Preprocesamiento

Minería de datos

Master Universitario en Tecnologías de Análisis de Datos Masivos
Escola Técnica Superior de Enxeñaría (ETSE)
Universidade de Santiago de Compostela

Contenidos de la presentación

- Limpieza de datos
  - Datos ausentes
  - Datos con ruido
  - Datos inconsistentes y con discrepancias
  - Variables con varianza cercana a cero
- Transformaciones de datos
  - Normalización
  - Discretización de variables numéricas a categóricas
  - Transformación de variables categóricas a numéricas
- Conclusiones

Limpieza de datos

- El resultado de una técnica de Minería de Datos depende de la calidad y cantidad de los datos.
  - La aplicación de técnicas de MD a datos de baja calidad generará conocimiento poco útil.
- Los conjuntos de datos están formados por objetos (ejemplos, instancias, tuplas, ...).
  - Pacientes, clientes, coches, estudiantes, ...
- Estos objetos se describen por medio de atributos (dimensiones, características, variables, ...)
  - Sexo, nombre, tipo, enfermedad, año de construcción, ...
  - Un atributo tiene asociado un tipo que define el dominio de los valores que pueden tomar.
- Los datos reales pueden contener gran cantidad de datos potencialmente incorrectos: fallos en los instrumentos de adquisición, error computacional o humano, error de transmisión, ...
- Por lo tanto, los errores pueden ser debidos a diferentes causas:
  - Datos ausentes: pueden faltar algunos atributos de interés, o algunos valores de los mismos.
  - Datos con ruido o errores, *outliers* e incluso datos duplicados.
  - Datos inconsistentes en la forma de discrepancias en códigos y nombres, o en datos duplicados.
  - Errores intencionados como forma de encubrir la falta de algunos datos.

Datos ausentes

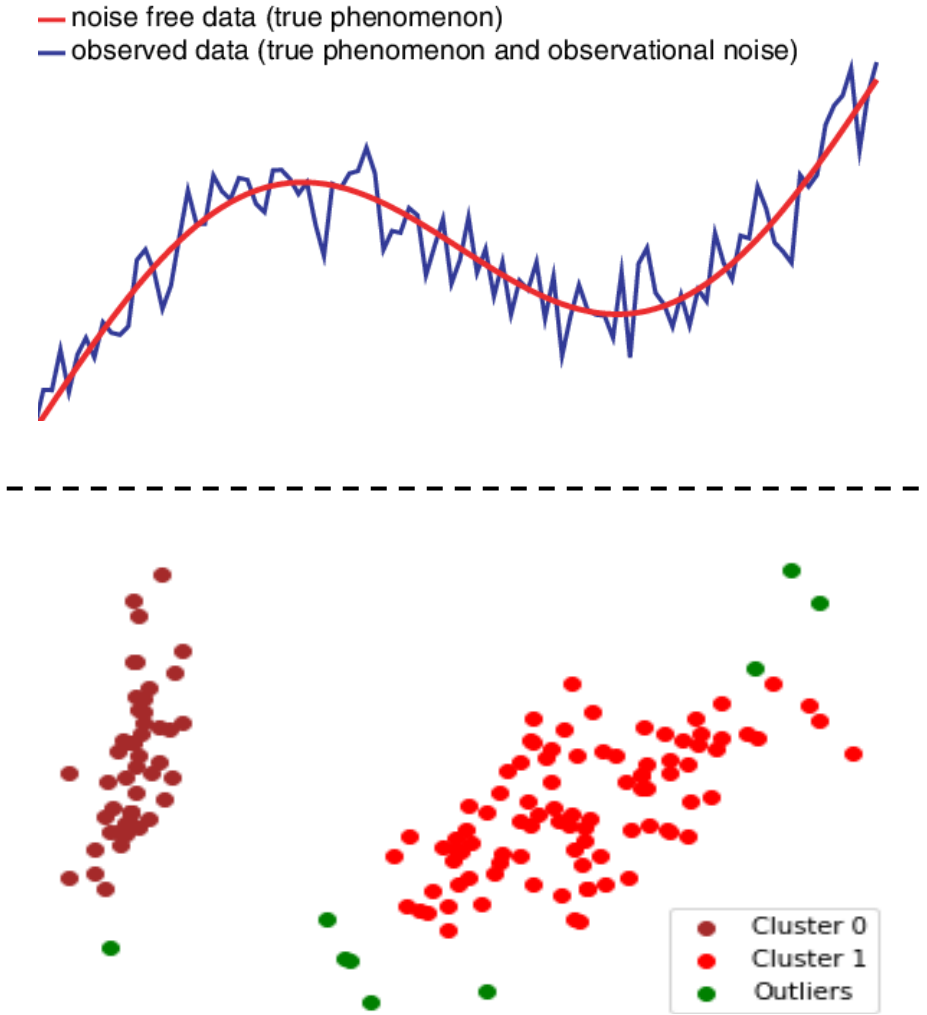
- Los datos ausentes pueden introducir varios problemas:
  - Pérdida de eficacia: se extraen menos patrones y las conclusiones pueden ser estadísticamente menos concluyentes.
  - Complicaciones a la hora de analizar los datos, ya que muchas técnicas no están preparadas para gestionarlos.
  - En el caso de que se requieran calcular valores agregados pueden impedir el cálculo.
  - Pueden producir sesgos en los modelos resultantes al aplicar los métodos de aprendizaje.
- Los datos ausentes pueden tomar distintas formas:
  - Nulos en las bases de datos
  - Datos sin valor o con un valor especial
  - Datos camuflados
- Para su tratamiento es necesario conocer su causa:
  - Algunos valores faltantes expresan situaciones relevantes. Por ejemplo, la falta de un teléfono puede indicar que la persona no quiere ser molestada.
  - Algunos datos realmente no existen.
  - Datos incompletos después de una combinación.

