

Ciencia de Datos

Ciencia que permite obtener conocimiento a partir de los datos.

Villalbilla, Ensanche de Vallecas, Villalbilla, Alcalá de Henares, Alcalá de Henares, Alcalá de Henares, Cifuentes, Cifuentes, El Casar, Fuente el Saz del Jarama , ND , Coslada, Daganzo de Arriba, Camarma de Esteruelas, Alcalá de Henares, Arganda del Rey, Coslada, Alcalá de Henares, Alcalá de Henares, Madrid, Madrid, Mejorada del Campo, Alcalá de Henares, Alcalá de Henares, Alcalá de Henares, ND, ND, ND, Alcalá de Henares, Guadalajara, Guadalajara, Torrejón de Ardoz, Torrejón de Ardoz, Torrejón de Ardoz, Torrejón de Ardoz, Alcalá de Henares, Torrejón de Ardoz, Guadalajara, ND, ND, Chiloheches, Alcalá de Henares, ND, ND, ND, Guadalajara, Guadalajara, Coslada, Guadalajara, Cabanillas del Campo, Alcalá de Henares, Madrid, ND, ND, Daganzo, Alcalá de Henares, Torres de la Alameda, Velilla de San Antonio, Daganzo, Guadalajara, ND, Guadalajara, Cobeña, Galapagos, Madrid, Alcalá de Henares, Madrid, Coslada, ND, Alcalá de Henares, ND, Alovera, Torrejón de Ardoz

16.5, 34.8, 20.7, 6.2, 4.4, 3.4, 24, 24, 32, 30, 33, 27, 15, 9.4, 2.1, 34, 24, 12, 4.4, 28, 31.4, 21.6, 3.1, 4.5, 5.1, 4, 3.2, 25, 4.5, 20, 34, 12, 12, 12, 12, 5, 19, 30, 5.5, 38, 25, 3.7, 9, 30, 13, 30, 30, 26, 30, 30, 1, 26, 22, 10, 9.7, 11, 24.1, 33, 17.2, 27, 24, 27, 21, 28, 30, 4, 46, 29, 3.7, 2.7, 8.1, 19, 16

Ciencia de Datos

...obtener conocimiento a partir de los datos

¿sobre qué?

Sobre lo que observamos de las cosas y de los hechos

Villalbilla, Ensanche de Vallecas, Villalbilla, Alcalá de Henares,...

Localidades de residencia de los alumnos

16.5, 34.8, 20.7, 6.2, ...

Distancias en Km desde su localidad de residencia a la universidad

N, S, N, S, ...

Posesión del carnet de conducir

Ciencia de Datos

lo que observamos de las cosas y de los hechos

Característica [1]: Dicho de una cualidad: Que da carácter o sirve para distinguir a alguien o algo de sus semejantes.

Cualitativas, Cuantitativas (Aritméticas) y Binarias (Lógicas)

Localidades, Distancias, Carnet

Atributo: Propiedad o característica de un objeto que puede variar, ya sea de un objeto a otro o de un tiempo a otro.

[2] Una característica cualitativa de un individuo, usualmente empleada para distinguirla de una variable o característica cuantitativa.

Variable: [2] Características medibles cuyo valor varía. Es aconsejable llamar así a características no medibles (medibles = observables), así como distinguir entre variable y variable aleatoria.

Propiedad [1]: Atributo o cualidad esencial de alguien o algo.

Cualidad [1]: Elemento o carácter distintivo de la naturaleza de alguien o algo.

Ciencia de Datos

...obtener conocimiento a partir de los datos

¿y que son los datos?

Dato [1]: Información dispuesta de manera adecuada para su tratamiento por un ordenador

Un dato es el valor obtenido para una característica en una observación

¿Cómo pueden ser los datos?

Cualitativos: Para las características cualitativas. En forma de nombres o niveles (Discretos).

= **Dato Nominal.** Proporciona suficiente información para nombrar la característica.

= y > **Dato Ordinal.** Proporciona suficiente información para ordenar las observaciones.

Cuantitativos: Para las características cuantitativas. [2] En forma de cantidades numéricas (sobre las que se pueden hacer operaciones aritméticas), como los conteos o mediciones.

+, -, *, / **Dato Discreto.** La característica tiene valores enteros.

+, -, *, / **Dato Continuo.** La característica tiene valores reales.

+, - **Dato Intervalo (Interval).** La característica pertenece a una escala de intervalo sin 0.

+, -, *, / **Dato Razón (Ratio).** Todas las operaciones aritméticas son posibles. Existe el 0.

Binarios / Lógicos:

Dato Binario / Lógico. Dato Binario[2] La variable toma dos valores, como éxito/fallo.

Ciencia de Datos

Cualitativos: Siempre son discretos.

= **Dato Nominal.** Localidad de residencia del estudiante. Pero también numéricos (pero no se pueden realizar operaciones con los números): DNI del estudiante.

= y > **Dato Ordinal.** Mayor curso en el que está matriculado.

Cuantitativos: Los datos de intervalo o proporción pueden ser discretos o continuos.

+, -, *, / **Dato Discreto.** Número de asignaturas en las que está matriculado el estudiante.

+, -, *, / **Dato Continuo.** Distancia desde su localidad de residencia hasta la Universidad.

+, - **Dato Intervalo (Interval).** Calificación Académica. El cero es arbitrario, se puede establecer en un nivel de conocimientos u otro. Un 3 no significa que se tengan exactamente la mitad de conocimientos que un 6, pero sí que se tienen 3 puntos menos.

+, -, *, / **Dato Proporción (Ratio).** Distancia desde su localidad de residencia hasta la Universidad. El 0 si significaría que se vive en el edificio donde se dan las clases. 10 km si significa que se vive a la mitad de distancia que 20.

Binarios / Lógicos:

Dato Binario / Lógico. Posesión del carnet de conducir.

Ciencia de Datos

...obtener conocimiento a partir de los datos

¿Cuántos datos tenemos?

Todos.

Entonces tenemos la **Población** [2]: Finita o infinita colección de individuos. Ha desplazado al término “Universo” Y es sinónimo del término “aggregate” (conjunto).

[1] Conjunto de los individuos o cosas sometido a una evaluación estadística.

Cocemos con certeza. Estadística **Descriptiva**.

Parte.

Entonces tenemos una **Muestra** [2]: Parte de una población, que es obtenida por algún proceso con el objeto de investigar las propiedades de la población de la que procede.

[1] Parte o porción extraída de un conjunto por métodos que permiten considerarla como representativa de él

Conocemos con certeza la muestra pero con **probabilidad** la población. **Inferencia** Estadística

Ciencia de Datos

El método científico tradicional obtenía el conocimiento a través de los siguientes pasos:

1. Definir el problema. Puede ser de ciencias naturales, sociales, ingeniería...
2. Recoger la información existente ya publicada sobre el problema.
3. Formular una o más hipótesis.
4. Recoger datos experimentales.
5. Analizar la información del nuevo conjunto de datos.
6. Establecer las conclusiones.

La Ciencia de Datos o Data Science amplia este proceso de obtención del conocimiento modificando el orden habitual y permitiendo la

Extracción Automática del Conocimiento

Es decir, que el paso 3 pueda ser una consecuencia del paso 5, o lo que es lo mismo, que del análisis de datos se obtendrían consecuencias no buscadas a priori

Ciencia de Datos

La Data Science se fundamenta en el conocimiento proveniente principalmente, al menos, de 4 áreas de conocimiento:

1. Bases de Datos.
2. Estadística.
3. Inteligencia Artificial.
4. Programación.

Y ha desarrollado, al menos, 3 áreas de conocimiento nuevas:

1. Data Warehousing.
2. Data Mining.
3. Visualización.

Ciencia de Datos

El término **Big Data** tiene su origen en la gran cantidad de datos que gracias a las nuevas tecnologías se obtienen actualmente en la casi todas las disciplinas del conocimiento y abarca todas las actividades relacionadas con el desarrollo y utilización de dichas enormes bases de datos, que fundamentalmente consiste en:

- Desarrollo las bases de datos.
- Extracción de la posible información contenida en las mismas.
- Visualización de los datos y de la información extraída.
- Posibles áreas de aplicación.

La Ciencia de Datos proporciona el conocimiento necesario para llevar a cabo dichas actividades.



Ciencia de datos

El número de datos a describir de una característica cuantitativa puede ser muy grande. Para facilitar su análisis estadístico se pueden agrupar utilizando **Clases de Equivalencia**.

Una clase de equivalencia es un conjunto de datos agrupados por un criterio. El número de clases de equivalencia con las que trabajamos depende del número de datos que tenemos.

Cuando se utilizan clases de equivalencia se utilizan los siguientes parámetros:

1. Rango. Con los datos ordenados por magnitud, diferencia entre el menor del mayor.
2. Límites. Son los números menor y mayor se cada clase.
3. Amplitud. Límite superior menos límite inferior de una clase.
4. Marca. Representa a la clase. Es el punto medio. Suma de los límites dividido entre 2.

Ciencia de datos

Obtener el rango de los datos, agruparlos en 5 clases de equivalencia, incluyendo en cada clase los valores en la misma decena, los límites y amplitud de cada clase, la marca de la clase

16.5, 34.8, 20.7, 6.2, 4.4, 3.4, 24, 24, 32, 30, 33, 27, 15, 9.4, 2.1, 34, 24, 12, 4.4, 28, 31.4, 21.6, 3.1, 4.5, 5.1, 4, 3.2, 25, 4.5, 20, 34, 12, 12, 12, 12, 5, 19, 30, 5.5, 38, 25, 3.7, 9, 30, 13, 30, 30, 26, 30, 30, 1, 26, 22, 10, 9.7, 11, 24.1, 33, 17.2, 27, 24, 27, 21, 28, 30, 4, 46, 29, 3.7, 2.7, 8.1, 19, 16.

Ciencia de datos

En el primer análisis de los datos se cuantifica la **Frecuencia** de aparición de los mismos. La frecuencia no tiene una sola definición sino que se define a través de los siguientes conceptos:

- **Frecuencia Absoluta f_i :** Número de apariciones de un dato. El término **Frecuencia** se suele referir a Frecuencia Absoluta, Frecuencia [2]: Número de ocurrencias de un determinado tipo de evento, o número de elementos de una población que caen en una clase específica.
- **Frecuencia Relativa f_{ri} :** Frecuencia absoluta de un dato dividida entre el número de datos. [2]: proporción del número total de ocurrencias o del número total de elementos.
- **Frecuencia Acumulada (Absoluta f_{ai} o Relativa f_{rai}):** Con los datos ordenados por magnitud, suma de las frecuencias absolutas o relativas de todos los datos inferiores al dato más la del dato.

Para datos agrupados en clases de equivalencia, el concepto de dato se cambia por el de clase.

Ciencia de datos

Se deben realizar los estudios estadísticos sobre los satélites menores, radio menor de 50 km. Radio en km: 13, 15, 16, 20, 20, 22, 27, 29, 30, 33, 34, 42

Agrupar los datos en 4 clases de equivalencia, incluyendo en cada clase los valores en la misma decena y dar utilizando la marca de la clase, la frecuencia absoluta, relativa y acumulada, tanto relativa como absoluta de cada clase

Solución:

Frecuencia Absoluta: 15: 3, 25: 5, 35: 3, 45: 1 (n=12)

Frecuencia Relativa: 15: 0.25, 25: 0.41, 35: 0.25, 45: 0.083

Frecuencia Acumulada (Absoluta / Relativa): 15: 3 / 0.25, 25: 8 / 0.66, 35: 11, 0.916, 45: 12 / 1

Distribución de Frecuencias (Absoluta o Relativa): (15, 3), (25, 5)...

Ciencia de datos

El segundo análisis de los datos se basa en el cálculo de la **Media** y la **Moda**.

[1] define **Media** como: 3. Mat. Número que resulta al efectuar una serie determinada de operaciones con un conjunto de números y que, **en determinadas condiciones**, puede representar por si solo a todo el conjunto. Recibe diferentes denominaciones según las operaciones que se realicen para obtenerlo, así: aritmética, geométrica, armónica.

La media más utilizada en Estadística es la **Aritmética**:

$$\bar{x}_a = \frac{\sum_{i=1}^n x_i}{n} = \frac{\sum_{j=1}^m f_j x_j}{\sum_{i=1}^m f_j}$$

El término **Media** se suele referir a **Media Aritmética**.

Existen otras medias, por ejemplo la **Geométrica** y la **Armónica**:

$$\bar{x}_g = (\prod_{i=1}^n x_i)^{1/n} \quad 1/\bar{x}_h = \frac{\sum_{i=1}^n 1/x_i}{n}$$

Para datos agrupados en clases de equivalencia, el concepto de dato se cambia por el de clase.

