$ y $Gini(S_2)$ son los índices Gini de los subconjuntos S_1 y S_2, respectivamente.

- ✓Ganancia de Gini
  - ➢La ganancia de Gini se calcula como la diferencia entre el índice Gini del conjunto original y el índice Gini ponderado de los subconjuntos:

$$\text{Ganancia de Gini} = Gini(S) - Gini_{ponderado}$$

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Índice Gini en nuestro ejemplo



- ✓ Para cada variable, calcularemos la impureza Gini antes y después de la división, y luego calcularemos la reducción de la impureza Gini para determinar la ganancia de información basada en el índice Gini.
- ✓ Paso 1: Índice Gini total
  - ➢ Para el conjunto total de 30 estudiantes, de los cuales 15 juegan cricket y 15 no juegan cricket:

$$Gini(S) = 1 - \left(\left(\frac{15}{30}\right)^2 + \left(\frac{15}{30}\right)^2\right) = 1 - (0.5^2 + 0.5^2) = 1 - (0.25 + 0.25) = 1 - 0.5 = 0.5$$

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Índice Gini en nuestro ejemplo

✓Paso 2: Índice Gini por variable

➢Variable Género

- **Mujeres S_1:** 10 (2 juegan cricket, 8 no juegan)
- **Hombres S_2:** 20 (13 juegan cricket, 7 no juegan)

$$Gini(\text{Mujeres}) = 1 - \left(\left(\frac{2}{10}\right)^2 + \left(\frac{8}{10}\right)^2\right) = 1 - (0.04 + 0.64) = 1 - 0.68 = 0.32$$

$$Gini(\text{Hombres}) = 1 - \left(\left(\frac{13}{20}\right)^2 + \left(\frac{7}{20}\right)^2\right) = 1 - (0.4225 + 0.1225) = 1 - 0.545 = 0.455$$

Índice Gini ponderado:

$$Gini(\text{Género}) = \frac{10}{30} \times 0.32 + \frac{20}{30} \times 0.455 = 0.1067 + 0.3033 = 0.41$$

Ganancia de Gini:

$$\text{Ganancia Gini}(\text{Género}) = 0.5 - 0.41 = 0.09$$

➢Variable Clase

- Clase IX: 14 (6 juegan cricket, 8 no juegan)
- Clase X: 16 (9 juegan cricket, 7 no juegan)

$$Gini(\text{IX}) = 1 - \left(\left(\frac{6}{14}\right)^2 + \left(\frac{8}{14}\right)^2\right) = 1 - (0.1837 + 0.3265) = 1 - 0.5102 = 0.4898$$

$$Gini(\text{X}) = 1 - \left(\left(\frac{9}{16}\right)^2 + \left(\frac{7}{16}\right)^2\right) = 1 - (0.3164 + 0.1914) = 1 - 0.5078 = 0.4922$$

Índice Gini ponderado:

$$Gini(\text{Clase}) = \frac{14}{30} \times 0.4898 + \frac{16}{30} \times 0.4922 = 0.2285 + 0.2624 = 0.4909$$

Ganancia de Gini:

$$\text{Ganancia Gini}(\text{Clase}) = 0.5 - 0.4909 = 0.0091$$

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Índice Gini en nuestro ejemplo

✓Paso 2: Índice Gini por variable

➢Variable Altura

- **< 5.5 pies: 12 (5 juegan cricket, 7 no juegan)**
- **>= 5.5 pies: 18 (10 juegan cricket, 8 no juegan)**

$$Gini(< 5.5 \text{ pies}) = 1 - \left(\left(\frac{5}{12}\right)^2 + \left(\frac{7}{12}\right)^2\right) = 1 - (0.1736 + 0.3403) = 1 - 0.5139 = 0.4861$$

$$Gini(>= 5.5 \text{ pies}) = 1 - \left(\left(\frac{10}{18}\right)^2 + \left(\frac{8}{18}\right)^2\right) = 1 - (0.3086 + 0.1975) = 1 - 0.5061 = 0.4939$$