Datos ausentes: Soluciones

- No hacer nada. Algunos métodos son robustos ante este hecho (por ejemplo, árboles de decisión).
- Filtrar (eliminar) aquellos atributos con valores nulos.
  - Es una solución extrema.
  - Necesaria en el caso de un alto porcentaje de nulos.
  - En otros casos podemos encontrar otro atributo dependiente con una mayor calidad.
- Filtrar (eliminar) el objeto:
  - Se suele hacer cuando es un problema de clasificación y el valor de la clase está ausente.
  - No es efectivo cuando el porcentaje de ausentes varía mucho entre atributos.
  - Puede introducir sesgos en los datos.
- Reemplazar el hueco por un valor (imputación).
  - Manualmente si no hay muchos o por una constante global.
  - Mediante técnicas específicas. Por ejemplo, determinar el sexo a partir del nombre o del código postal.
  - Por un valor que preserve la media o la varianza para datos numéricos o la moda para nominales

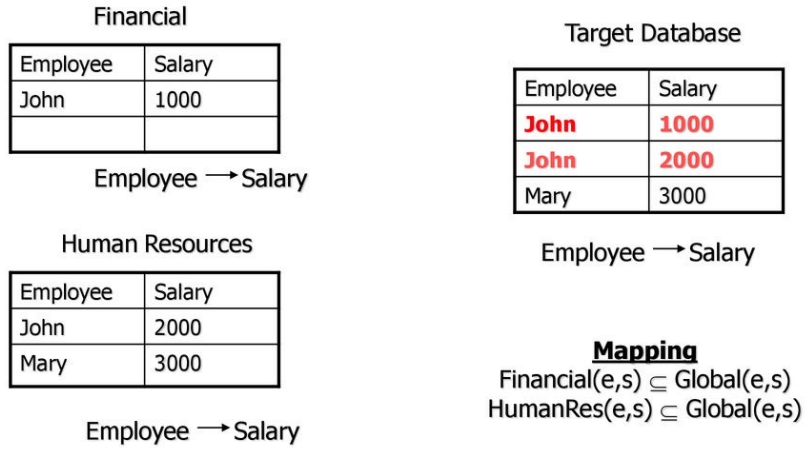
Datos con ruido

- Entendemos por ruido un error o varianza aleatoria en una medición de una variable.
- Existen varios métodos para suavizar los datos para eliminar el ruido.
  - Discretización. Este método permite suavizar un conjunto de valores ordenados consultando su vecindad.
    - Los valores ordenados se distribuyen en una serie de categorías con el mismo número de elementos (*equal frecuency*) o el mismo tamaño (*equal width*).
    - Se sustituyen los valores de cada categoría un un valor: media (*smooth by means*), mediana (*smooth by median*) o el extremo más cercano (*smooth by bin boundaries*).
  - Regresión. Se realiza un proceso de regresión para ajustar la función y sustituir los valores por los predichos por la función.
  - Clustering. El proceso de agrupamiento nos permite identificar los *outliers*.