Ciencia de datos

Para los datos de las distancias desde su casa hasta la Universidad de los estudiantes de Ciencia de datos calcular la media aritmética para los datos agrupados en las cinco clases de las decenas y para los datos sin agrupar:

(16.5, 34.8, 20.7, 6.2, 4.4, 3.4, 24, 24, 32, 30, 33, 27, 15, 9.4, 2.1, 34, 24, 12, 4.4, 28, 31.4, 21.6, 3.1, 4.5, 5.1, 4, 3.2, 25, 4.5, 20, 34, 12, 12, 12, 5, 19, 30, 5.5, 38, 25, 3.7, 9, 30, 13, 30, 30, 26, 30, 30, 1, 26, 22, 10, 9.7, 11, 24.1, 33, 17.2, 27, 24, 27, 21, 28, 30, 4, 46, 29, 3.7, 2.7, 8.1, 19, 16)

$$\bar{x}_a = \frac{\sum_{j=1}^5 f_j x_j}{\sum_{j=1}^5 f_j} = \begin{cases} \frac{22.5+14.15+20.25+16.35+1.45}{\sum_{j=1}^5 f_j} \\ \sum_{i=1}^5 f_i = 22+14+20+16+1=73 \end{cases} \quad \bar{x}_a = \frac{1425}{73} = 19.52$$

$$\bar{x}_a = \frac{\sum_{i=1}^{73} x_i}{n} = \begin{cases} \frac{16.5+34.8+\dots+19+16}{73} \\ \end{cases} \quad \bar{x}_a = \frac{1353}{73} = 18.53$$

$$\bar{x}_a = \frac{\sum_{j=1}^{47} f_j x_j}{\sum_{j=1}^{47} f_j} = \begin{cases} \frac{1.1+\dots+3.7.2+\dots+30.8+\dots+1.46}{F_n} \\ \sum_{j=1}^{47} f_j = 1+\dots+2+\dots+30+\dots+1=73 \end{cases} \quad \bar{x}_a = \frac{1353}{73} = 18.53$$

Ciencia de datos

Para los satélites menores de Urano, con los datos sin agrupar y agrupados por decenas calcular la media aritmética:

(13, 15, 16, 20, 20, 22, 27, 29, 30, 33, 34, 42)

Solución:

$$\bar{x}_a = \frac{\sum_{i=1}^{11} f_i x_i}{F_n} = \begin{cases} \frac{1.13+15+16+2.20+22+27+29+30+33+34+42}{F_n} \\ F_n = \sum_{i=1}^{11} f_i = 1+1+1+2+1+1+1+1+1+1+1=12 \end{cases}$$

$$\bar{x}_a = \frac{301}{12} = 25.08$$

$$\bar{x}_a = \frac{\sum_{i=1}^4 f_i x_i}{F_n} = \begin{cases} \frac{3.15+5.25+3.35+1.45}{F_n} \\ F_n = \sum_{i=1}^4 f_i = 3+5+3+1=12 \end{cases}$$

$$\bar{x}_a = \frac{320}{12} = 26.67$$

Ciencia de datos

Moore estableció en 1965 que el número de transistores en un circuito integrado se duplicaría cada año. En 1975 modificó su propia ley y estableció que se duplicaría cada dos años ¿Por qué modificó su ley? Lo que Moore sabía sobre la evolución de los procesadores entre 1971 y 1073 era: (Nombre del procesador, fecha de creación, número de transistores): 4004, 1971, 2300; 8008, 1972, 3.500; 8080, 1973, 4500; 8086,

En 1971 el procesador 4004 tenía 2300 transistores. En 1972 el procesador 8008 tenía 3500. En consecuencia, el índice de crecimiento durante 1971 fue: $3500 = 2300 \cdot x_1 \rightarrow x_1 = 1,52$. En 1973 el procesador 8080 tenía 4500 transistores. En consecuencia, el índice de crecimiento durante 1972 fue: $4500 = 3500 \cdot x_2 \rightarrow x_2 = 1,29$

La media a aplicar es geométrica porque los índices multiplican el valor anterior

$$\bar{x}_g = \left(\prod_{i=1}^2 x_i^{f_i} \right)^{1/F_2} = \begin{cases} \sqrt[2]{1,52^1 \cdot 1,29^1} \\ F_2 = \sum_{i=1}^2 f_i = 1+1=2 \end{cases} \quad \bar{x}_a = \sqrt{1,96} = 1,4$$

$$1,4 < 2 \quad / \quad 1,4 \cdot 1,4 = 1,96 \sim 2 \quad \text{Tenía razones para modificar su ley}$$

Ciencia de datos

Moda [2]: es aquel valor de la variable (característica) que es observado por el mayor número de elementos de la población. Puede haber más de una moda en la población.

Para los datos de las distancias desde su casa hasta la Universidad de los estudiantes de Ciencia de datos calcular la moda:

(16.5, 34.8, 20.7, 6.2, 4.4, 3.4, 24, 24, 32, 30, 33, 27, 15, 9.4, 2.1, 34, 24, 12, 4.4, 28, 31.4, 21.6, 3.1, 4.5, 5.1, 4, 3.2, 25, 4.5, 20, 34, 12, 12, 12, 12, 5, 19, 30, 5.5, 38, 25, 3.7, 9, 30, 13, 30, 30, 26, 30, 30, 1, 26, 22, 10, 9.7, 11, 24.1, 33, 17.2, 27, 24, 27, 21, 28, 30, 4, 46, 29, 3.7, 2.7, 8.1, 19, 16)

Moda: $x_{MOD} = 30$ con Frecuencia Absoluta 8

Para los satélites menores de Urano calcular la moda:

(13, 15, 16, 20, 20, 22, 27, 29, 30, 33, 34, 42)

Moda: $x_{MOD} = 20$

Ciencia de datos

El tercer análisis de los datos se basa en el **cálculo de las medidas de dispersión**.

Las medidas de dispersión pueden ser:

- **Dispersión Absoluta.** Se calcula a través de las medidas **Desviación**, **Varianza**, para la media y **Rango**, para las medidas de ordenación.
- **Dispersión Relativa.** Cociente entre la media aritmética y la dispersión absoluta. Si la dispersión absoluta se ha calculado mediante la desviación típica, la dispersión relativa se denomina **Coeficiente de Variación de Pearson**.

Ciencia de datos

Las desviación más utilizada en Estadística es la **Estándar o Típica**:

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} = \sqrt{\frac{\sum_{j=1}^m f_j (x_j - \bar{x})^2}{\sum_{j=1}^m f_j}}$$

Tchebychev: En cualquier distribución entre la media y k desviaciones típicas se encuentran, al menos, el $(1 - \frac{1}{k^2})$. 100% de los datos.

Existen otras desviaciones, por ejemplo la **Desviación Media**:

$$s = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n} = \frac{\sum_{j=1}^m f_j |x_j - \bar{x}|}{\sum_{j=1}^m f_j}$$

Para datos agrupados en clases de equivalencia, el concepto de dato se cambia por el de clase.

Ciencia de datos

Para los satélites menores de Urano: Calcular la desviación típica con los datos sin agrupar y agrupados.

(13, 15, 16, 20, 20, 22, 27, 29, 30, 33, 34, 42)

Solución:

$$\bar{x}_a = \frac{\sum_{i=1}^{12} x_i}{n} = \frac{13+15+16+20+20+22+27+29+30+33+34+42}{12} = \frac{301}{12} = 25.08$$

$$s = \sqrt{\frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{\sum_{i=1}^n f_i}} = \sqrt{\frac{(13 - 25.08)^2 + \dots + 2(20 - 25.08)^2 + (42 - 25.08)^2}{12}} = 8.48$$

$$s = \sqrt{\frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{\sum_{i=1}^n f_i}} = \sqrt{\frac{3.(15 - 26.66)^2 + 5(25 - 26.66)^2 + 3(35 - 26.66)^2 + (45 - 26.66)^2}{12}} = 7.07$$

Ciencia de datos

[1] define **Varianza** como: 1.f. Estad. Media de las desviaciones cuadráticas de una variable aleatoria, referidas al valor medio de esta:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = \frac{\sum_{j=1}^m f_j (x_j - \bar{x})^2}{\sum_{j=1}^m f_j}$$

Calcular la varianza de los satélites menores de Urano.

(13, 15, 16, 20, 20, 22, 27, 29, 30, 33, 34, 42)

Solución:

$$\Rightarrow s^2 = \frac{\sum_{i=1}^n f_i (x_i - \bar{x})^2}{\sum_{i=1}^n f_i} = 71.91$$

Ciencia de datos

El cuarto análisis de los datos se basa en las **medidas de ordenación**.

Las medidas de ordenación son la **mediana** y los **cuantiles**.

- **Mediana** (\tilde{x}). [1] la define como: 9.f. Mat. Elemento de una serie ordenada de valores crecientes de forma que la divide en dos partes iguales, superiores e inferiores a él.
- **Cuantiles**. Elementos que permiten dividir un conjunto ordenado de datos en un conjunto de partes de igual tamaño. Pueden ser Cuartiles, Deciles y Percentiles:
 - **Cuartil** (\tilde{x}_c ; $c = \frac{1}{4}, \frac{2}{4}, \frac{3}{4}$): Elementos de una serie ordenada de valores crecientes de forma que la dividen en cuatro partes iguales.
 - **Decil** (\tilde{x}_d ; $d = \frac{1}{10}, \frac{2}{10}, \dots, \frac{9}{10}$): Elementos de una serie ordenada de valores crecientes de forma que la dividen en diez partes iguales.
 - **Percentil** (\tilde{x}_p ; $p = \frac{1}{100}, \frac{2}{100}, \dots, \frac{99}{100}$): Elementos de una serie ordenada de valores crecientes de forma que la dividen en cien partes iguales.

Ciencia de datos

Para datos no agrupados, una vez ordenados por magnitud:

La **mediana** se calcula de la siguiente manera:

- Si n es par: $\tilde{x} = \frac{x_{n/2} + x_{(n/2)+1}}{2}$
- Si n es impar, es el valor central: $\tilde{x} = x_{(n+1)/2}$

Los **cuantiles** se calculan (para d y p sería igual cambiando c) de la siguiente manera:

- si $nc \notin \mathbb{N}$: $\tilde{x}_{c,d,p} = x_{[nc]+1}$ $[nc]$ parte entera de nc
- si $nc \in \mathbb{N}$: $\tilde{x}_{c,d,p} = \frac{x_{nc} + x_{nc+1}}{2}$



Ciencia de datos

Medida de dispersión para las medidas de ordenación.

[1] define **Rango** como: 5.m.Estad. Amplitud de la variación de un fenómeno entre un límite menor y uno mayor claramente especificados.

Los rangos más utilizados en estadística son:

- Rango Intercuartílico: $Rc = \tilde{x}_{3/4} - \tilde{x}_{1/4}$
- Rango Interdecilico: $Rd = \tilde{x}_{9/10} - \tilde{x}_{1/10}$
- Rango Interpercentilico: $Rp = \tilde{x}_{90/100} - \tilde{x}_{10/100}$

Ciencia de datos

Calcular la mediana de los radios de los satélites menores de Urano.
(13, 15, 16, 20, 20, 22, 27, 29, 30, 33, 34, 42)

Solución:

$$n = 12, \text{ par: } \tilde{x} = \frac{x_{n/2} + x_{(n/2)+1}}{2} = \frac{x_6 + x_7}{2} = \frac{22 + 27}{2} = 24.5$$

Ciencia de datos

Calcular los cuartiles de los radios de los satélites menores de Urano.
(13, 15, 16, 20, 20, 22, 27, 29, 30, 33, 34, 42)

$$\text{Cuartiles: (2-mediana)} n = 12 \rightarrow \frac{12.1}{2} = 6 \in \mathbb{N} \rightarrow \tilde{x}_{\frac{1}{2}} = \frac{x_6 + x_7}{2} = \frac{22+27}{2} = 24.5$$

$$(1) n = 12 \rightarrow \frac{12.1}{4} = 3 \in \mathbb{N} \rightarrow \tilde{x}_{\frac{1}{4}} = \frac{x_3 + x_4}{2} = \frac{16+20}{2} = 18$$

$$(3) \tilde{x}_{\frac{3}{4}} = 31.5$$

Ciencia de datos

¿Por qué se estudia probabilidad al estudiar Data Science?

Dos razones fundamentalmente:

- Gran parte de los fundamentos teóricos de muchas técnicas utilizadas en la Ciencia de Datos utilizan conceptos de probabilidad. Por lo que es necesario conocer los conceptos básicos de probabilidad para entenderlos.
- Cuando los datos que se estudian son una muestra de una población, el objetivo del estudio es inducir las propiedades de la población a partir de la muestra. Para poder aplicar las teorías de la inducción se necesitan conceptos de probabilidad.

Ciencia de datos

[1] define **Experimento** como: hacer operaciones destinadas a descubrir, comprobar o demostrar determinados fenómenos o principios.

A partir de la definición de experimento puede definirse el **Experimento Aleatorio** como aquel en el que el resultado no está determinado.

[1] define **Suceso** como: En un experimento aleatorio, subconjunto del total de resultados posibles.

A partir de la definición de suceso puede definirse el **Suceso Elemental** como cada uno de los resultados más simples que pueden darse en la realización de un experimento aleatorio.

Ciencia de datos

A partir de la definición de suceso elemental se puede aplicar al estudio de sucesos la Teoría de Conjuntos, y se tienen las siguientes definiciones :

- **Espacio Muestral, E:** Conjunto cuyos elementos son los sucesos elementales de un experimento aleatorio.
- **Partes de E:** Conjunto cuyos elementos son todos los posibles subconjuntos de E.
- **Suceso Complementario \bar{A} :** De un suceso dado A es el que se verifica siempre que no se verifica A.
- **Suceso Seguro:** Aquel que siempre ocurre al realizar el experimento.
- **Suceso Imposible:** Aquel que nunca ocurre al realizar el experimento. Es el conjunto vacío (\emptyset)

Ciencia de datos

Identificar a través de un ejemplo basado en el lanzamiento de un dado los conceptos de: Experimento, suceso, suceso elemental, espacio muestral, suceso complementario, suceso imposible, y Partes de E.

Solución:

Experimento: Aleatorio: Lanzar un dado.

Suceso: Obtener un 3; Obtener un número par,...

Suceso Elemental: Obtener un 2; Obtener un 5;...

Espacio Muestral: {1, 2, 3, 4, 5, 6}.

Suceso Complementario: $A = \{2\}$ es $(A)^c = \{1, 3, 4, 5, 6\}$, no se verifica A.

Suceso Imposible: \emptyset

Partes de E: $\{\emptyset, 1, 2, 3, 4, 5, 6, \{1,2\}, \{1,3\}, \{1,4\}, \{1,5\}, \{1,6\}, \{2,3\}, \{2,4\}, \{2,5\}, \{2,6\}, \{3,4\}, \{3,5\}, \{3,6\}, \{4,5\}, \{4,6\}, \{5,6\}, \{1,2,3\}, \{1,2,4\}, \{1,2,5\}, \{1,2,6\}, \{1,3,4\}, \{1,3,5\}, \{1,3,6\}, \{1,4,5\}, \{1,4,6\}, \{1,5,6\}, \{2,3,4\}, \{2,3,5\}, \{2,3,6\}, \{2,4,5\}, \{2,4,6\}, \{2,5,6\}, \{3,4,5\}, \{3,4,6\}, \{3,5,6\}, \{4,5,6\}, \{1,2,3,4\}, \{1,2,3,5\}, \{1,2,3,6\}, \{1,2,4,5\}, \{1,2,4,6\}, \{1,2,5,6\}, \{1,3,4,5\}, \{1,3,4,6\}, \{1,3,5,6\}, \{1,4,5,6\}, \{2,3,4,5\}, \{2,3,4,6\}, \{2,3,5,6\}, \{2,4,5,6\}, \{3,4,5,6\}, \{1,2,3,4,5\}, \{1,2,3,4,6\}, \{1,2,3,5,6\}, \{1,2,4,5,6\}, \{1,3,4,5,6\}, \{2,3,4,5,6\}, \{1,2,3,4,5,6\}\}$

Ciencia de datos

A partir de la definición de suceso elemental se puede aplicar al estudio de sucesos la Teoría de Conjuntos, y se tienen las siguientes operaciones:

- **Unión o suma de conjuntos:** Sean A y B dos conjuntos, se define la unión de A y B, $A \cup B$ como el conjunto cuyos elementos pertenecen a A o a B.
- **Intersección de conjuntos:** Sean A y B dos conjuntos, se define la intersección de A y B, $A \cap B$ como el conjunto cuyos elementos pertenecen a A y a B.
- **Diferencia de conjuntos:** Sean A y B dos conjuntos, se define la diferencia de A y B, $A - B$ como el conjunto cuyos elementos pertenecen a A y no a B.
- **Diferencia simétrica de conjuntos:** Sean A y B dos conjuntos, se define la diferencia simétrica de A y B, $A \Delta B$ como el conjunto cuyos elementos pertenecen a A y a B, menos los que pertenecen a ambos a la vez.