Índice Gini ponderado:

$$Gini(\text{Altura}) = \frac{12}{30} \times 0.4861 + \frac{18}{30} \times 0.4939 = 0.1944 + 0.2963 = 0.4907$$

Ganancia de Gini:

$$\text{Ganancia Gini}(\text{Altura}) = 0.5 - 0.4907 = 0.0093$$

✓Conclusión

➢Comparando las ganancias de Gini:

- Ganancia Gini(Género) = 0.09
- Ganancia Gini(Clase) = 0.0091
- Ganancia Gini(Altura) = 0.0093

La variable Género tiene la mayor ganancia de Gini (0.09), por lo tanto, es la variable más significativa para la mejor división de la población en este conjunto de datos.

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Índice Chi Cuadrado

- ✓ El índice Chi Cuadrado es una medida estadística utilizada para evaluar la independencia entre dos variables categóricas.
- ✓ Se basa en la comparación de las frecuencias observadas en una tabla de contingencia con las frecuencias esperadas bajo la hipótesis nula de independencia.
- ✓ Fórmula del Índice Chi Cuadrado
  - ➢ El índice Chi Cuadrado se calcula mediante la siguiente fórmula:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

donde:

- O_i son las frecuencias observadas.
- E_i son las frecuencias esperadas bajo la hipótesis nula.
- La suma se realiza sobre todas las celdas de la tabla de contingencia.

- ✓ Cálculo de Frecuencias Esperadas E_i
  - ➢ Las frecuencias esperadas E_i se calculan utilizando:

$$E_i = \frac{(F_{row} \times F_{col})}{N}$$

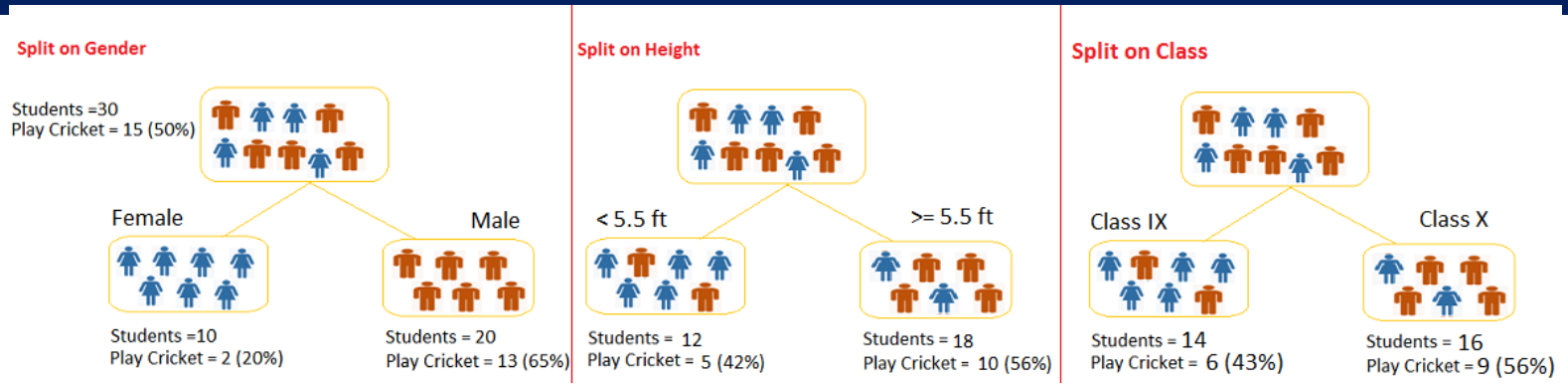
donde:

- F_{row} es el total de la fila.
- F_{col} es el total de la columna.
- N es el tamaño total de la muestra.

- ✓ Interpretación del Índice Chi Cuadrado
  - ➢ Un valor alto de χ^2 indica que las diferencias entre las frecuencias observadas y esperadas son grandes, lo que sugiere que las variables no son independientes.
  - ➢ Un valor bajo de χ^2 indica que las frecuencias observadas están cerca de las esperadas, lo que sugiere que las variables son independientes.