Datos con inconsistencias y discrepancias

- Las inconsistencias en los datos pueden ser debidas a:
  - Formularios de entrada de datos mal diseñados o errores en los dispositivos de entrada.
  - Error humano en la introducción de datos o errores deliberados.
  - Obsolescencia de los datos, o que los datos hayan sido recogidos para otros usos.
  - Uso inconsistente del formato de datos o de los códigos.



- ¿Cómo detectarlos?
  - Uso de metadatos: Dominio y tipo de los atributos, valores permitidos, longitudes permitidas, análisis de su distribución.
  - Uso inconsistente de los formatos, por ejemplo, el uso de diferentes formatos para las fechas.
  - En los casos que se pueda aplicar: la regla de la unicidad, la regla de la correlatividad y la regla de la nulidad.
- Para resolver este problema podemos utilizar dos tipos de herramientas:
  - Las herramientas de depuración de datos (data scrubbing) utilizan conocimiento del dominio para detectar y corregir errores.
  - Las herramientas de auditoría de datos se centran en encontrar discrepancias mediante un análisis que permita descubrir reglas y relaciones en los datos y detectar las violaciones a las mismas.

Variables con varianza cercana a cero

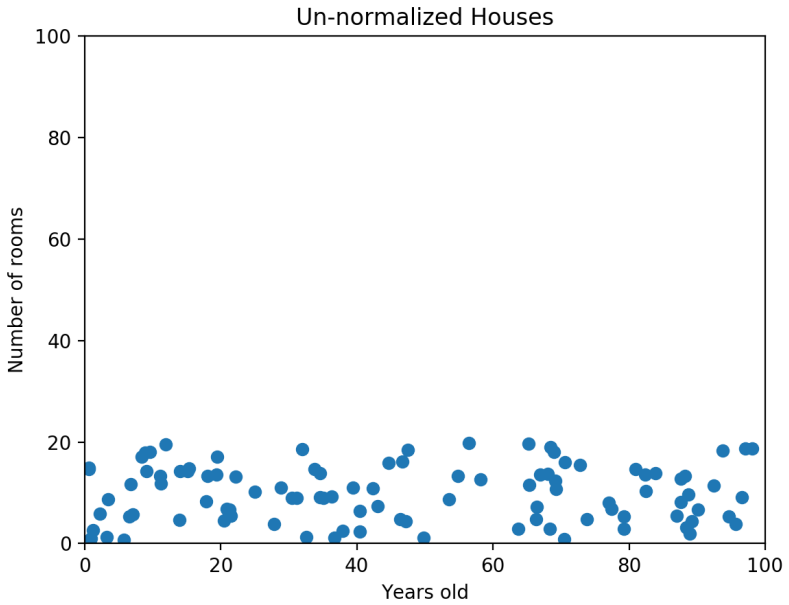
- En muchas situaciones podemos tener variables que tienen un solo valor (variables de varianza cero). En este caso hay modelos que no pueden tratar con este tipo de variables o muestran un comportamiento inestable.
- En otros casos pueden existir variables en las que un valor se presenta con una baja frecuencia, es decir, variables con varianza cercana a cero o muy desbalanceadas.
  - Estas variables se pueden convertir en variables con varianza cero cuando validamos por validación cruzada o bootstrap, afectando al resultado de la técnica elegida.
- Debido a esto, en muchos casos se suelen detectar y eliminar aquellas variables con varianza cercana a cero. Para detectarlas se utilizan dos métricas de forma conjunta:
  - La ratio entre la frecuencia del valor más frecuente y la frecuencia del segundo valor más frecuente (ratio de frecuencia): 1 para variables balanceadas, grande para variables mal balanceadas.
  - El porcentaje de valores únicos sobre el total de objetos, que se aproximará a cero a medida que la granulariadad de la variable aumenta.
- Si la ratio de frecuencia supera un límite establecido y el porcentaje de valores único cae por debajo de un límite establecido, podemos considerar la variable como una variable con varianza cercana a cero.