Ciencia de datos

Identificar a través de un ejemplo basado en los sucesos: B: obtener un número par y C: obtener un número distinto de 1 o 2 del lanzamiento de un dado los conceptos de: Unión, intersección, diferencia y diferencia simétrica de conjuntos (sucesos).

Solución:

Unión: B: obtener un número par: {2, 4, 6}; C: obtener un número distinto de 1 o 2: {3, 4, 5, 6}. La unión de los conjuntos B y C, $B \cup C$, es $B \cup C = \{2, 3, 4, 5, 6\}$.

Intersección: La intersección de los sucesos B={2, 4, 6} y C={3, 4, 5, 6}, $B \cap C$, es $B \cap C = \{4, 6\}$.

Diferencia: La diferencia de los sucesos B y C, $B - C$, es $B - C = \{2\}$

Diferencia Simétrica: la diferencia simétrica de los sucesos B y C, $B - C$, es B

Ciencia de datos

[1] define **Probabilidad** como:

3. f. Mat. En un proceso aleatorio, razón entre el número de casos favorables y el número de casos posibles.

La definición clásica de Probabilidad o Regla de Laplace define la Probabilidad como:

La probabilidad de que aparezca un determinado suceso A, es el cociente entre el número de casos favorables a ese suceso y el número total de casos.

$$P(A) = \frac{n_A}{N}$$

Ciencia de datos

Se ha formado un equipo de investigación y análisis en Data Science formado por 10 estudiantes procedentes de las titulaciones de Informática y de Sistemas de Información. De Informática se han elegido 4 chicas y un chico; y de Sistemas de Información 2 chicas y tres chicos.

Determinar el espacio muestral del experimento escoger al azar de entre los miembros del equipo de investigación (a.1) un chico o una chica, (a.2) escoger un estudiante de una titulación y (a.3) escoger un chico o una chica teniendo en cuenta la titulación.

Calcular la probabilidad de escoger al azar de entre los miembros del equipo de investigación (b.1) un estudiante que sea de informática, (b.2) un estudiante que sea de sistemas de información, (b.3) un chico, (b.4) una chica

(a.1) {chico, chica}, (a.2) {i, si}, (a.3) {(chico, i), (chico, si), (chica, i), (chica, si)}

$$(b.1) P(i) = \frac{n_i}{N} = \frac{5}{10} = 0.5 \quad (b.2) P(si) = \frac{n_{si}}{N} = \frac{5}{10} = 0.5$$

Ciencia de datos

Las propiedades de una **Probabilidad** son:

1. $0 \leq P(A) \leq 1$
2. $P(\bar{A}) = 1 - P(A)$
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
4. Si $A \subset B = P(A) \leq P(B)$
5. $P(E) = 1$
6. $P(\emptyset) = 0$

son análogas a las de la frecuencia relativa

Ciencia de datos

Sabiendo como es la constitución del equipo de investigación en Data Science formado por estudiantes:

a) Verificar que la probabilidad de elegir al azar de entre los miembros del equipo un estudiante que sea de I, de SI, chico, chica, son menores que 1.

b) Verificar que las probabilidades de elegir al azar de entre los miembros del equipo un estudiante que sea de I y SI; chico o chica, son complementarias ¿lo son que sea de la SI y chico?

a) Según la propiedad 1: $0 \leq P(A) \leq 1 \rightarrow P(SI) = 0.5 \quad P(I) = 0.5 \quad P(chica) = 0.6 \quad P(chico) = 0.4$

b) Según la propiedad 2: $P(\bar{A}) = 1 - P(A) \rightarrow P(SI) = 1 - P(I) = 0.5; \quad P(chica) = 1 - P(chico) = 0.6$; No

Ciencia de datos

c) Calcular la probabilidad de elegir al azar de entre los miembros del equipo (a) un estudiante que sea chica siendo de I, y (b) chica o de I.

d) Verificar que la probabilidad de que un estudiante sea chica siendo de la I es menor que sea de I

e) Calcular (a) la probabilidad de que uno de los 10 estudiantes sea chico o chica y (b) que no sea de ninguna titulación

c) (a) $P(vo) = \frac{n_{chica\ I}}{N} = \frac{4}{10} = 0.4$ y según la propiedad 3 (b):

$P(A \cup B) = P(A) + P(B) - P(A \cap B) \rightarrow P(chica \cup I) = P(chica) + P(I) - P(chica \cap I) \rightarrow 0.6 + 0.5 - 0.4 = 0.7$

d) Según la propiedad 4 Si $A \subset B = P(A) \leq P(B) \rightarrow P(chica\ I) = 0.4 \leq P(I) = 0.5$

e) Según la propiedad 5 $P(E) = 1$ y según 6 $P(\emptyset)$

Ciencia de datos

Desde el punto de vista de como están relacionados entre si, los sucesos pueden ser:

- Independientes: Un conjunto de sucesos son independientes si la ocurrencia de cualquiera de ellos no modifica la probabilidad de ocurrencia del resto.
- Dependientes: Un conjunto de sucesos son dependientes si la ocurrencia de cualquiera de ellos modifica la probabilidad de ocurrencia del resto.

La probabilidad de sucesos independientes viene dada por la ecuación (para dos sucesos):

$$P(A \cap B) = P(A)P(B)$$

La probabilidad de sucesos dependientes se denomina *condicionada* y viene dada por la ecuación (para dos sucesos):

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

Donde $P(A|B)$ es la probabilidad de que ocurra A habiendo ocurrido B

Ciencia de datos

Desde el punto de vista de como están relacionados entre si, los sucesos pueden ser:

- Independientes: Un conjunto de sucesos son independientes si la ocurrencia de cualquiera de ellos no modifica la probabilidad de ocurrencia del resto.
- Dependientes: Un conjunto de sucesos son dependientes si la ocurrencia de cualquiera de ellos modifica la probabilidad de ocurrencia del resto.

La probabilidad de sucesos independientes viene dada por la ecuación (para dos sucesos):

$$P(A \cap B) = P(A)P(B)$$

La probabilidad de sucesos dependientes se denomina *condicionada* y viene dada por la ecuación (para dos sucesos):

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

Donde $P(A|B)$ es la probabilidad de que ocurra A habiendo ocurrido B

Ciencia de datos

Se va a elegir un estudiante como responsable del equipo. (a) ¿Qué probabilidad tiene el elegido de ser de Informática y chico? (b) ¿Qué probabilidad tiene una chica de cualquiera de las dos titulaciones de ser elegida como representante?

Calculamos primero $P(I)=0.5$

La probabilidad de ser chico siendo de Informática es $P(chico|I)=1/5=0,2$

La probabilidad de ser chico y de informática es:

$$P(chico \cap I) = P(chico|I)P(I) = 0,2 \cdot 0,5 = 0,1 \rightarrow 10\%$$

Para resolver el apartado (b) se aplica la propiedad 3 de la probabilidad y la probabilidad condicionada como en el (a) :

$$\begin{aligned} P((chica \cap I) \cup (chica \cap SI)) &= P(chica|I)P(I) + P(chica|SI)P(SI) \\ &= 0,8 \cdot 0,5 + 0,4 \cdot 0,5 = 0,6 \rightarrow 60\% \end{aligned}$$

Tema. Asociación

Los estudios de asociación buscan **encontrar patrones de aparición conjunta de sucesos**, es decir, ver si la probabilidad aparición conjunta de varios sucesos excede unos umbrales.

La primera medida del grado de asociación entre varios sucesos se denomina

soporte, s

y establece la probabilidad de aparición de los elementos del conjunto $P(E)$ que engloban a dichos sucesos en toda la muestra estudiada.

Formalmente se define como:

$$\forall \{A_i\}_{i=1}^{\infty} \subset P(E) \text{ con } A_i \cap A_j = \emptyset \forall i \neq j, s: P(E) \rightarrow \mathbb{R}^+ / s(A_i \cup A_j) = \frac{n_{A_i \cup A_j}}{n_T}$$

Tema. Asociación

Los estudios de asociación buscan **encontrar patrones de aparición conjunta de sucesos**, es decir, ver si la probabilidad aparición conjunta de varios sucesos excede unos umbrales.

La primera medida del grado de asociación entre varios sucesos se denomina

soporte, s

y establece la probabilidad de aparición de los elementos del conjunto $P(E)$ que engloban a dichos sucesos en toda la muestra estudiada.

Formalmente se define como:

$$\forall \{A_i\}_{i=1}^{\infty} \subset P(E) \text{ con } A_i \cap A_j = \emptyset \forall i \neq j, s: P(E) \rightarrow \mathbb{R}^+ / s(A_i \cup A_j) = \frac{n_{A_i \cup A_j}}{n_T}$$

El **umbral de aceptación de un soporte** para concluir que existe asociación entre sucesos **no es fijo**, sino que será **fijado arbitrariamente a priori**, y dependerá de los objetivos del estudio.

Es esencial tener presente el hecho de que los sucesos elementales no son equiprobables y en consecuencia para poder establecer la probabilidad de aparición de los elementos de $P(E)$ es **imprescindible disponer de una muestra de sucesos**.

Tema. Asociación

A partir del espacio muestral E de un experimento aleatorio obtenemos el conjunto partes de E , $P(E)$, formado por los 2^n subconjuntos de que se pueden formar con los elementos de E .

Ejercicio 1: Se deben establecer los conjuntos E y $P(E)$ para una cesta de la compra con los siguientes productos: Pan, Agua, Café, Leche y Naranjas.

Solución:

$$E = \{\text{Pan, Agua, Café, Leche, Naranjas}\}$$

$$P(E) = \{\emptyset, \{\text{Pan}\}, \{\text{Agua}\}, \{\text{Café}\}, \{\text{Leche}\}, \{\text{Naranjas}\}, \{\text{Pan, Agua}\}, \{\text{Pan, Café}\}, \{\text{Pan, Leche}\}, \{\text{Pan, Naranjas}\}, \{\text{Agua, Café}\}, \{\text{Agua, Leche}\}, \{\text{Agua, Naranjas}\}, \{\text{Café, Leche}\}, \{\text{Café, Naranjas}\}, \{\text{Leche, Naranjas}\}, \{\text{Pan, Agua, Café}\}, \{\text{Pan, Agua, Leche}\}, \{\text{Pan, Agua, Naranjas}\}, \{\text{Pan, Café, Leche}\}, \{\text{Pan, Café, Naranjas}\}, \{\text{Pan, Leche, Naranjas}\}, \{\text{Agua, Café, Leche}\}, \{\text{Agua, Café, Naranjas}\}, \{\text{Café, Leche, Naranjas}\}, \{\text{Pan, Agua, Café, Leche}\}, \{\text{Pan, Agua, Café, Naranjas}\}, \{\text{Pan, Agua, Leche, Naranjas}\}, \{\text{Pan, Café, Leche, Naranjas}\}, \{\text{Agua, Café, Leche, Naranjas}\}, \{\text{Pan, Agua, Café, Leche, Naranjas}\}\}$$

$$\text{Número de elementos de } P(E): 25 = 32$$

Tema. Asociación

Ejercicio: A partir de la muestra de seis cestas de la compra: {Pan,Agua,Leche,Naranjas}, {Pan,Agua,Café,Leche}, {Pan,Agua,Leche}, {Pan,Café,Leche}, {Pan,Agua}, {Leche}.

- a. ¿Cuál es el soporte de los conjuntos disjuntos $A_1=\{\text{Pan,Agua}\}$ y $A_2=\{\text{Leche}\}$. 
- b. ¿Qué otras asociaciones tendrían el mismo soporte que la estudiada?

Solución a:

El soporte se define como: $s(A_1 \cup A_2) = \frac{n_{A_1 \cup A_2}}{n_T} = p(A_1 \cup A_2)$.

Por la muestra obtenemos $n_{A_1 \cup A_2}$ y n_T

El conjunto: $A_1 \cup A_2 = \{\text{Pan, Agua, Leche}\}$, se encuentra en los sucesos:
{Pan,Agua,Leche,Naranjas}, {Pan,Agua,Café,Leche}, {Pan,Agua,Leche}.

Por lo tanto $n_{A_1 \cup A_2} = 3$

Los sucesos que forman la muestra son: {Pan,Agua,Leche,Naranjas},
{Pan,Agua,Café,Leche}, {Pan,Agua,Leche}, {Pan,Café,Leche}, {Pan,Agua}, {Leche}.

Por lo tanto $n_T = 6$

Por lo que el soporte es: $s = \frac{n_{A_1 \cup A_2}}{n_T} = \frac{3}{6} = 0.5 \equiv 50\%$

Solución b:

$A_1 = \{\text{Pan, Leche}\}$ y $A_2 = \{\text{Agua}\}$; $A_1 = \{\text{Leche, Agua}\}$ y $A_2 = \{\text{Pan}\}$.

Tema. Asociación

Medir únicamente el soporte no es suficiente para establecer el grado en el que los sucesos estudiados se encuentran asociados. 

La segunda medida del grado de asociación entre varios sucesos se denomina

confianza, c

y establece la probabilidad de aparición de los elementos del conjunto P (E) que engloban a dichos sucesos en el subconjunto de la muestra compuesto únicamente por los sucesos en los que se de uno de los dos conjuntos cuya asociación se está estudiando.

Formalmente se define como:

$$\forall \{A_i\}_{i=1}^{\infty} \subset P(E) \text{ con } A_i \cap A_j = \emptyset \forall i \neq j, c: P(E) \rightarrow \mathbb{R}^+ / c(A_i \cup A_j) = \frac{n_{A_i \cup A_j}}{n_{A_i}}$$

Al igual que para el soporte, El **umbral de aceptación de una confianza** para concluir que existe asociación entre sucesos **no es fijo, sino que será fijado arbitrariamente a priori**, y dependerá de los objetivos del estudio.

Tema. Asociación

Ejercicio: A partir de la muestra de seis cestas de la compra: {Pan,Agua,Leche,Naranjas}, {Pan,Agua,Café,Leche}, {Pan,Agua,Leche}, {Pan,Café,Leche}, {Pan,Agua}, {Leche}.

- a. ¿Cuál es la confianza de los conjuntos disjuntos $A_1=\{\text{Pan,Agua}\}$ y $A_2=\{\text{Leche}\}$.
- b. ¿Qué otras asociaciones tendrían la misma confianza que la estudiada?
- c. ?¿Cuál es la confianza de $A_2=\{\text{Leche}\} \rightarrow A_1=\{\text{Pan,Agua}\}$?