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Índice Chi Cuadrado en nuestro ejemplo



✓Paso 1: Datos del Ejemplo

➢Tenemos 30 estudiantes con las siguientes variables:

- Género (hombre/mujer)
  - Mujeres: 10 (2 juegan cricket, 8 no juegan)
  - Hombres: 20 (13 juegan cricket, 7 no juegan)
- Clase (IX/X)
  - Clase IX: 14 (6 juegan cricket, 8 no juegan)
  - Clase X: 16 (9 juegan cricket, 7 no juegan)
- Altura (< 5.5 pies / >= 5.5 pies)
  - < 5.5 pies: 12 (5 juegan cricket, 7 no juegan)
  - >= 5.5 pies: 18 (10 juegan cricket, 8 no juegan)

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Índice Chi Cuadrado en nuestro ejemplo

✓Paso 2: Cálculo del Índice Chi Cuadrado por Variable

Variable Género

Género	Juegan (O)	No Juegan (O)	Totales (F_row)
Mujeres	2	8	10
Hombres	13	7	20
Totales	15	15	30

Calculamos los números esperados E_i:

Género	Juegan (E)	No Juegan (E)
Mujeres	$\frac{10\times15}{30} = 5$	$\frac{10\times15}{30} = 5$
Hombres	$\frac{20\times15}{30} = 10$	$\frac{20\times15}{30} = 10$

- Ahora, calculamos X^2para la variable Genero:

$$\chi^2 = \frac{(2-5)^2}{5} + \frac{(8-5)^2}{5} + \frac{(13-10)^2}{10} + \frac{(7-10)^2}{10} = \frac{9}{5} + \frac{9}{5} + \frac{9}{10} + \frac{9}{10} = 1.8 + 1.8 + 0.9 + 0.9 = 5.4$$

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Índice Chi Cuadrado en nuestro ejemplo

✓Paso 2: Cálculo del Índice Chi Cuadrado por Variable

Variable Clase

Clase	Juegan (O)	No Juegan (O)	Totales (F_row)
IX	6	8	14
X	9	7	16
Totales	15	15	30

Calculamos los números esperados E_i:

Clase	Juegan (E)	No Juegan (E)
IX	$\frac{14\times15}{30} = 7$	$\frac{14\times15}{30} = 7$
X	$\frac{16\times15}{30} = 8$	$\frac{16\times15}{30} = 8$

- Ahora, calculamos X^2para la variable Genero:

$$\chi^2 = \frac{(6-7)^2}{7} + \frac{(8-7)^2}{7} + \frac{(9-8)^2}{8} + \frac{(7-8)^2}{8} = 0.1429 + 0.1429 + 0.125 + 0.125 = 0.5358$$

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Índice Chi Cuadrado en nuestro ejemplo

✓Paso 2: Cálculo del Índice Chi Cuadrado por Variable

Variable Altura

Altura	Juegan (O)	No Juegan (O)	Totales (F_row)
< 5.5	5	7	12
>= 5.5	10	8	18
Totales	15	15	30

Calculamos los números esperados E_i:

Altura	Juegan (E)	No Juegan (E)
< 5.5	$\frac{12\times15}{30} = 6$	$\frac{12\times15}{30} = 6$
>= 5.5	$\frac{18\times15}{30} = 9$	$\frac{18\times15}{30} = 9$

- Ahora, calculamos X^2para la variable Genero:

$$\chi^2 = \frac{(5-6)^2}{6} + \frac{(7-6)^2}{6} + \frac{(10-9)^2}{9} + \frac{(8-9)^2}{9} = 0.1667 + 0.1667 + 0.1111 + 0.1111 = 0.5556$$

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Índice Chi Cuadrado en nuestro ejemplo

- ✓ **Paso 3: Comparación de Valores Chi Cuadrado**
  - ➢ X^2(Genero)=5.4
  - ➢ X^2(Clase)=0.5358
  - ➢ X^2(Altura)=0.5556

Conclusión: La variable Género tiene el valor de X^2 más alto (5.4), por lo que es la variable más significativa para la división de la población en este conjunto de datos.