Transformaciones de datos

- Las técnicas de transformación nos permiten preparar los datos de forma apropiada para poder aplicar las distintas técnicas de minería de datos.
- La mayor parte de las técnicas de transformación de datos son aplicaciones sobreyectivas, es decir, a cada valor original le hace corresponder un valor transformado, pero varios valores originales pueden estar asociados a un mismo valor transformado.
- Entre las técnicas de transformación de datos tenemos:
  - Suavizado: para eliminar el ruido.
  - Agregación: cuando queremos resumir o agregar datos. Por ejemplo, acumular las ventas mensuales en las anuales. Este tipo de transformación se suele realizar en la construcción de los cubos de datos.
  - Generalización: de datos de bajo nivel o primitivos a datos de nivel más alto. Para ello es necesario la existencia de jerarquías conceptuales que definan el nivel de abstracción de los conceptos.
  - Creación de atributos a partir de los ya existentes (algunas técnicas las veremos en el siguiente capítulo).
  - Normalización que permite escalar los datos a un determinado rango, por ejemplo, [0, 1] o [-1, 1].

Normalización

- La idea básica consiste en escalar los valores de una variable a un rango determinado.
- Existen técnicas de minería de datos que requieren que los datos estén normalizados (máquinas de soporte de vectores o técnicas de agrupamiento) o que mejoran su rendimiento si previamente se normalizan los datos (redes neuronales).
- En las técnicas basadas en el concepto de distancia, la normalización evita que los variables con rangos mayores predominen sobre las de rangos menores.
- Existen numerosos métodos de normalización de los que destacamos: Normalización min-max, normalización por transformada z, normalización por escalado decimal.



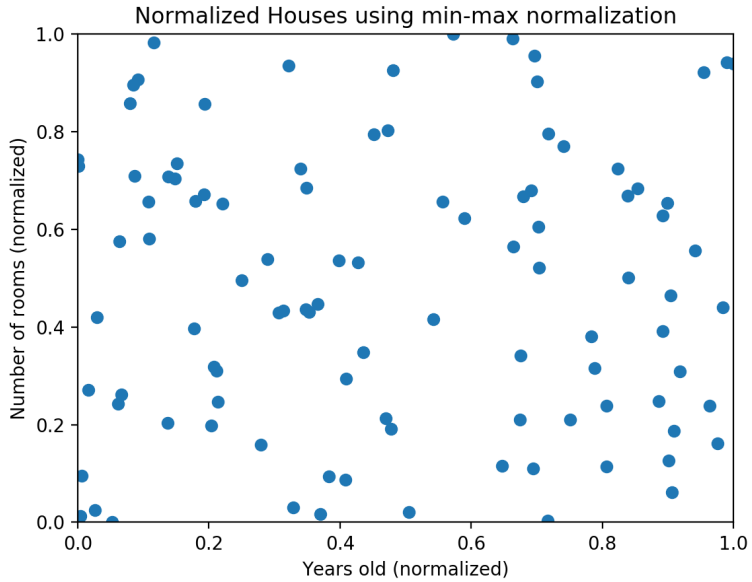
Cuando los datos aparecen así de aplastados, sabemos que tenemos un problema. El algoritmo de aprendizaje automático debería darse cuenta de que hay una gran diferencia entre una casa con 2 habitaciones y una casa con 20 habitaciones. Pero, como dos casas pueden tener 100 años de diferencia, la diferencia en el número de habitaciones contribuye menos a la diferencia global.