Solución a:

La confianza se define como: $c(A_1 \cup A_2) = \frac{n_{A_1 \cup A_2}}{n_{A_1}}$

Por la muestra obtenemos $n_{A_1 \cup A_2}$ y n_{A_1} y como sabemos $n_{A_1 \cup A_2} = 3$

Los sucesos de la muestra sucesos en los que aparece el conjunto $A_1=\{\text{Pan,Agua}\}$, son: {Pan,Agua,Leche,Naranjas} \ {Pan,Agua,Café,Leche}, {Pan,Agua,Leche}, {Pan, Agua}.

Por lo tanto $n_{A_1} = 4$

Por lo que la confianza es: $c = \frac{n_{A_1 \cup A_2}}{n_{A_1}} = \frac{3}{4} = 0.75 \equiv 75\%$

Solución b:

Ninguna, la confianza de $A_1=\{\text{Pan, Agua}\} \rightarrow A_2=\{\text{Leche}\} \neq A_2=\{\text{Leche}\} \rightarrow A_1=\{\text{Pan, Agua}\}$

Solución c:

$0.6 \equiv 60\%$

Tema. Asociación

Un análisis de asociación tiene una gran dificultad debido al elevado número de asociaciones que pueden darse incluso para espacios muestrales pequeños, por lo que es imprescindible utilizar algoritmos que permitan reducir el número de asociaciones a analizar.

Para n sucesos elementales, el número de asociaciones es: $3^n - 2^{n+1} + 1$

Ejercicio: A partir del espacio muestral: $E=\{\text{Pan, Agua, Café, Leche, Naranjas}\}$

- a. Calcular el numero de posible asociaciones a estudiar
- b. ¿Y para un espacio muestral de 10 elementos?
- c. ¿y de 20 elementos?

Solución a:

$$3^5 - 2^6 + 1 = 180$$



Solución b:

57002

Solución c:

3500000000

Tema. Asociación

Los dos pasos que suelen emplear los algoritmos de asociación están basados en el cálculo del soporte y la confianza en dos pasos secuenciales.

- A. **Identificación de las asociaciones frecuentes. Cálculo del Soporte, s.** Se fija un umbral de soporte y se identifican las asociaciones de los elementos de $P(E)$ que lo alcanzan o superan. El paso 2 solo se aplica sobre dichos sucesos.

Si el umbral fijado de aceptación del soporte fuese del 40%, la asociación $A1=\{Pan, Agua\} \rightarrow A2 = \{Leche\}$ sería aceptada, porque tiene un soporte del 50%, pero la asociación $A1=\{Pan, Café\} \rightarrow A2=\{Leche\}$ no lo sería, ya que su soporte es del 33%.

- B. **Identificación de las asociaciones de confianza. Cálculo de la Confianza, c.** Se fija un umbral de confianza y se identifican las asociaciones con soporte que lo alcanzan o superan. La confianza identifica que sucesos están asociados y el sentido de la asociación.

Si el umbral fijado de aceptación de la confianza fuese del 70%, la asociación $A1=\{Pan, Agua\} \rightarrow A2=\{Leche\}$ sería aceptada, porque tiene una confianza del 75% pero la asociación $A1=\{Leche\} \rightarrow A2=\{Pan, Agua\}$ no lo sería, ya que su confianza es del 60%.

Tema. Asociación

Algoritmo Apriori

El algoritmo apriori sigue el proceso genérico de dos pasos:

- A. **Identificación de las asociaciones frecuentes. Cálculo del Soporte, s.** Se basa en el hecho de que la medida de soporte es **antimonótona**: si un suceso es frecuente, su soporte supera el umbral, todos los subconjuntos de dicho conjunto son también frecuentes, ya que su soporte es mayor o igual que el del conjunto que los contiene.
- B. **Identificación de las asociaciones de confianza. Cálculo de la Confianza, c.** El algoritmo utiliza la función **ap-genrules**, que está basada en el teorema: Sean dos conjuntos A y B, si la asociación $A \rightarrow B - A$ no supera el umbral de confianza, entonces cualquier asociación $A' \rightarrow B - A'$, donde A' es cualquier subconjunto de A ($A' \subseteq A$) tampoco lo alcanzará.

Tema. Asociación

Algoritmo Apriori

Ejercicio: A partir de la muestra: {Pan,Agua,Leche,Naranjas}, {Pan,Agua,Café,Leche}, {Pan,Agua,Leche}, {Pan,Café,Leche}, {Pan,Agua}, {Leche}.

Comprobar para el suceso {Pan,Agua,Leche} que la medida de soporte es antimonótona.

Solución:

Para que se verifique la propiedad, si el suceso {Pan,Agua,Leche} es frecuente, los sucesos {Pan}, {Agua}, {Leche}, {Pan,Agua}, {Pan, Leche}, y {Agua,Leche}, deben serlo.

El soporte de {Pan,Agua,Leche} a partir de la muestra es: $3/6 = 0.5$

Si observamos el conjunto {Pan, Agua} vemos que aparece tres veces junto a Leche, y además aparece una vez más en solitario, por lo que su soporte es $4/6 = 0.67$, y en consecuencia es mayor que el de {Pan,Agua,Leche}.

Lo mismo se puede comprobar para el resto, todos son $> 50\%$, excepto {Agua,Leche}, que es del 50%

Asociación

Algoritmo Apriori

Paso A:

Subpaso A.1. Se pasa por todos los sucesos elementales para calcular su soporte y se eliminan los que no alcancen el umbral fijado.

Ejercicio: Si se fija el umbral de aceptación del soporte en el 50% ¿Cuáles son los sucesos elementales candidatos de la muestra: {Pan,Agua,Leche,Naranjas}, {Pan,Agua,Café,Leche}, {Pan,Agua,Leche}, {Pan,Café,Leche}, {Pan,Agua}, {Leche}?

Solución:

Los sucesos elementales son: {{Pan}, {Agua}, {Café}, {Leche}, {Naranjas}}.

Como hay 6 sucesos en la muestra para alcanzar el umbral el suceso tendrá que aparecer 3 o más veces en la muestra, de forma que su soporte sea 0,5 o más.

{Pan}: 5 veces, por lo tanto lo seleccionamos; {Agua}: 4 veces, lo seleccionamos; {Café} solo 2 veces, no lo seleccionamos; {Leche}: 5 veces y lo seleccionamos; {Naranjas}: 1 vez y no lo seleccionamos.

Por lo tanto después del subpaso A.1 hemos reducido el conjunto de sucesos elementales a: {{Pan}, {Agua}, {Leche}}.

Asociación

Algoritmo Apriori

Paso A:

Subpaso A.2. Se realizarán dos pasos sucesivos en cada dimensión, comenzando por los conjuntos de dos elementos y terminando cuando no sea posible identificar una dimensión en la que haya un soporte igual o mayor que el umbral. Consta de dos subpasos:

Subpaso A.2.1. Se aplica la función **apriori-gen** para identificar los sucesos candidatos en cada dimensión. Funcionamiento:

Se utilizan como base los conjuntos seleccionados para la dimensión anterior a través del método $F_{k-1} \times F_{k-1}$ que genera los conjunto candidatos en la dimensión k uniendo pares de conjuntos candidatos de la dimensión anterior k-1, pero solo aquellos pares en los que coinciden sus primeros k-2 elementos:

Sean $A = \{a_1, a_2, \dots, a_{k-1}\}$ y $B = \{b_1, b_2, \dots, b_{k-1}\}$ dos sucesos frecuentes identificados en la dimensión k-1. A y B solo se unirán para formar un suceso candidato en la dimensión k si se cumplen dos condiciones:

1. $a_i = b_i$ para $i = 1, 2, \dots, k-2$
 2. $a_{k-1} \neq b_{k-1}$
-

Asociación

Algoritmo Apriori

Paso A. Subpaso A.2.1.

Ejercicio: Utilizando el método $F_{k-1} \times F_{k-1}$ determinar los posibles sucesos candidatos de dos y tres dimensiones (suponiendo que todos los de dos tienen suficiente soporte) a partir de los sucesos elementales aceptados en el paso A.1: $\{\{\text{Pan}\}, \{\text{Agua}\}, \{\text{Leche}\}\}$

Solución:

$k=2 \rightarrow k-1=1$. Para poder unir dos conjuntos de un solo elemento deben coincidir los primeros k-2, 2-2=0 elementos y no coincidir los elementos $a(k-1) \neq b(k-1)$, que en este caso son $a(1) \neq b(1)$.

Por lo tanto consistirá en unir los sucesos elementales frecuentes, dichos conjuntos son: $\{\text{Pan, Agua}\}, \{\text{Pan, Leche}\}, \{\text{Agua, Leche}\}$

$k=3 \rightarrow k-1=2$. Para poder unir dos conjuntos de dos elementos deben coincidir los primeros k-2, 3-2=1 elemento y no coincidir los elementos $a(k-1) \neq b(k-1)$, que en este caso son $a(2) \neq b(2)$.

Por lo que solo se pueden unir, porque cumplen las dos condiciones: $\{\text{Pan, Agua}\}, \{\text{Pan, Leche}\}$, pero no se podría unir con ninguno de los dos conjuntos el conjunto $\{\text{Agua, Leche}\}$ porque en los dos casos $a(1) \neq b(1)$. En consecuencia el conjunto unión es $\{\text{Pan, Agua, Leche}\}$

Asociación

Algoritmo Apriori

Paso A:

Subpaso A.2.2. Se calcula el soporte de los sucesos candidatos identificados en el paso A.2.1. Tiene tres subpasos. Funcionamiento:

A.2.2.1. Partición de los sucesos candidatos con un hash tree

A.2.2.2. Partición de los sucesos de la muestra con el mismo hash tree.

A.2.2.3. Comparación de ambas particiones. Todas las hojas coincidentes incrementan en una unidad el numerador en el cálculo del soporte del suceso candidato.



Asociación

Algoritmo Apriori

Paso A. Subpaso A.2.2.1.

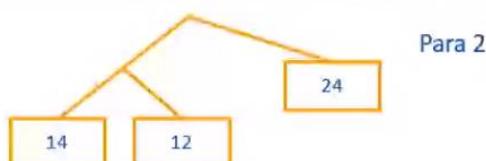
Ejercicio: Realizar una partición de los sucesos candidatos de dos y tres dimensiones utilizando un hash tree con función de partición: $h(p) = p \bmod 3$. Los sucesos candidatos son: {Pan, Agua}, {Pan, Leche}, {Agua, Leche}, {Pan, Agua, Leche}.

Solución:

Se numeran los sucesos elementales y se cambian los valores por números: Pan≡1, Agua≡2, Cafe≡3, Leche≡4, Naranjas≡5. Los sucesos candidatos serán ahora: {12}, {14}, {24}, {124}.

A partir de la función de partición los índices quedan agrupados según el resto que de su división entre 3, por ejemplo, la división de 1, 4 y 7 entre 3 dará 1, por lo que estarán en el mismo nodo.

Los nodos son {1, 4, 7}, {2, 5, 8} y {3, 6, 9} y se organizan de izquierda a derecha. Los sucesos que tengan como segundo dígito un 4 estarán primer nodo, al la izquierda, y los que tengan un 2, estarán en el nodo central. Según esto tenemos:



Asociación

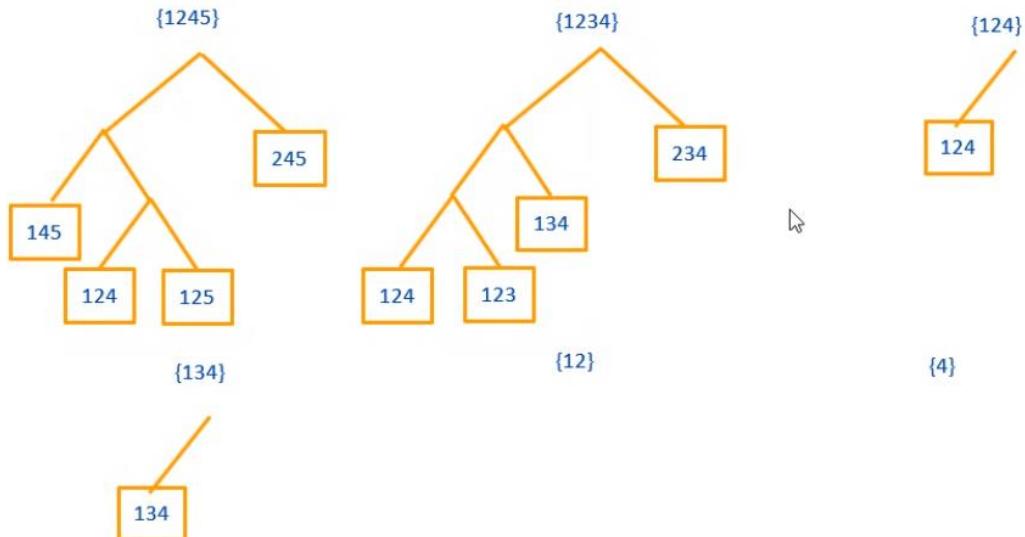
Algoritmo Apriori

Paso A. Subpaso A.2.2.2.

Ejercicio: Realizar una partición de los sucesos de la muestra utilizando el mismo hash tree.

Solución: {Pan,Agua,Leche,Naranjas}, {Pan,Agua,Café,Leche}, {Pan,Agua,Leche}, {Pan,Café,Leche}, {Pan,Agua}, {Leche} $\equiv \{1245\}, \{1234\}, \{124\}, \{134\}, \{12\}, \{4\}$

Para la dimensión 3 tenemos:



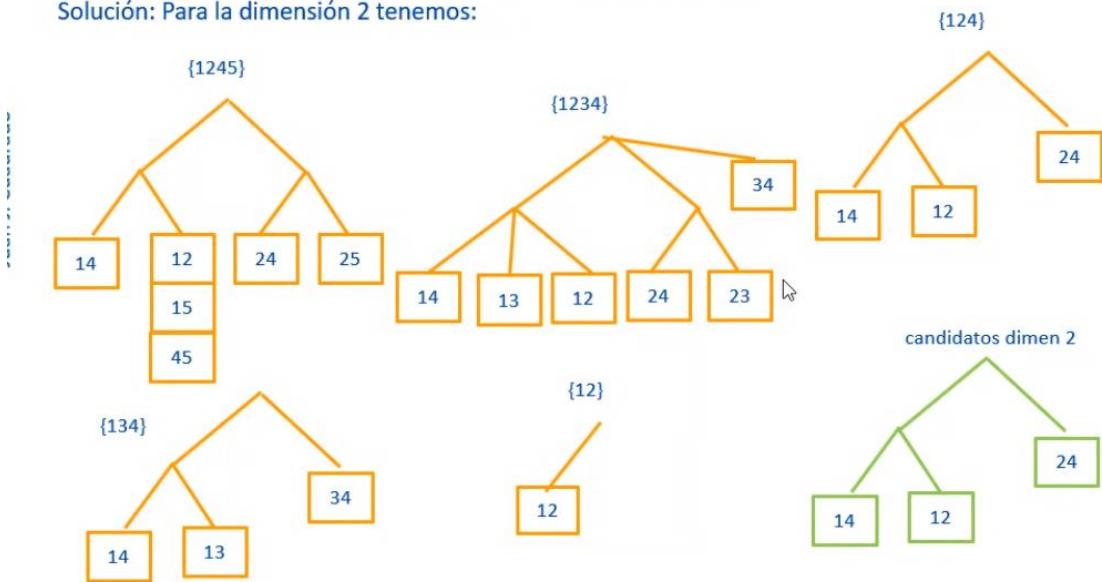
Asociación

Algoritmo Apriori

Paso A. Subpaso A.2.2.3.