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Ganancia de Información y Entropía

✓ Entropía:

- Mide la cantidad de incertidumbre o desorden en la distribución de clases de los datos en un nodo del árbol.
- En términos simples, es una medida de la pureza del nodo. Si todos los ejemplos en un nodo pertenecen a la misma clase, la entropía es 0 (sin incertidumbre). Si las clases están igualmente distribuidas, la entropía es máxima.
- Para un nodo con m clases posibles, la entropía H se calcula como:

$$H = -\sum_{i=1}^{m} p_i \log_2 p_i$$

Donde:

- p_i es la proporción de ejemplos que pertenecen a la clase i.

✓ Ganancia de información:

- Se usa para decidir qué atributo dividir en cada nodo del árbol. La idea es elegir el atributo que proporciona la mayor reducción en la incertidumbre, o sea, el que más "información" gana. La ganancia de información se calcula como la reducción en la entropía después de aplicar un atributo para dividir los datos.
- Para un atributo A en un nodo, la ganancia de información Gain(A) se calcula como:

$$\text{Gain}(A) = H(\text{nodo}) - \sum_{v \in \text{values}(A)} \frac{|D_v|}{|D|} H(D_v)$$

Donde:

- $H(\text{nodo})$ es la entropía del nodo antes de la partición,
- $\text{values}(A)$ son los posibles valores del atributo A,
- $|D_v|$ es el número de ejemplos en el subconjunto donde el atributo A tiene el valor v,
- $|D|$ es el número total de ejemplos en el nodo original,
- $H(D_v)$ es la entropía del subconjunto de datos con el valor v para el atributo A.

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Ganancia de Información y Entropía

- ✓ La entropía ponderada se refiere a la entropía de un nodo después de que los datos se han dividido en función de un atributo específico.
- ✓ Es una medida de la incertidumbre promedio en los subconjuntos resultantes después de la partición.
- ✓ Cálculo de la Entropía Ponderada
  - ➢ Para calcular la entropía ponderada, se considera la entropía de cada uno de los subconjuntos resultantes de la partición, ponderada por el tamaño de cada subconjunto en relación con el tamaño total del nodo original.
  - ➢ Supongamos que tenemos un nodo que se divide en varios subconjuntos basados en un atributo A con valores $\{v1, v2, \ldots, vk\}$. La entropía ponderada Hponderada después de la partición se calcula como:

$$H_{\text{ponderada}} = \sum_{i=1}^{k} \frac{|D_{v_i}|}{|D|} H(D_{v_i})$$

Donde:

- k es el número de valores distintos del atributo A,
- $|D_{v_i}|$ es el número de ejemplos en el subconjunto donde el atributo A tiene el valor v_i,
- $|D|$ es el número total de ejemplos en el nodo original,
- $H(D_{v_i})$ es la entropía del subconjunto de datos correspondiente al valor v_i del atributo A.

- ✓ Interpretación
  - ➢ Entropía Ponderada: Proporciona una medida de la incertidumbre en los subconjuntos resultantes en función de su tamaño relativo. Si un atributo divide los datos en subconjuntos que son relativamente homogéneos (baja entropía en cada subconjunto) y los subconjuntos tienen tamaños significativos, la entropía ponderada será baja.
  - ➢ Uso en Ganancia de Información: La entropía ponderada se usa para calcular la ganancia de información. La ganancia de información es la diferencia entre la entropía del nodo original y la entropía ponderada después de la partición. Un atributo que reduce significativamente la entropía ponderada en comparación con la entropía del nodo original es considerado útil para la clasificación, ya que proporciona más información sobre la estructura de los datos.