Normalización min-max

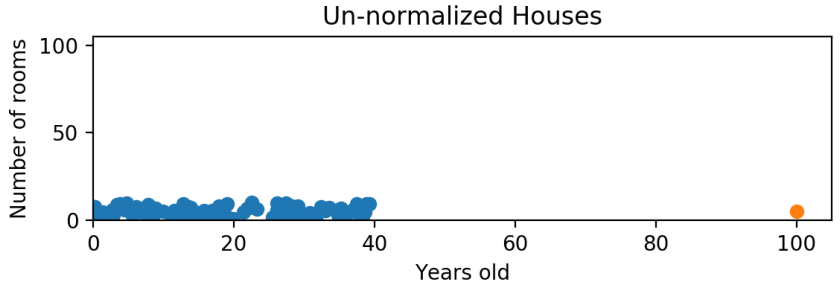
- Se realiza una transformación lineal sobre los datos originales. Supongamos que tenemos una variable A cuyo rango es $[min_A, max_A]$. Esta trasformación nos va a permitir transformar los valores v de la variable A en unos nuevos valores v' en el rango $[min'_A, max'_A]$, mediante la transformación:

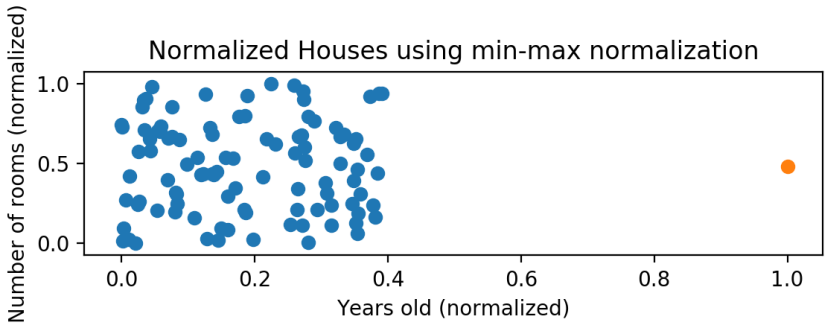
$$v' = \frac{v - min_A}{max_A - min_A} (max'_A - min'_A) + min'_A$$

- Esta transformación mantiene la relación entre los datos originales.
- Sensible a los outliers.



Ejemplo: Tenemos 99 valores entre 0 y 40, y uno único igual a 100.





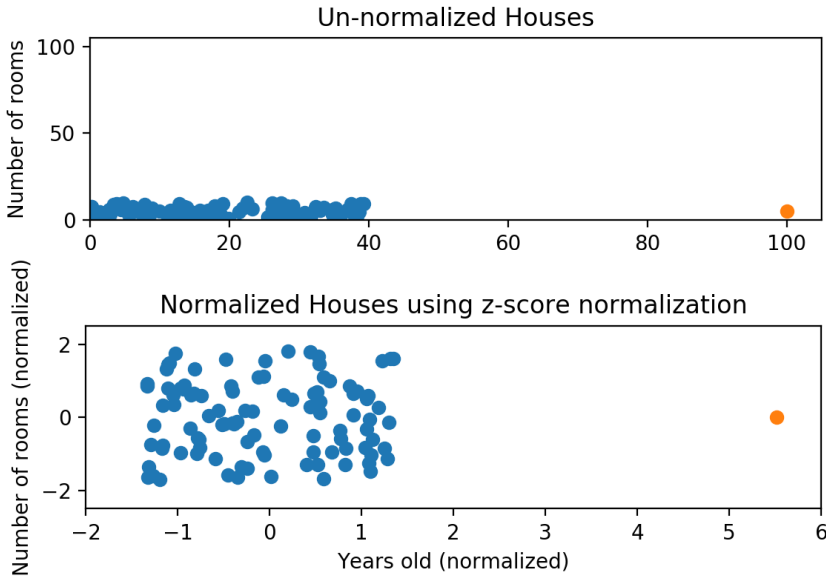
Normalización por transformada z (z-score)

- Los valores de una variable A son normalizados por su media $\bar{A}$ y su desviación típica σ_A :

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

- Este método se suele utilizar cuando los rangos de las variables son desconocidos, o existen valores anormales que dominan en la normalización min-max.
- Es un centrado y un escalado (media=0 y desviación típica=1):
  - Datos independientes de la unidad o de la escala.
  - Variables con la misma varianza y media.
- Es un cambio de unidad y no tiene efecto al comparar variables
- Las relaciones de correlación se mantienen.
- Sensible a los outliers.