Ejercicio: Comparación de ambas particiones. Todas las hojas coincidentes incrementan en una unidad el numerador en el cálculo del soporte del suceso candidato.

Solución: Para la dimensión 2 tenemos:



Asociación

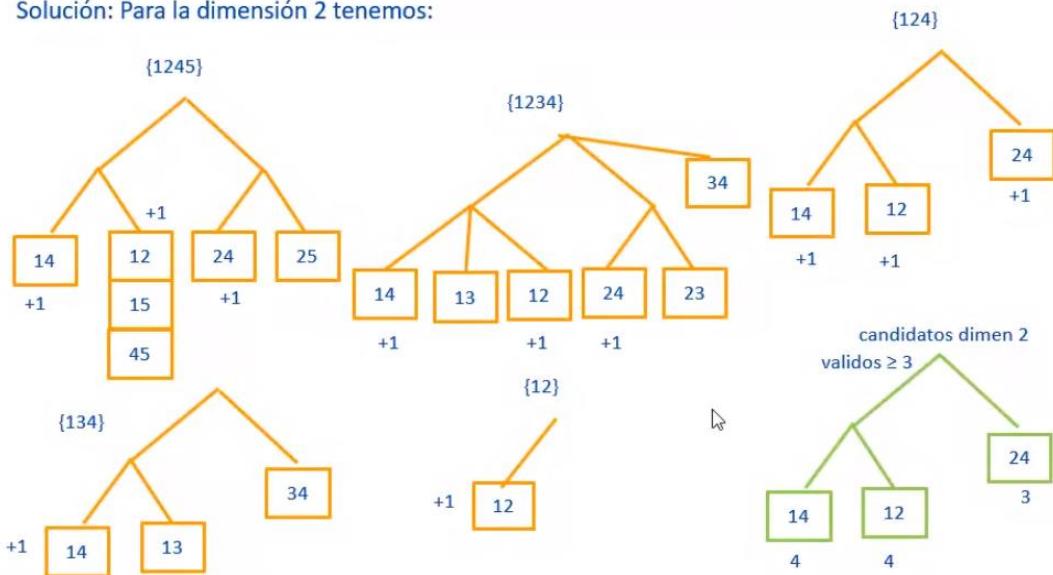
Algoritmo Apriori

Paso A. Subpaso A.2.2.3.

Ejercicio: Comparación de ambas particiones. Todas las hojas coincidentes incrementan en una unidad el numerador en el cálculo del soporte del suceso candidato.

Solución: Para la dimensión 2 tenemos:

Juan J. Cuadrado



Asociación

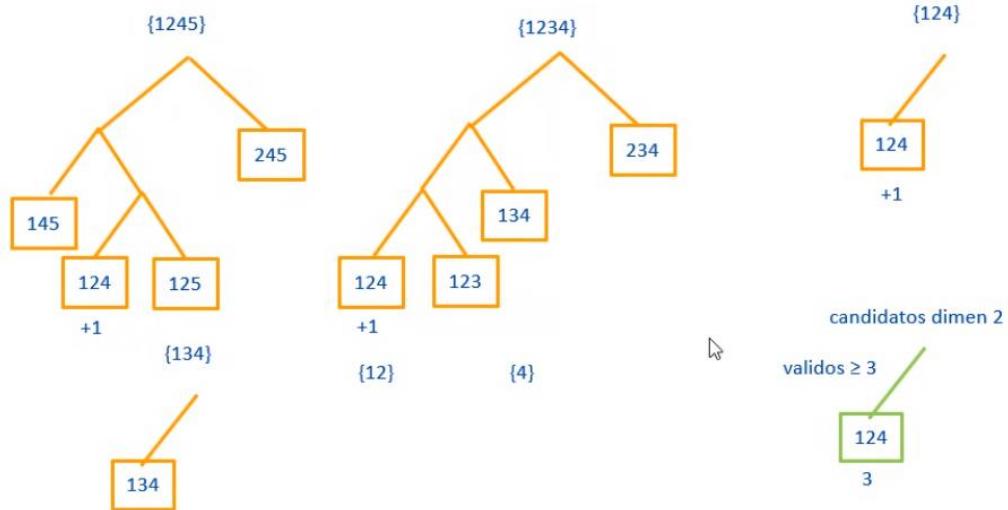
Algoritmo Apriori

Paso A. Subpaso A.2.2.3.

Ejercicio: Comparación de ambas particiones. Todas las hojas coincidentes incrementan en una unidad el numerador en el cálculo del soporte del suceso candidato.

Solución: Para la dimensión 2 tenemos:

Juan J. Cuadrado



Asociación

Algoritmo Apriori

Paso B: Se aplica únicamente a los sucesos que han superado el umbral de soporte. Se fija un umbral de confianza y se identifican las asociaciones que lo alcanzan o superan y su sentido.

Para cada suceso seleccionado de dimensión k habrá $2^k - 2$ posibles reglas de asociación.

Ejercicio: ¿Sobre qué asociaciones se aplicará el Paso B?

Solución: Los sucesos con soporte son: $\{\{12\}, \{14\}, \{24\}, \{124\}\}$ o lo que es lo mismo $\{\{\text{Pan, Agua}\}, \{\text{Pan, Leche}\}, \{\text{Agua, Leche}\}, \{\text{Pan, Agua, Leche}\}\}$.

Las asociaciones que se pueden establecer son:

- Dimensión 2: ($2^2 - 2 = 2$, para cada conjunto): $\{\text{Pan}\} \rightarrow \{\text{Agua}\}$, $\{\text{Agua}\} \rightarrow \{\text{Pan}\}$, $\{\text{Pan}\} \rightarrow \{\text{Leche}\}$, $\{\text{Leche}\} \rightarrow \{\text{Pan}\}$, $\{\text{Agua}\} \rightarrow \{\text{Leche}\}$, $\{\text{Leche}\} \rightarrow \{\text{Agua}\}$.
- Dimensión 3: ($2^3 - 2 = 6$): $\{\text{Pan, Agua}\} \rightarrow \{\text{Leche}\}$, $\{\text{Leche}\} \rightarrow \{\text{Pan, Agua}\}$, $\{\text{Pan, Leche}\} \rightarrow \{\text{Agua}\}$, $\{\text{Agua}\} \rightarrow \{\text{Pan, Leche}\}$, $\{\text{Agua, Leche}\} \rightarrow \{\text{Pan}\}$, $\{\text{Pan}\} \rightarrow \{\text{Agua, Leche}\}$.



Asociación

Algoritmo Apriori

Paso B: Se aplica únicamente a los sucesos que han superado el umbral de soporte. Se fija un umbral de confianza y se identifican las asociaciones que lo alcanzan o superan y su sentido.

Para cada suceso seleccionado de dimensión k habrá $2^k - 2$ posibles reglas de asociación.

Ejercicio: ¿Sobre qué asociaciones se aplicará el Paso B?

Solución: Los sucesos con soporte son: $\{\{12\}, \{14\}, \{24\}, \{124\}\}$ o lo que es lo mismo $\{\{\text{Pan, Agua}\}, \{\text{Pan, Leche}\}, \{\text{Agua, Leche}\}, \{\text{Pan, Agua, Leche}\}\}$.

Las asociaciones que se pueden establecer son:

- Dimensión 2: ($2^2 - 2 = 2$, para cada conjunto): $\{\text{Pan}\} \rightarrow \{\text{Agua}\}$, $\{\text{Agua}\} \rightarrow \{\text{Pan}\}$, $\{\text{Pan}\} \rightarrow \{\text{Leche}\}$, $\{\text{Leche}\} \rightarrow \{\text{Pan}\}$, $\{\text{Agua}\} \rightarrow \{\text{Leche}\}$, $\{\text{Leche}\} \rightarrow \{\text{Agua}\}$.
- Dimensión 3: ($2^3 - 2 = 6$): $\{\text{Pan, Agua}\} \rightarrow \{\text{Leche}\}$, $\{\text{Leche}\} \rightarrow \{\text{Pan, Agua}\}$, $\{\text{Pan, Leche}\} \rightarrow \{\text{Agua}\}$, $\{\text{Agua}\} \rightarrow \{\text{Pan, Leche}\}$, $\{\text{Agua, Leche}\} \rightarrow \{\text{Pan}\}$, $\{\text{Pan}\} \rightarrow \{\text{Agua, Leche}\}$.



Asociación

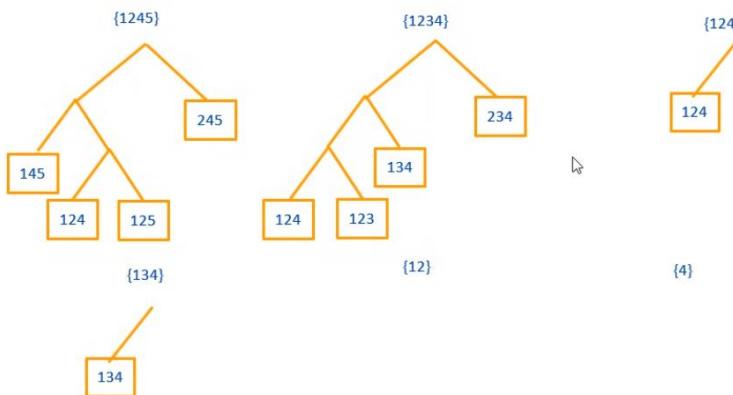
Algoritmo Apriori

Paso A. Subpaso A.2.2.2.

Ejercicio: Realizar una partición de los sucesos de la muestra utilizando el mismo hash tree.

Solución: {Pan, Agua, Leche, Naranjas}, {Pan, Agua, Café, Leche}, {Pan, Agua, Leche}, {Pan, Café, Leche}, {Pan, Agua}, {Leche} $\equiv \{1245\}, \{1234\}, \{124\}, \{134\}, \{12\}, \{4\}$

Para la dimensión 3 tenemos:



Asociación

Algoritmo Apriori

Paso B:

Ejercicio: Aplicar la función ap-genrules para establecer asociaciones con un umbral de confianza del 80%.

Solución: ap-genrules está basada en el teorema: Sean dos conjuntos A y B, si la asociación $A \rightarrow B - A$ no supera el umbral de confianza, entonces cualquier asociación $A' \rightarrow B - A'$, donde A' es cualquier subconjunto de A ($A' \subseteq A$) tampoco lo alcanzará.

Tomamos el conjunto de mayor tamaño y le llamamos B, B={Pan, Agua, Leche}.

Tomamos A={Pan, Agua}. Aplicamos: $A \rightarrow B - A$. Donde B-A={Leche}. Por lo tanto:

$$\{\text{Pan, Agua}\} \rightarrow \{\text{Leche}\}. \text{ Confianza: } c = \frac{n_B}{n_A} = \frac{3}{4} = 0.75 < 0.8$$



Asociación

Algoritmo Apriori

Paso B:

Ejercicio: Aplicar la función ap-genrules para establecer asociaciones con un umbral de confianza del 80%.

Solución: ap-genrules está basada en el teorema: Sean dos conjuntos A y B, si la asociación $A \rightarrow B - A$ no supera el umbral de confianza, entonces cualquier asociación $A' \rightarrow B - A'$, donde A' es cualquier subconjunto de A ($A' \subseteq A$) tampoco lo alcanzará.

Tomamos el conjunto de mayor tamaño y le llamamos B, B={Pan, Agua, Leche}.

Tomamos A={Pan, Agua}. Aplicamos: $A \rightarrow B - A$. Donde $B-A=\{\text{Leche}\}$. Por lo tanto:

$$\{\text{Pan, Agua}\} \rightarrow \{\text{Leche}\}. \text{Confianza: } c = \frac{n_B}{n_A} = \frac{3}{4} = 0.75 < 0.8$$

Como no supera el umbral de confianza $A' \rightarrow B - A'$ tampoco lo alcanza.

A' son los subconjuntos de A: $A'_1 = \{\text{Pan}\}$, $A'_2 = \{\text{Agua}\}$

Por lo tanto $\{\text{Agua}\} \rightarrow \{\text{Pan, Leche}\}$ y $\{\text{Pan}\} \rightarrow \{\text{Agua, Leche}\}$ no lo alcanzan. $\{\text{Pan, Leche}\}$ y $\{\text{Agua, Leche}\}$ son $B - A'$.

Asociación

Algoritmo Apriori

Paso B:

Ejercicio: Aplicar la función ap-genrules para establecer asociaciones con un umbral de confianza del 80%.

Solución: ap-genrules está basada en el teorema: Sean dos conjuntos A y B, si la asociación $A \rightarrow B - A$ no supera el umbral de confianza, entonces cualquier asociación $A' \rightarrow B - A'$, donde A' es cualquier subconjunto de A ($A' \subseteq A$) tampoco lo alcanzará.

Tomamos el conjunto de mayor tamaño y le llamamos B, B={Pan, Agua, Leche}.

Tomamos A={Pan, Agua}. Aplicamos: $A \rightarrow B - A$. Donde $B-A=\{\text{Leche}\}$. Por lo tanto:

$$\{\text{Pan, Agua}\} \rightarrow \{\text{Leche}\}. \text{Confianza: } c = \frac{n_B}{n_A} = \frac{3}{4} = 0.75 < 0.8$$

Como no supera el umbral de confianza $A' \rightarrow B - A'$ tampoco lo alcanza.

A' son los subconjuntos de A: $A'_1 = \{\text{Pan}\}$, $A'_2 = \{\text{Agua}\}$

Por lo tanto $\{\text{Agua}\} \rightarrow \{\text{Pan, Leche}\}$ y $\{\text{Pan}\} \rightarrow \{\text{Agua, Leche}\}$ no lo alcanzan. $\{\text{Pan, Leche}\}$ y $\{\text{Agua, Leche}\}$ son $B - A'$.

Lo comprobamos:

$$cA'_1 = \frac{n_B}{n_{A'_1}} = \frac{3}{5} = 0.6 < 0.8 \quad cA'_2 = \frac{n_B}{n_{A'_2}} = \frac{3}{4} = 0.75 < 0.8$$

Asociación

Algoritmo Apriori

Paso B:

Ejercicio: Aplicar la función ap-genrules para establecer asociaciones con un umbral de confianza del 80%.