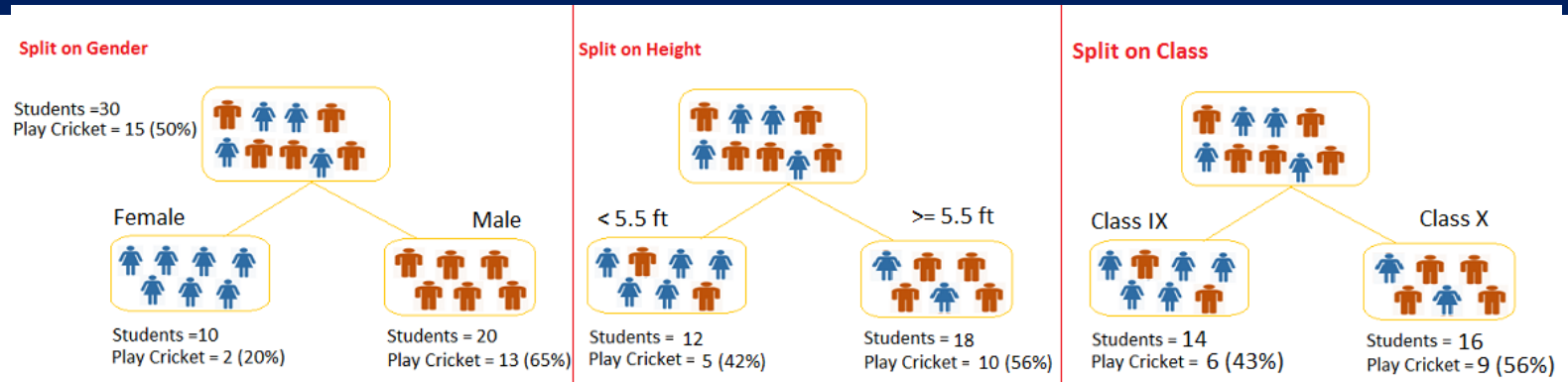
© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Ganancia de Información y Entropía

- ✓En resumen, en la construcción de un árbol de decisión, se calcula la entropía para medir la pureza de los nodos y la ganancia de información para seleccionar el mejor atributo para dividir los datos en cada paso del árbol.
- ✓El objetivo es construir un árbol que tenga la menor incertidumbre (entropía) posible en sus nodos.

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Cálculo de la Ganancia de Información y Entropía en nuestro ejemplo



✓ Paso 1: Datos del Ejemplo

➢ Tenemos 30 estudiantes con las siguientes variables:

- Género (hombre/mujer)
  - Mujeres: 10 (2 juegan cricket, 8 no juegan)
  - Hombres: 20 (13 juegan cricket, 7 no juegan)
- Clase (IX/X)
  - Clase IX: 14 (6 juegan cricket, 8 no juegan)
  - Clase X: 16 (9 juegan cricket, 7 no juegan)
- Altura (< 5.5 pies / >= 5.5 pies)
  - < 5.5 pies: 12 (5 juegan cricket, 7 no juegan)
  - >= 5.5 pies: 18 (10 juegan cricket, 8 no juegan)

✓ Paso 2: Cálculo de la Entropía del Conjunto Completo

➢ Para el conjunto completo:

- Juegan cricket: 15/30 = 0.5
- No juegan cricket: 15/30 = 0.5

$$H(S) = -(0.5 \log_2(0.5) + 0.5 \log_2(0.5)) = -(0.5 \cdot -1 + 0.5 \cdot -1) = 1$$

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Cálculo de la Ganancia de Información y Entropía en nuestro ejemplo

✓Paso 3: Cálculo de la Entropía y la Ganancia de Información por Variable

Variable Género

- Mujeres: 10 (2 juegan cricket, 8 no juegan)
- Hombres: 20 (13 juegan cricket, 7 no juegan)

Calculamos la entropía para cada subconjunto:

$$H(\text{Mujeres}) = -\left(\frac{2}{10}\log_2\frac{2}{10} + \frac{8}{10}\log_2\frac{8}{10}\right) = -\left(0.2\log_2 0.2 + 0.8\log_2 0.8\right) = -(0.2\cdot -2.32 + 0.8\cdot -0.32) = 0.72$$

$$H(\text{Hombres}) = -\left(\frac{13}{20}\log_2\frac{13}{20} + \frac{7}{20}\log_2\frac{7}{20}\right) = -(0.65\log_2 0.65 + 0.35\log_2 0.35) = 0.93$$