Aunque los datos siguen "aplastados", los puntos están ahora más o menos en la misma escala para ambas características: casi todos los puntos están entre -2 y 2 tanto en el eje x como en el eje y.



Normalización por escalado decimal

- Se basa en el desplazamiento del punto decimal de los valores del atributo.
  - El número de posiciones que se desplaza el punto decimal depende del valor absoluto máximo de la variable A.
  - El cálculo de los nuevos valores se realiza de la siguiente fórmula:

$$v' = \frac{v}{10^j}$$

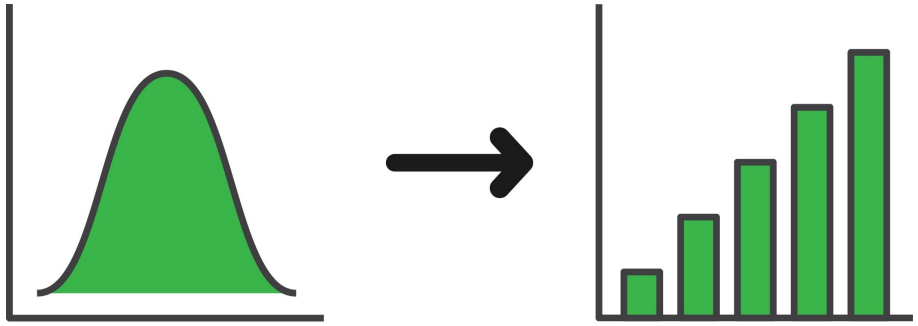
donde j es el entero más pequeño que hace que $\max|v'| < 1$.

Ejemplo: Supongamos que tenemos la siguiente tabla de salarios y queremos normalizarla por escalado decimal

Salario	Fórmula	Valor normalizado
40.000,00	40.000 / 100.000	0.4
56.000,00	56.000 / 100.000	0.56
31.000,000	31.000 / 100.000	0.31

Discretización

- La discretización (cuantización o "binning") es la conversión de un valor numérico en un valor nominal ordenado (que representa un intervalo o "bin").
  - Por ejemplo, convertir una nota en la escala [0,10] en una serie de valores ordenados [suspenso, aprobado, notable, sobresaliente, matrícula de honor].

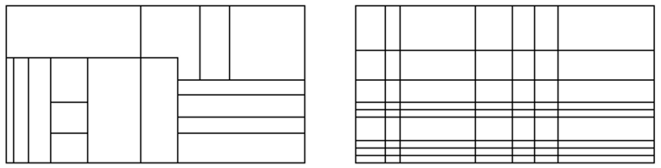


- ¿Por qué discretizar?
  - Puede mejorar la calidad del conocimiento descubierto.
  - Se considera un mecanismo de reducción de la dimensionalidad de los datos. Podemos pasamos de tener un dominio numérico grande a un subconjunto de categorías.
  - Puede mejorar el mantenimiento de los datos.
  - Algunas técnicas de minería de datos solo aceptan atributos discretos.
  - Reduce el tiempo de ejecución de ciertas técnicas como las reglas de asociación, clasificación y predicción.
  - Cuando existen ciertos umbrales significativos.