Solución: Por último tomamos A={Agua, Leche}.

Aplicamos: $A \rightarrow B - A$. Donde $B-A=\{\text{Pan}\}$. Por lo tanto:

$$\{\text{Agua, Leche}\} \rightarrow \{\text{Pan}\}. \text{ Calculamos su confianza: } c = \frac{n_B}{n_A} = \frac{3}{3} = 1 > 0.8$$

Sí supera el umbral de confianza, así que además de aceptar la asociación no podemos aplicar la segunda premisa y hay que analizar $A' \rightarrow B - A'$ uno a uno. A' son los subconjuntos de A: $A'_1 = \{\text{Agua}\}$, $A'_2 = \{\text{Leche}\}$



Asociación

Algoritmo Apriori

Paso B:

Ejercicio: Aplicar la función ap-genrules para establecer asociaciones con un umbral de confianza del 80%.

Solución: Tomamos ahora A={Pan, Leche}.

Aplicamos: $A \rightarrow B - A$. Donde $B-A=\{\text{Agua}\}$. Por lo tanto:

$$\{\text{Pan, Leche}\} \rightarrow \{\text{Agua}\}. \text{ Calculamos su confianza: } c = \frac{n_B}{n_A} = \frac{3}{4} = 0.75 < 0.8$$

Como no supera el umbral de confianza $A' \rightarrow B - A'$ tampoco lo alcanza. A' son los subconjuntos de A: $A'_1 = \{\text{Pan}\}$, $A'_2 = \{\text{Leche}\}$

Por lo tanto $\{\text{Pan}\} \rightarrow \{\text{Agua, Leche}\}$ y $\{\text{Leche}\} \rightarrow \{\text{Agua, Pan}\}$ no lo alcanzan. $\{\text{Pan, Agua}\}$ y $\{\text{Agua, Leche}\}$ son $B - A'$.



Lo comprobamos:

$$cA'_1 = 0.6 < 0.8 \text{ es igual que en caso anterior} \quad cA'_2 = \frac{n_B}{n_{A'_2}} = \frac{3}{5} = 0.6 < 0.8$$

Asociación

Algoritmo Apriori

Paso B:

Ejercicio: Aplicar la función ap-genrules para establecer asociaciones con un umbral de confianza del 80%.

Solución: Por último tomamos A={Agua, Leche}.

Aplicamos: $A \rightarrow B - A$. Donde $B-A=\{\text{Pan}\}$. Por lo tanto:

$$\{\text{Agua, Leche}\} \rightarrow \{\text{Pan}\}. \text{ Calculamos su confianza: } c = \frac{n_B}{n_A} = \frac{3}{3} = 1 > 0.8$$

Sí supera el umbral de confianza, así que además de aceptar la asociación no podemos aplicar la segunda premisa y hay que analizar $A' \rightarrow B - A'$ uno a uno. A' son los subconjuntos de A: $A'_1 = \{\text{Agua}\}$, $A'_2 = \{\text{Leche}\}$

Por lo tanto hay que analizar $\{\text{Agua}\} \rightarrow \{\text{Pan, Leche}\}$ y $\{\text{Leche}\} \rightarrow \{\text{Agua, Pan}\}$. $\{\text{Pan, Leche}\}$ y $\{\text{Agua, Pan}\}$ son $B - A'$.

Lo comprobamos:

$$cA'_1 \text{ es igual que en caso anterior} \quad cA'_2 \text{ es igual que en caso anterior}$$

Asociación

Algoritmo Apriori

Paso B:

Ejercicio: Aplicar la función ap-genrules para establecer asociaciones con un umbral de confianza del 80%.

Solución: Tomamos ahora los conjuntos de dimensión 2.

Comenzamos por $B=\{\text{Pan, Agua}\}$.

Tomamos $A=\{\text{Pan}\}$

Aplicamos la primera premisa: $A \rightarrow B - A$. Donde $B-A=\{\text{Agua}\}$. Por lo tanto tenemos:

$$\{\text{Pan}\} \rightarrow \{\text{Agua}\}. \text{ Calculamos su confianza: } c = \frac{n_B}{n_A} = \frac{4}{5} = 0.8 = 0.8$$



Asociación

Algoritmo Apriori

Paso B:

Ejercicio: Aplicar la función ap-genrules para establecer asociaciones con un umbral de confianza del 80%.

Solución: Tomamos ahora los conjuntos de dimensión 2.

Comenzamos por $B = \{\text{Pan}, \text{Agua}\}$.

Tomamos $A = \{\text{Pan}\}$

Aplicamos la primera premisa: $A \rightarrow B - A$. Donde $B-A = \{\text{Agua}\}$. Por lo tanto tenemos:

$$\{\text{Pan}\} \rightarrow \{\text{Agua}\}. \text{ Calculamos su confianza: } c = \frac{n_B}{n_A} = \frac{4}{5} = 0.8 = 0.8$$

En este caso **si supera el umbral de confianza**, así que además de aceptar la asociación no podemos aplicar la segunda premisa y hay que analizar $A' \rightarrow B - A'$ uno a uno. Pero nos damos cuenta que en esta dimensión no hay A' porque no hay subconjuntos de A , por lo que hay que analizar todos los casos.

Asociación

Algoritmo Apriori

Paso B:

Ejercicio: Aplicar la función ap-genrules para establecer asociaciones con un umbral de confianza del 80%.

Solución: Tomamos ahora los conjuntos de dimensión 2.

$$\{\text{Agua}\} \rightarrow \{\text{Pan}\}: c = \frac{n_B}{n_A} = \frac{4}{4} = 1 > 0.8 \quad \text{Sí}$$

$$\{\text{Pan}\} \rightarrow \{\text{Leche}\}: c = \frac{n_B}{n_A} = \frac{4}{5} = 0.8 = 0.8 \quad \text{Sí}$$

$$\{\text{Leche}\} \rightarrow \{\text{Pan}\}: c = \frac{n_B}{n_A} = \frac{4}{5} = 0.8 = 0.8 \quad \text{Sí}$$

$$\{\text{Agua}\} \rightarrow \{\text{Leche}\}: c = \frac{n_B}{n_A} = \frac{3}{4} = 0.75 < 0.8 \quad \text{No}$$

$$\{\text{Leche}\} \rightarrow \{\text{Agua}\}: c = \frac{n_B}{n_A} = \frac{3}{5} = 0.6 < 0.8 \quad \text{No}$$

Asociación

Algoritmo Apriori

Paso B:

Ejercicio: Aplicar la función ap-genrules para establecer asociaciones con un umbral de confianza del 80%.

Solución: Por lo tanto las asociaciones seleccionadas por presentan un 50% de soporte y un 80% de confianza son:

{Agua, Leche} → {Pan}

{Pan} → {Agua}

{Agua} → {Pan}

{Pan} → {Leche}

{Leche} → {Pan}



Clasificación Supervisada

Los estudios de clasificación buscan **definir un modelo de clasificación** que permita obtener el valor desconocido de un suceso elemental a partir del valor resto de sucesos elementales. Para poder definir el modelo es **imprescindible disponer de una muestra de sucesos**.

Se denominan **supervisados** aquellos en los que **se conoce el conjunto de posibles valores que podrá tener el valor desconocido**.

Las técnicas más utilizadas son:

1. Árboles de decisión
2. Reglas de decisión
3. Random Forest
4. Máquinas de vectores de soporte (SVM)
5. Naïve Bayes
6. Redes neuronales
7. Regresión

Clasificación Supervisada

Ejercicio: Se tiene un conjunto de calificaciones académicas formadas por cuatro notas: Teoría, Laboratorio, Prácticas y Calificación Global. Las calificaciones de Teoría, Laboratorio, Prácticas, tendrán los valores A, B, C y D, donde A será la mayor calificación posible y D la menor. La calificación global tendrá dos valores, Aprobado, Ap, y Suspenso, Ss. Establecer los sucesos elementales, definir las clases de equivalencia de un suceso clasificador y explicar que haría la función clasificadora.

Solución:

$E = \langle \text{Teoría, Laboratorio, Prácticas, Calificación Global} \rangle$

Clasificador = Calificación Global, que tiene dos clases de equivalencia complementarias y disjuntas: Aprobados y Suspensos

La función clasificadora en función de los valores A, B, C y D de los primeros tres sucesos definirá el valor del cuarto, Aprobado o Suspensos.

Clasificación Supervisada

Arboles de Decisión

El **Algoritmo de Hunt** sigue un proceso genérico de 1 a k pasos, siendo k el número de sucesos elementales:

- A. Selección de un suceso elemental para el nodo inicial o raíz en el primer paso, o para un nodo intermedio en el resto de los pasos. No puede ser el clasificador.
- B. Clasificación de los sucesos. Si se pueden clasificar completamente los sucesos se trata de un nodo final y se termina la clasificación. Si no lo permite se trata de un nodo interno y se vuelve al paso A.

Clasificación Supervisada

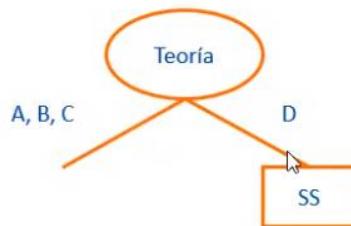
Ejercicio: A partir de la muestra de sucesos que permitirá buscar la definición de la función de clasificación está formada por los siguientes nueve sucesos: 1. {A, A, B, Ap}; 2. {A, B, D, Ss}; 3. {D, C, C, Ss}; 4. {D, D, A, Ss}; 5. {B, C, B, Ss}; 6. {C, B, B, Ap}; 7. {B, B, A, Ap}; 8. {C, D, C, Ss}; 9. {B, A, C, Ss}. Aplicar el algoritmo de Hunt para la construcción de un árbol de decisión, sin tener en cuenta la ganancia de información. {Teoría, Laboratorio, Prácticas, Calificación Global}

Clasificación Supervisada

Ejercicio: A partir de la muestra de sucesos que permitirá buscar la definición de la función de clasificación está formada por los siguientes nueve sucesos: 1. {A, A, B, Ap}; 2. {A, B, D, Ss}; 3. {D, C, C, Ss}; 4. {D, D, A, Ss}; 5. {B, C, B, Ss}; 6. {C, B, B, Ap}; 7. {B, B, A, Ap}; 8. {C, D, C, Ss}; 9. {B, A, C, Ss}. Aplicar el algoritmo de Hunt para la construcción de un árbol de decisión, sin tener en cuenta la ganancia de información. {Teoría, Laboratorio, Prácticas, Calificación Global}

Solución:

Como Clasificación Global es el clasificador no lo podemos tomar como nodo inicial, así que tomamos cualquiera de los otros tres. Tomamos teoría y clasificamos:



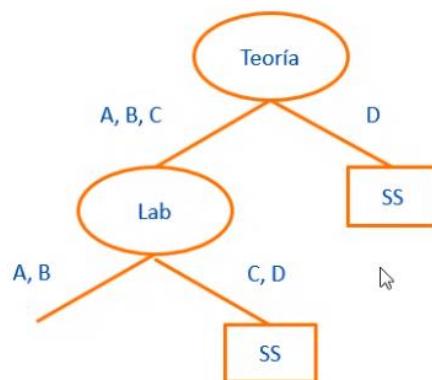
Teoría no es un nodo final porque si tiene una D es suspenso pero si tiene una A, B o C no se sabe.

Clasificación Supervisada

Ejercicio: 1. {A, A, B, Ap}; 2. {A, B, D, Ss}; 3. {D, C, C, Ss}; 4. {D, D, A, Ss}; 5. {B, C, B, Ss}; 6. {C, B, B, Ap}; 7. {B, B, A, Ap}; 8. {C, D, C, Ss}; 9. {B, A, C, Ss}. {Teoría, Laboratorio, Prácticas, Calificación Global}

Solución:

Tomamos Laboratorio como el siguiente nodo:

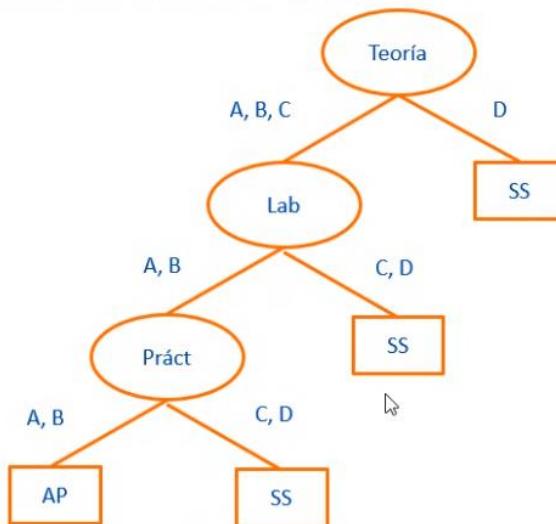


Clasificación Supervisada

Ejercicio: 1. {A, A, B, Ap}; 2. {A, B, D, Ss}; 3. {D, C, C, Ss}; 4. {D, D, A, Ss}; 5. {B, C, B, Ss}; 6. {C, B, B, Ap}; 7. {B, B, A, Ap}; 8. {C, D, C, Ss}; 9. {B, A, C, Ss}. {Teoría, Laboratorio, Prácticas, Calif Global}

Solución:

Por último tomamos Prácticas en el siguiente nodo:



Prácticas si es nodo final porque si tiene una C o D es suspenso y si tiene una A o B es aprobado. El modelo permite clasificar sucesos sin saber la calificación global.

Clasificación Supervisada

Arboles de Decisión

La Optimización de los árboles se realiza a través de la Ganancia de Información:

La ganancia de información de cada división mide la diferencia de impureza entre el nodo padre y los nodos hijos una vez realizada una división.

Cuanta mayor ganancia de información se tenga mejor será la división realizada.

La ecuación de cálculo es: $\Delta_I = I_{padre} - \sum_{j=1}^k \frac{N(n_j)}{N} I(n_j)$

- I_{padre} es la impureza del nodo padre
- $I(n_j)$ es la impureza del nodo hijo j
- $N(n_j)$ es el número de sucesos asociados con el nodo hijo j
- N es el número total de sucesos en el nodo padre

Clasificación Supervisada

Arboles de Decisión

La Impureza de un nodo se puede medir a través de diferentes unidades de medida:

- Entropía: $Ent(nodo) = -\sum_{i=0}^c f_{i \text{ nodo}} \log_2(f_{i \text{ nodo}})$
- Error: $Err(nodo) = 1 - \max_i(f_{i \text{ nodo}})$
- Gini: $Gin(nodo) = 1 - \sum_{i=0}^{c-1} (f_{i \text{ nodo}})^2$

$f_{i \text{ nodo}}$: frecuencia relativa de la clase de equivalencia i en el nodo analizado.