Calculamos la entropía ponderada:

$$H_{\text{ponderada}}(\text{Género}) = \frac{10}{30}\cdot 0.72 + \frac{20}{30}\cdot 0.93 = 0.24\cdot 0.72 + 0.67\cdot 0.93 = 0.72$$

Ganancia de Información para Género:

$$\text{Ganancia de Información}(\text{Género}) = H(S) - H_{\text{ponderada}}(\text{Género}) = 1 - 0.85 = 0.15$$

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Cálculo de la Ganancia de Información y Entropía en nuestro ejemplo

✓Paso 3: Cálculo de la Entropía y la Ganancia de Información por Variable

Variable Clase

- Clase IX: 14 (6 juegan cricket, 8 no juegan)
- Clase X: 16 (9 juegan cricket, 7 no juegan)

Calculamos la entropía para cada subconjunto:

$$H(\text{IX}) = -\left(\frac{6}{14}\log_2\frac{6}{14} + \frac{8}{14}\log_2\frac{8}{14}\right) = -(0.43\log_2 0.43 + 0.57\log_2 0.57) = 0.99$$

$$H(\text{X}) = -\left(\frac{9}{16}\log_2\frac{9}{16} + \frac{7}{16}\log_2\frac{7}{16}\right) = -(0.56\log_2 0.56 + 0.44\log_2 0.44) = 0.99$$

Calculamos la entropía ponderada:

$$H_{\text{ponderada}}(\text{Clase}) = \frac{14}{30}\cdot 0.99 + \frac{16}{30}\cdot 0.99 = 0.4667 + 0.5333 = 0.99$$

Ganancia de Información para Clase:

$$\text{Ganancia de Información}(\text{Clase}) = H(S) - H_{\text{ponderada}}(\text{Clase}) = 1 - 0.99 = 0.01$$

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Cálculo de la Ganancia de Información y Entropía en nuestro ejemplo

✓Paso 3: Cálculo de la Entropía y la Ganancia de Información por Variable

Variable Altura

- < 5.5 pies: 12 (5 juegan cricket, 7 no juegan)
- **= 5.5 pies: 18 (10 juegan cricket, 8 no juegan)**

Calculamos la entropía para cada subconjunto:

$$H(<5.5) = -\left(\frac{5}{12}\log_2\frac{5}{12} + \frac{7}{12}\log_2\frac{7}{12}\right) = -(0.42\log_2 0.42 + 0.58\log_2 0.58) = 0.99$$

$$H(\geq 5.5) = -\left(\frac{10}{18}\log_2\frac{10}{18} + \frac{8}{18}\log_2\frac{8}{18}\right) = -(0.56\log_2 0.56 + 0.44\log_2 0.44) = 0.99$$

Calculamos la entropía ponderada:

$$H_{\text{ponderada}}(\text{Altura}) = \frac{12}{30}\cdot 0.99 + \frac{18}{30}\cdot 0.99 = 0.396 + 0.594 = 0.99$$

Ganancia de Información para Altura:

$$\text{Ganancia de Información}(\text{Altura}) = H(S) - H_{\text{ponderada}}(\text{Altura}) = 1 - 0.99 = 0.01$$

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Cálculo de la Ganancia de Información y Entropía en nuestro ejemplo

✓Conclusión

➢La ganancia de información de las tres variables:

- Género = 0.15
- Clase = 0.01
- Altura = 0.01

La variable Género tiene la mayor ganancia de información (0.15), por lo que es la más significativa para dividir la población en este conjunto de datos.

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Cálculo de Reducción en la varianza (regresión)

- ✓Los algoritmos anteriores se aplicaban para problemas de clasificación con variables objetivo categóricas. La reducción en la varianza es un algoritmo usado para variables objetivo continuas (problemas de regresión).
- ✓Este algoritmo usa la fórmula estándar de la varianza para escoger el criterio de división. La división con la varianza más baja se escoge para dividir la población:

$$\text{Varianza} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

donde:

- X_i es el valor de la variable (1 si juega cricket, 0 si no juega cricket),
- $\bar{X}$ es la media de los valores,
- n es el número total de datos en el nodo.