Tipos de discretización

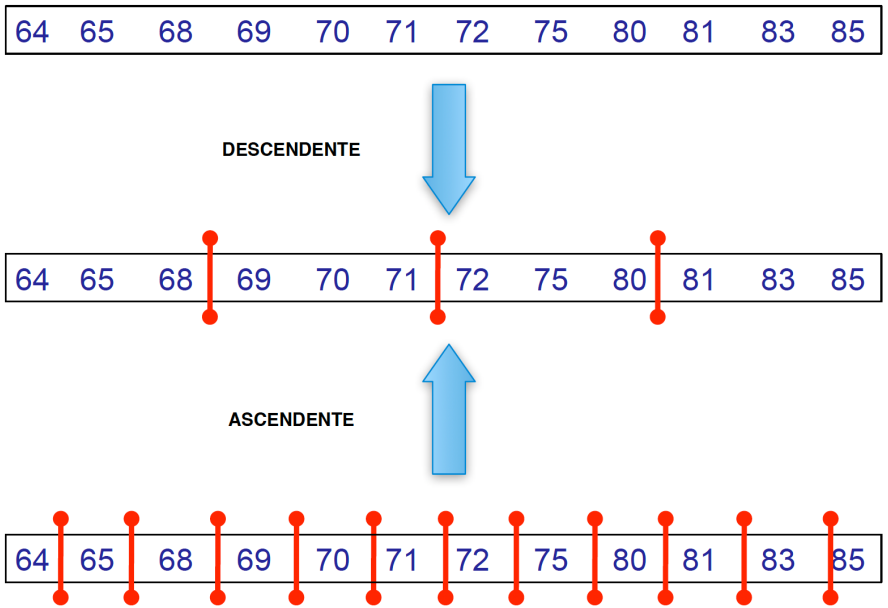
- **Supervisada o no supervisada**
  - Si la técnica de clasificación utiliza la información sobre la clase estaremos en un caso de discretización supervisada.
  - Al utilizar la información de la distribución de clases, este tipo de discretización puede facilitar las tareas de clasificación.
  - En otro caso, hablaremos de discretización no supervisada.
- **Local o global**
  - Los métodos globales aplican los mismos puntos de corte a todas las instancias.
  - Los métodos locales utilizan diferentes puntos de corte a diferentes conjuntos de instancias.
- **Ascendente (bottom-up) o descendente (top-down)**
  - Top-down (splitting). Se comienza seleccionando uno o más puntos para dividir el rango del atributo. Se va repitiendo el proceso con cada nuevo intervalo hasta que no se pueda dividir más.
  - Bottom-up (merging). Se van fusionando puntos cercanos entre sí para formar intervalos y repetir el proceso con los nuevos intervalos.

Ejemplo: diferencia entre local y global



LOCAL GLOBAL

Ejemplo: diferencia ascendente vs descendente



Técnicas más comunes

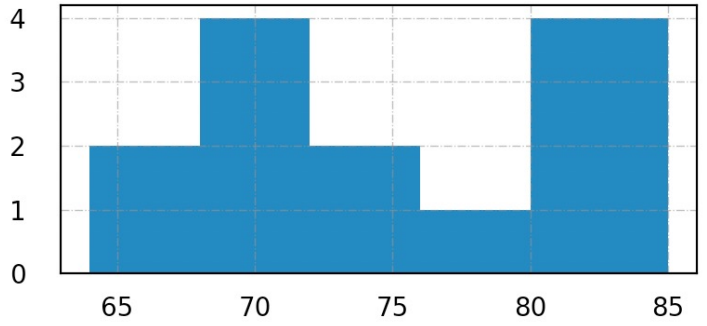
- Binning (descendente, no supervisada) - (introducido al hablar del suavizado).
- **Análisis del histograma** (descendente, no supervisada).
- **Discretización Basada en la Entropía** (descendente supervisada).
- **Fusión de intervalos mediante análisis χ^2** (ascendente, supervisado).
- **Análisis de clúster** (ascendente o descendente, no supervisado).
- Otras técnicas:
  - Análisis de árboles de decisión
  - Análisis de correlación

Binning with equal width

- Se divide el rango de valores en intervalos de la misma longitud.
- Para determinar la longitud de los intervalos $\boldsymbol{w = (V_{max} - V_{min})/N}$ donde N es el número de intervalos y V_{max} y V_{min} son el valor máximo y mínimo que toma el atributo
- Los límites de los intervalos: $V_{min} + w$, $V_{min} + 2w$, …, $V_{min} + (N-1)w$
- Puede verse alterada por la presencia de outliers y datos sesgados.