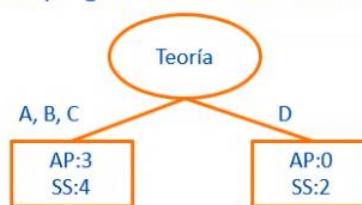
c: número de clases de equivalencia.

Clasificación Supervisada

Ejercicio: A partir de la muestra de calificaciones: 1. {A, A, B, Ap}; 2. {A, B, D, Ss}; 3. {D, D, C, Ss}; 4. {D, B, A, Ss}; 5. {B, C, B, Ss}; 6. {C, B, B, Ap}; 7. {B, B, A, Ap}; 8. {C, D, C, Ss}; 9. {B, A, C, Ss}. {Teoría, Laboratorio, Prácticas, Calificación Global}. Realizar el paso A del algoritmo de Hunt para la construcción de un árbol de decisión, optimizando la elección del suceso elemental elegido para el nodo inicial. Utilizando los tres métodos de medición de la impureza para el suceso elemental teoría y el Gini de Laboratorio y Prácticas y la ganancia de información de los tres.

Solución:

Comenzamos por Teoría:



Calculamos la ganancia de información basada en la Entropía:

$$Np: Ent(p) = -\sum_{i=0}^1 f_{ip} \log_2 f_{ip} = -\left(\frac{3}{9}\right) \log_2 \left(\frac{3}{9}\right) - \left(\frac{6}{9}\right) \log_2 \left(\frac{6}{9}\right) = 0,904, 3 \text{ AP y } 6 \text{ SS}$$

$$N1: Ent(1) = -\sum_{i=0}^1 f_{ip} \log_2 f_{i1} = -\left(\frac{3}{7}\right) \log_2 \left(\frac{3}{7}\right) - \left(\frac{4}{7}\right) \log_2 \left(\frac{4}{7}\right) = 0,98$$

$$N2: Ent(2) = -\sum_{i=0}^1 f_{ip} \log_2 f_{i2} = -\left(\frac{0}{2}\right) \log_2 \left(\frac{0}{2}\right) - \left(\frac{2}{2}\right) \log_2 \left(\frac{2}{2}\right) = 0$$

$$\Delta_I = I_{padre} - \sum_{j=1}^k \frac{N(n_j)}{N} I(n_j) = 0,904 - \left(\left(\frac{7}{9}\right) \cdot 0,98 + \left(\frac{2}{9}\right) \cdot 0 \right) = 0,904 - 0,77 = 0,13$$

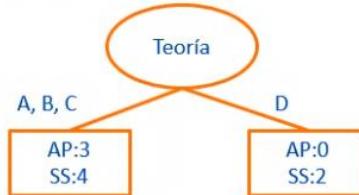


Clasificación Supervisada

Ejercicio: A partir de la muestra de calificaciones: 1. {A, A, B, Ap}; 2. {A, B, D, Ss}; 3. {D, D, C, Ss}; 4. {D, B, A, Ss}; 5. {B, C, B, Ss}; 6. {C, B, B, Ap}; 7. {B, B, A, Ap}; 8. {C, D, C, Ss}; 9. {B, A, C, Ss}. {Teoría, Laboratorio, Prácticas, Calificación Global}. Realizar el paso A del algoritmo de Hunt para la construcción de un árbol de decisión, optimizando la elección del suceso elemental elegido para el nodo inicial. Utilizando los tres métodos de medición de la impureza para el suceso elemental teoría y el Gini de Laboratorio y Prácticas y la ganancia de información de los tres.

Solución:

Comenzamos por Teoría:



Calculamos la ganancia de información basada en el **Error**:

$$Np: Err(p) = 1 - \max_i(f_{ip}) = 1 - \max_i\left(\frac{3}{9}, \frac{6}{9}\right) = 1 - 0,67 = 0,33, 3 \text{ AP y } 6 \text{ SS}$$

$$N1: Err(1) = 1 - \max_i(f_{i1}) = 1 - \max_i\left(\frac{3}{7}, \frac{4}{7}\right) = 0,43$$

$$N2: Err(2) = 1 - \max_i(f_{i2}) = 1 - \max_i\left(\frac{0}{2}, \frac{2}{2}\right) = 0$$

$$\Delta_I = I_{padre} - \sum_{j=1}^k \frac{N(n_j)}{N} I(n_j) = 0,33 - \left(\left(\frac{7}{9}\right) \cdot 0,43 + \left(\frac{2}{9}\right) \cdot 0 \right) = 0,33 - 0,33 = 0$$

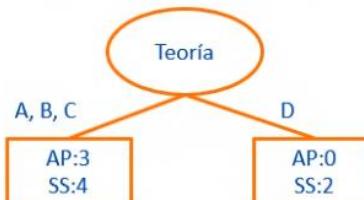
▷

Clasificación Supervisada

Ejercicio: A partir de la muestra de calificaciones: 1. {A, A, B, Ap}; 2. {A, B, D, Ss}; 3. {D, D, C, Ss}; 4. {D, B, A, Ss}; 5. {B, C, B, Ss}; 6. {C, B, B, Ap}; 7. {B, B, A, Ap}; 8. {C, D, C, Ss}; 9. {B, A, C, Ss}. {Teoría, Laboratorio, Prácticas, Calificación Global}. Realizar el paso A del algoritmo de Hunt para la construcción de un árbol de decisión, optimizando la elección del suceso elemental elegido para el nodo inicial. Utilizando los tres métodos de medición de la impureza para el suceso elemental teoría y el Gini de Laboratorio y Prácticas y la ganancia de información de los tres.

Solución:

Comenzamos por Teoría:



Calculamos la ganancia de información basada en el **Gini**:

$$Np: Gin(p) = 1 - \sum_{i=0}^{c-1} f_{ip}^2 = 1 - \left(\left(\frac{3}{9}\right)^2 + \left(\frac{6}{9}\right)^2 \right) = 1 - 0,55 = 0,45, 3 \text{ AP y } 6 \text{ SS}$$

$$N1: Gin(1) = 1 - \sum_{i=0}^{c-1} f_{i1}^2 = 1 - \left(\left(\frac{3}{7}\right)^2 + \left(\frac{4}{7}\right)^2 \right) = 0,49$$

$$N2: Gin(2) = 1 - \sum_{i=0}^{c-1} f_{i2}^2 = 1 - \left(\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2 \right) = 0$$

$$\Delta_I = I_{padre} - \sum_{j=1}^k \frac{N(n_j)}{N} I(n_j) = 0,45 - \left(\left(\frac{7}{9}\right) \cdot 0,49 + \left(\frac{2}{9}\right) \cdot 0 \right) = 0,45 - 0,38 = 0,07$$

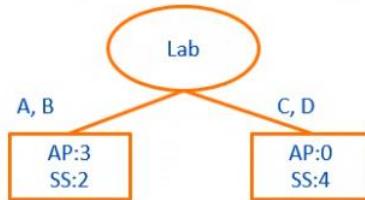
▷

Clasificación Supervisada

Ejercicio: A partir de la muestra de calificaciones: 1. {A, A, B, Ap}; 2. {A, B, D, Ss}; 3. {D, D, C, Ss}; 4. {D, B, A, Ss}; 5. {B, C, B, Ss}; 6. {C, B, B, Ap}; 7. {B, B, A, Ap}; 8. {C, D, C, Ss}; 9. {B, A, C, Ss}. {Teoría, Laboratorio, Prácticas, Calificación Global}. Realizar el paso A del algoritmo de Hunt para la construcción de un árbol de decisión, optimizando la elección del suceso elemental elegido para el nodo inicial. Utilizando los tres métodos de medición de la impureza para el suceso elemental teoría y el Gini de Laboratorio y Prácticas y la ganancia de información de los tres.

Solución:

Calculamos para Lab:



Calculamos la ganancia de información basada en el **Gini**:

$$Np: \text{Gin}(p) = 1 - \sum_{i=0}^{c-1} f_{ip}^2 = 1 - \left(\left(\frac{3}{9}\right)^2 + \left(\frac{6}{9}\right)^2 \right) = 1 - 0,55 = 0,45, 3 \text{ AP y } 6 \text{ SS}$$

$$N1: \text{Gin}(1) = 1 - \sum_{i=0}^{c-1} f_{i1}^2 = 1 - \left(\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2 \right) = 0,48$$

$$N2: \text{Gin}(2) = 1 - \sum_{i=0}^{c-1} f_{i2}^2 = 1 - \left(\left(\frac{0}{4}\right)^2 + \left(\frac{4}{4}\right)^2 \right) = 0$$

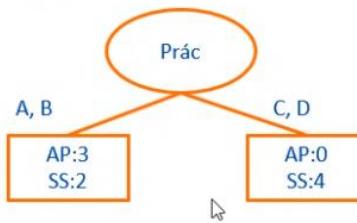
$$\Delta_I = I_{\text{padre}} - \sum_{j=1}^k \frac{N(n_j)}{N} I(n_j) = 0,45 - \left(\left(\frac{5}{9}\right) \cdot 0,48 + \left(\frac{4}{9}\right) \cdot 0 \right) = 0,45 - 0,27 = 0,18 > 0,07$$

Clasificación Supervisada

Ejercicio: A partir de la muestra de calificaciones: 1. {A, A, B, Ap}; 2. {A, B, D, Ss}; 3. {D, D, C, Ss}; 4. {D, B, A, Ss}; 5. {B, C, B, Ss}; 6. {C, B, B, Ap}; 7. {B, B, A, Ap}; 8. {C, D, C, Ss}; 9. {B, A, C, Ss}. {Teoría, Laboratorio, Prácticas, Calificación Global}. Realizar el paso A del algoritmo de Hunt para la construcción de un árbol de decisión, optimizando la elección del suceso elemental elegido para el nodo inicial. Utilizando los tres métodos de medición de la impureza para el suceso elemental teoría y el Gini de Laboratorio y Prácticas y la ganancia de información de los tres.

Solución:

Calculamos para Prácticas:



Calculamos la ganancia de información basada en el **Gini**:

$$Np: \text{Gin}(p) = 1 - \sum_{i=0}^{c-1} f_{ip}^2 = 1 - \left(\left(\frac{3}{9}\right)^2 + \left(\frac{6}{9}\right)^2 \right) = 1 - 0,55 = 0,45, 3 \text{ AP y } 6 \text{ SS}$$

$$N1: \text{Gin}(1) = 1 - \sum_{i=0}^{c-1} f_{i1}^2 = 1 - \left(\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2 \right) = 0,48$$

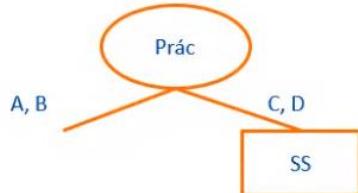
$$N2: \text{Gin}(2) = 1 - \sum_{i=0}^{c-1} f_{i2}^2 = 1 - \left(\left(\frac{0}{4}\right)^2 + \left(\frac{4}{4}\right)^2 \right) = 0$$

$$\Delta_I = I_{\text{padre}} - \sum_{j=1}^k \frac{N(n_j)}{N} I(n_j) = 0,45 - \left(\left(\frac{5}{9}\right) \cdot 0,48 + \left(\frac{4}{9}\right) \cdot 0 \right) = 0,45 - 0,27 = 0,18 > 0,07$$

Clasificación Supervisada

Ejercicio: A partir de la muestra de calificaciones: 1. {A, A, B, Ap}; 2. {A, B, D, Ss}; 3. {D, D, C, Ss}; 4. {D, B, A, Ss}; 5. {B, C, B, Ss}; 6. {C, B, B, Ap}; 7. {B, B, A, Ap}; 8. {C, D, C, Ss}; 9. {B, A, C, Ss}. {Teoría, Laboratorio, Prácticas, Calificación Global}. Realizar el paso A del algoritmo de Hunt para la construcción de un árbol de decisión, optimizando la elección del suceso elemental elegido para el nodo inicial. Utilizando los tres métodos de medición de la impureza para el suceso elemental teoría y el Gini de Laboratorio y Prácticas y la ganancia de información de los tres.

Solución:



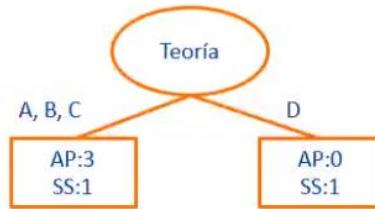
A partir de los cálculos de Gini para Teoría, Laboratorio y Prácticas, 0.07, 0.18, 0.18, respectivamente, vemos que la mayor ganancia de información se obtiene con Laboratorio o Prácticas, por lo que para el nodo inicial seleccionamos cualquiera de los dos, vamos a seleccionar prácticas.

Clasificación Supervisada

Ejercicio: A partir de la muestra de calificaciones: 1. {A, A, B, Ap}; 2. {A, B, D, Ss}; 3. {D, D, C, Ss}; 4. {D, B, A, Ss}; 5. {B, C, B, Ss}; 6. {C, B, B, Ap}; 7. {B, B, A, Ap}; 8. {C, D, C, Ss}; 9. {B, A, C, Ss}. {Teoría, Laboratorio, Prácticas, Calificación Global}. Realizar, en su caso, el paso B del algoritmo de H para la construcción de un árbol de decisión del ejercicio anterior utilizando la medida impureza Error.

Solución:

Comenzamos por Teoría:



Calculamos la ganancia de información basada en el Error:

$$\text{Np: } Err(p) = 1 - \max_i(f_{ip}) = 1 - \max_i\left(\frac{3}{5}, \frac{2}{5}\right) = 1 - 0,6 = 0,4, 3 \text{ AP y 2 SS}$$

$$\text{N1: } Err(1) = 1 - \max_i(f_{i1}) = 1 - \max_i\left(\frac{3}{4}, \frac{1}{4}\right) = 1 - 0,75 = 0,25$$

$$\text{N2: } Err(2) = 1 - \max_i(f_{i2}) = 1 - \max_i\left(\frac{0}{1}, \frac{1}{1}\right) = 0$$

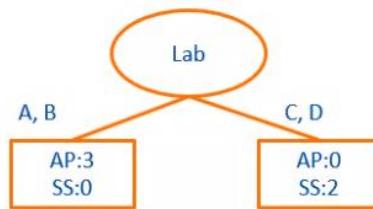
$$\Delta I = I_{\text{padre}} - \sum_{j=1}^k \frac{N(n_j)}{N} I(n_j) = 0,4 - \left(\left(\frac{4}{5}\right) \cdot 0,25 + \left(\frac{1}{5}\right) \cdot 0 \right) = 0,4 - 0,2 = 0,2$$

Clasificación Supervisada

Ejercicio: A partir de la muestra de calificaciones: 1. {A, A, B, Ap}; 2. {A, B, D, Ss}; 3. {D, D, C, Ss}; 4. {D, B, A, Ss}; 5. {B, C, B, Ss}; 6. {C, B, B, Ap}; 7. {B, B, A, Ap}; 8. {C, D, C, Ss}; 9. {B, A, C, Ss}. {Teoría, Laboratorio, Prácticas, Calificación Global}. Realizar, en su caso, el paso B del algoritmo de Hunt para la construcción de un árbol de decisión del ejercicio anterior utilizando la medida de impureza Error.