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Cálculo de Reducción en la varianza (regresión)

✓ **Considerando nuestro ejemplo y asumiendo que vamos a convertir a valores numéricos de 1 para “play cricket” y valor de 0 para “No play cricket”.**

Criterio	Mean	Varianza	
Nodo General	$\bar{X} = \frac{15}{30} = 0.5$	$\text{Varianza}_{\text{Total}} = \frac{1}{30}\left[15 \times (1-0.5)^2 + 15 \times (0-0.5)^2\right]$	0.25
Nodo Mujeres	$\bar{X}_{\text{Mujeres}} = \frac{2}{10} = 0.2$	$\text{Varianza}_{\text{Mujeres}} = \frac{1}{10}\left[2 \times (1-0.2)^2 + 8 \times (0-0.2)^2\right]$	0.160
Nodo Hombres	$\bar{X}_{\text{Hombres}} = \frac{13}{20} = 0.65$	$\text{Varianza}_{\text{Hombres}} = \frac{1}{20}\left[13 \times (1-0.65)^2 + 7 \times (0-0.65)^2\right]$	0.228
Genero ponderado	$\text{Varianza}_{\text{Ponderada, Género}} = \frac{10}{30} \times \text{Varianza}_{\text{Mujeres}} + \frac{20}{30} \times \text{Varianza}_{\text{Hombres}}$		0.205
Nodo Clase IX	$\bar{X}_{\text{Clase IX}} = \frac{6}{14} \approx 0.429$	$\text{Varianza}_{\text{Clase IX}} = \frac{1}{14}\left[6 \times (1-0.429)^2 + 8 \times (0-0.429)^2\right]$	0.244
Nodo Clase X	$\bar{X}_{\text{Clase X}} = \frac{9}{16} = 0.5625$	$\text{Varianza}_{\text{Clase X}} = \frac{1}{16}\left[9 \times (1-0.5625)^2 + 7 \times (0-0.5625)^2\right]$	0.244
Clase ponderado	$\text{Varianza}_{\text{Ponderada, Clase}} = \frac{14}{30} \times \text{Varianza}_{\text{Clase IX}} + \frac{16}{30} \times \text{Varianza}_{\text{Clase X}}$		0.244
Nodo Altura <5.5	$\bar{X}_{< 5.5 \text{ pies}} = \frac{5}{12} \approx 0.417$	$\text{Varianza}_{< 5.5 \text{ pies}} = \frac{1}{12}\left[5 \times (1-0.417)^2 + 7 \times (0-0.417)^2\right]$	0.242
Nodo Altura >=5.5	$\bar{X}_{>= 5.5 \text{ pies}} = \frac{10}{18} \approx 0.556$	$\text{Varianza}_{>= 5.5 \text{ pies}} = \frac{1}{18}\left[10 \times (1-0.556)^2 + 8 \times (0-0.556)^2\right]$	0.244
Altura ponderado	$\text{Varianza}_{\text{Ponderada, Altura}} = \frac{12}{30} \times \text{Varianza}_{< 5.5 \text{ pies}} + \frac{18}{30} \times \text{Varianza}_{>= 5.5 \text{ pies}}$		0.245

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

Cálculo de Reducción en la varianza (regresión)

✓Conclusión

- ➢Las varianzas ponderadas para cada división son:
  - Género: 0.205
  - Clase: 0.244
  - Altura: 0.245
- ➢En este caso, la división con la varianza ponderada más baja es la división por género (0.205), lo que la hace la mejor opción para dividir la población según estos datos.

© 2024 EVC - Facultad de Ingeniería – Departamento de Sistemas

PREGUNTAS??

Dr. Edwin Valencia Castillo
Departamento de Sistemas
Facultad de Ingeniería
Universidad Nacional de Cajamarca
2024