Solución:

Calculamos para Lab:



Calculamos la ganancia de información basada en el Error:

$$Np: Err(p) = 1 - \max_i(f_{ip}) = 1 - \max_i\left(\frac{3}{5}, \frac{2}{5}\right) = 1 - 0,6 = 0,4, 3 \text{ AP y } 2 \text{ SS}$$

$$N1: Err(1) = 1 - \max_i(f_{i1}) = 1 - \max_i\left(\frac{3}{3}, \frac{0}{3}\right) = 1 - 1 = 0$$

$$N2: Err(2) = 1 - \max_i(f_{i2}) = 1 - \max_i\left(\frac{0}{2}, \frac{2}{2}\right) = 0$$

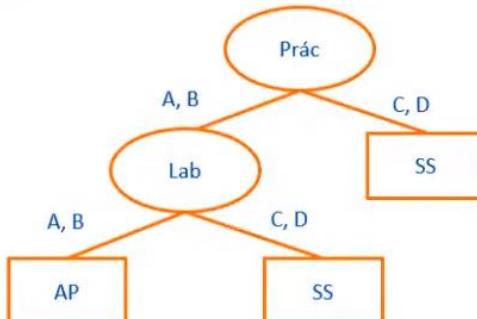
$$\Delta_I = I_{padre} - \sum_{j=1}^k \frac{N(n_j)}{N} I(n_j) = 0,4 - \left(\left(\frac{4}{5}\right) \cdot 0 + \left(\frac{1}{5}\right) \cdot 0 \right) = 0,4 - 0 = 0,4$$

Clasificación Supervisada

Ejercicio: A partir de la muestra de calificaciones: 1. {A, A, B, Ap}; 2. {A, B, D, Ss}; 3. {D, D, C, Ss}; 4. {D, B, A, Ss}; 5. {B, C, B, Ss}; 6. {C, B, B, Ap}; 7. {B, B, A, Ap}; 8. {C, D, C, Ss}; 9. {B, A, C, Ss}. {Teoría, Laboratorio, Prácticas, Calificación Global}. Realizar, en su caso, el paso B del algoritmo de Hunt para la construcción de un árbol de decisión del ejercicio anterior utilizando la medida de impureza Error.

Solución:

A partir de los cálculos de Error para Teoría y Laboratorio, 0.2 y 0.4, respectivamente, vemos que la mayor ganancia de información se obtiene con Laboratorio, por lo que para el nodo siguiente seleccionamos laboratorio.



Y ya no hace falta seguir porque el modelo ya permite clasificar los sucesos. Como puede verse se ha optimizado claramente el modelo anterior porque solo con dos nodos pueden clasificarse los sucesos.

Clasificación Supervisada

El análisis conjunto de dos características se realiza sobre los datos de ambas a la vez.

En el primer análisis de los pares de datos se cuantifica su **Frecuencia de aparición**.

- **Frecuencia Absoluta f_i :** Número de apariciones de un par de datos.
- **Frecuencia Relativa f_{ri} :** Frecuencia absoluta de un par de datos dividida entre el número de pares de datos.

Definiciones asociadas al concepto de frecuencia de dos variables:

- **Tabla de doble entrada:** Representa los valores de ambas variables y sus frecuencias. Si las variables son cualitativas se denomina **Tabla de Contingencia**.
- **Distribución marginal de una variable:** Frecuencias absolutas de los datos de la variable.

$$f(x_i) = \sum_{j=1}^m f(x_i, y_j) \quad f(y_j) = \sum_{i=1}^n f(x_i, y_j)$$

Clasificación Supervisada

Ejercicio: Obtener la frecuencia absoluta y relativa, la tabla de doble entrada y las distribuciones marginales de los siguientes pares de datos procedentes de los planetas Mercurio, Venus, Tierra y Marte, respectivamente:

Radio Ecuatorial, Densidad: (2, 5), (6, 5), (6, 5), (3, 3). Los datos se han redondeado.

Solución:

Numero de pares: (n=4)

Frecuencia Absoluta: (2, 5): 1 / (6, 5): 2 / (3, 3): 1

Frecuencia Relativa: (2, 5): 0.25 / (6, 5): 0.5 / (3, 3): 0.25

Tabla de doble entrada y distribuciones marginales:

		Radio					
		2	3	4	5	6	f_{den}
Densidad	2						
	3		1				1
	4						
	5	1				2	3
	6						
	f_{rad}	1	1			2	4



MEDIDAS DE DEPENDENCIA

Las medidas más utilizadas son:

- **Covarianza:** mide lo relacionado que están dos variables. **No está normalizada.**
- **Correlación:** mide la relación entre las dos variables. **Si está normalizada.**

Covarianza	Correlación [0,1]
$S_{xy} = \frac{(\sum_{i=0}^n x_i \cdot y_i)}{n} - (\bar{x} \cdot \bar{y})$	$r_{xy} = \frac{S_{xy}}{\sigma_x \cdot \sigma_y}$
<ul style="list-style-type: none">• $S_{xy} > 0 \rightarrow$ Dependencia directa positiva.• $S_{xy} = 0 \rightarrow$ No existe relación lineal.• $S_{xy} < 0 \rightarrow$ Dependencia inversa negativa.	<ul style="list-style-type: none">• $[0,1] \rightarrow$ relación ascendente.• $0 \rightarrow$ no hay correlación.• $[-1,0] \rightarrow$ relación descendente. <p>Se consideran fuertemente correlacionadas a partir de 0'8.</p>

EJEMPLO COVARIANZA

Calcular covarianza para los datos:

$$E = \{2,5\}, \{6,5\}, \{6,5\}, \{3,3\}.$$

$$\begin{aligned} S_{xy} &= \frac{(\sum_{i=0}^n x_i \cdot y_i)}{n} - (\bar{x} \cdot \bar{y}) \\ &= \frac{(2 \cdot 5) + (6 \cdot 5) + (6 \cdot 5) + (3 \cdot 3)}{4} - \left(\frac{2+6+6+3}{4} \right) \cdot \left(\frac{5+5+5+3}{4} \right) \\ &= \frac{10+30+30+9}{4} - 4'25 \cdot 4'5 = 19'75 - 19'125 = 0'625 \end{aligned}$$

Clasificación Supervisada

Las medidas de dependencia más utilizadas son la **Covarianza** y la **Correlación**.

Correlación:

$$r_{xy} = \frac{S_{xy}}{s_x s_y}$$

Los valores de r_{xy} están en el intervalo [-1 y 1]

- Si la correlación lineal es perfecta, es decir si los valores de x e y están sobre una recta el valor de r será -1 si tiene pendiente negativa y 1 si la tiene positiva.
- Si r es igual a 0 no hay dependencia lineal entre las variables, lo cual implica o bien que las variables son independientes, o bien que hay una dependencia no lineal entre las mismas.

Clasificación Supervisada

Las medidas de dependencia más utilizadas son la **Covarianza** y la **Correlación**.

Correlación:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Ejercicio: Calcular la correlación del Radio Ecuatorial y la Densidad: (2,4, 5,4), (6,1, 5,2), (6,4, 5,5), (3,4, 3,9). Los datos no se han redondeado.

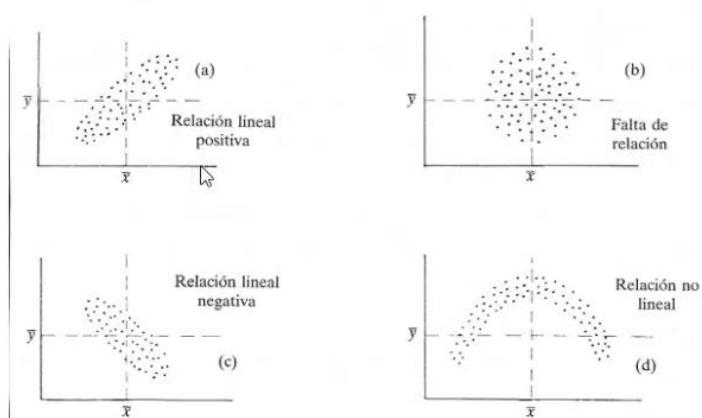
Solución:

$$s_{xy} = \frac{2,4 \cdot 5,4 + 6,1 \cdot 5,2 + 6,4 \cdot 5,5 + 3,4 \cdot 3,9}{4} - \left(\frac{2,4 + 6,1 + 6,4 + 3,4}{4} \right) \left(\frac{5,4 + 5,2 + 5,5 + 3,9}{4} \right) = 0,41$$
$$s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} = \sqrt{\frac{(2 - 4,575)^2 + \dots + (2 - 4,575)^2}{4}} = 1,715$$
$$s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}} = \sqrt{\frac{(5,4 - 5)^2 + \dots + (3,9 - 5)^2}{4}} = 0,6442$$
$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{0,41}{1,715 \cdot 0,6442} = 0,371$$

Clasificación Supervisada

El **análisis de regresión** trata de encontrar funciones que permitan establecer la relación entre dos o más variables.

El primer paso en el análisis de regresión consistirá en la visualización de los pares de datos mediante un gráfico denominado **diagramas de dispersión** o **nube de puntos**.



Clasificación Supervisada

El segundo paso en el análisis de regresión consistirá en la obtención de la **función de regresión**, que será aquella que esté más cerca de la mayoría de los pares de datos observados.

La función más utilizada es un relación lineal (recta) entre las variables, cuya ecuación es:

$$y = a + bx$$

El método de obtención o ajuste de la función más utilizado es el de **mínimos cuadrados**, que se fundamenta en minimizar la siguiente ecuación:

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Clasificación Supervisada

El segundo paso en el análisis de regresión consistirá en la obtención de la **función de regresión**, que será aquella que esté más cerca de la mayoría de los pares de datos observados.

La función más utilizada es un relación lineal (recta) entre las variables, cuya ecuación es:

$$y = a + bx$$

El método de obtención o ajuste de la función más utilizado es el de **mínimos cuadrados**, que se fundamenta en minimizar la siguiente ecuación:

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Los valores de b y a obtenidos a partir del método de mínimos cuadrados son:

$$b = \frac{s_{xy}}{s_x^2} \quad a = \bar{y} - b\bar{x}$$



Clasificación Supervisada

Ejercicio: Obtener una regresión lineal entre el Radio Ecuatorial y la Densidad de los planetas.
Datos: (2.4, 5.4), (6.1, 5.2), (6.4, 5.5), (3.4, 3.9).

Solución:

$$s_{xy} = 0,41; \quad s_x = 1,715; \quad \bar{x} = 4,575; \quad \bar{y} = 5$$

$$b = \frac{s_{xy}}{s_x^2} = \frac{0,41}{1,715^2} = 0,14$$

$$a = \bar{y} - b\bar{x} = 5 - 0,14 \cdot 4,575 = 4,36$$

$$y = 4,36 + 0,14x$$



Clasificación Supervisada

Una vez obtenida la regresión lineal (recta) debemos analizar en qué medida establece correctamente la relación entre las variables. El primer análisis que vamos a ver es el ANOVA. Para realizarlo calculamos:

- La dispersión, SSR , de los valores de y calculados a través de la función de regresión, \hat{y}_i :

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

- La dispersión, SSy , de los valores de y observados, y_i :

$$SSy = \sum_{i=1}^n (y_i - \bar{y})^2$$

- Y a partir de los dos valores anteriores la correlación cuadrada r^2 :

$$r^2 = \frac{SSR}{SSy}$$

Clasificación Supervisada

Una vez obtenida la regresión lineal (recta) debemos analizar en qué medida establece correctamente la relación entre las variables. El primer análisis que vamos a ver es el ANOVA. Para realizarlo calcularemos:

- La dispersión, SSR , de los valores de y calculados a través de la función de regresión, \hat{y}_i :
$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$
- La dispersión, SSy , de los valores de y observados, y_i :
$$SSy = \sum_{i=1}^n (y_i - \bar{y})^2$$
- A partir de los dos valores anteriores la correlación cuadrada r^2 :
$$r^2 = \frac{SSR}{SSy}$$

r^2 tendrá un valor que variará entre 0 y 1

Clasificación Supervisada

Ejercicio: Realizar un ANOVA para establecer la bonanza de la regresión lineal entre el Radio Ecuatorial y la Densidad de los planetas. Datos: (2.4, 5.4), (6.1, 5.2), (6.4, 5.5), (3.4, 3.9).

Solución:

La recta calculada es: $y = 4,36 + 0,14x$

A partir de la misma calculamos los valores de y para los correspondientes valores de x .

$$y(2,4) = 4,36 + 0,14 \cdot 2,4 = 4,696$$

$$y(6,1) = 4,36 + 0,14 \cdot 6,1 = 5,214$$

$$y(6,4) = 4,36 + 0,14 \cdot 6,4 = 5,256$$

$$y(3,4) = 4,36 + 0,14 \cdot 3,4 = 4,836$$

La media del valor de y , la conocemos y es $\bar{y} = 5$

Aplicamos las ecuaciones de cálculo:

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = (4,696 - 5)^2 + (5,214 - 5)^2 + (5,256 - 5)^2 + (4,836 - 5)^2 = 0.23$$

$$SSy = \sum_{i=1}^n (y_i - \bar{y})^2 = (5,4 - 5)^2 + (5,2 - 5)^2 + (5,5 - 5)^2 + (3,9 - 5)^2 = 1.66$$

$$r^2 = \frac{SSR}{SSy} = \frac{0,23}{1,66} = 0,14$$

Clasificación Supervisada

Ejercicio: Realizar un análisis de desviación típica residual para establecer la bonanza de la regresión lineal entre el Radio Ecuatorial y la Densidad de los planetas. Datos: (2.4, 5.4), (6.1, 5.2), (6.4, 5.5), (3.4, 3.9).

Solución:

$$y_1 = 5.4; y_2 = 5.2; y_3 = 5.5; y_4 = 3.9$$

$$\hat{y}_1 = 4.696; \hat{y}_2 = 5.214; \hat{y}_3 = 5.256; \hat{y}_4 = 4.836$$

Aplicamos la ecuación de cálculo:

$$s_r = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} = \sqrt{\frac{(5.4 - 4.696)^2 + (5.2 - 5.214)^2 + (5.5 - 5.256)^2 + (3.9 - 4.836)^2}{4}}$$

$$s_r = 0.598$$