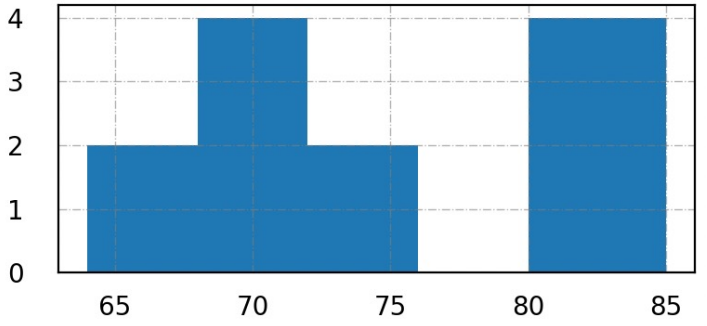
Ejemplo: Dividir los valores de temperatura [64, 65, 68, 69, 70, 71, 72, 75, 76, 80, 81, 83, 85] en 5 intervalos de la misma longitud

La longitud es 21/5=4.2
bin1 = [64-68]
bin2 = (68-72]
bin3 = (72-76]
bin4 = (77-80]
bin5 = (81-85]



Ejercicio: Dividir los valores de temperatura [64, 65, 68, 69, 70, 71, 72, 75, 80, 81, 83, 85] en 5 intervalos de la misma longitud

La longitud es 21/5=4.2
bin1 = [64-68]
bin2 = (68-72]
bin3 = (72-76]
bin4 = (76-80]
bin5 = (80-85]



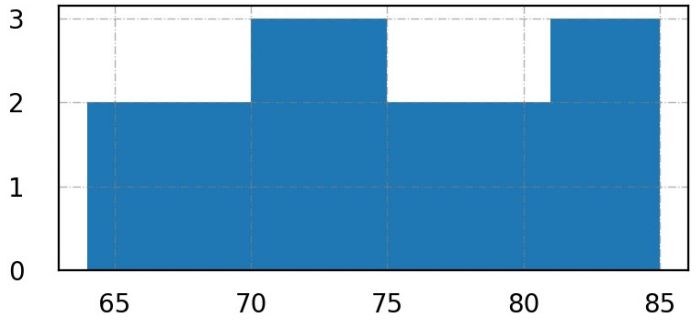
Como se puede ver, pueden existir intervalos sin elementos

Binning with equal depth (equal frequency)

- Se divide el rango de valores en intervalos que contengan aproximadamente el mismo número de elementos.
- Para saber cuántos elementos debe tener cada intervalo, se divide el número total de instancias por el número de intervalos.
- Para determinar cuáles son los valores en los que realizar la partición, se suele utilizar el punto medio entre los dos extremos de los intervalos.
- En el caso de que valores repetidos caigan en intervalos distintos habrá que tomar la decisión de a qué intervalo se asignan dichos valores, permitiendo que existan intervalos con un número de valores alejados de la media.

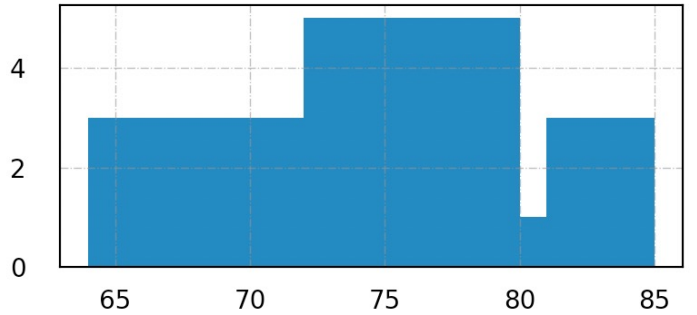
Ejemplo: Dividir los valores de temperatura [64, 65, 68, 69, 70., 71, 72, 75, 80, 81, 83, 85] en 5 intervalos con la misma frecuencia

La frecuencia es 12/5=2.4
bin1 = [64, 65]
bin2 = [68, 69]
bin3 = [70,71,72]
bin4 = [75, 80]
bin5 = [81, 83, 85]



Ejemplo: Dividir los valores de temperatura [64, 65, 68, 69, 70., 71, 72, 75, 75, 75, 75, 80, 81, 83, 85] en 5 intervalos con la misma frecuencia

La frecuencia es 15/5=3
bin1 = [64, 65, 68]
bin2 = [69, 70, 71]
bin3 = [72, 75, 75, 75, 75]
bin4 = [80]
bin5 = [81, 83, 85]

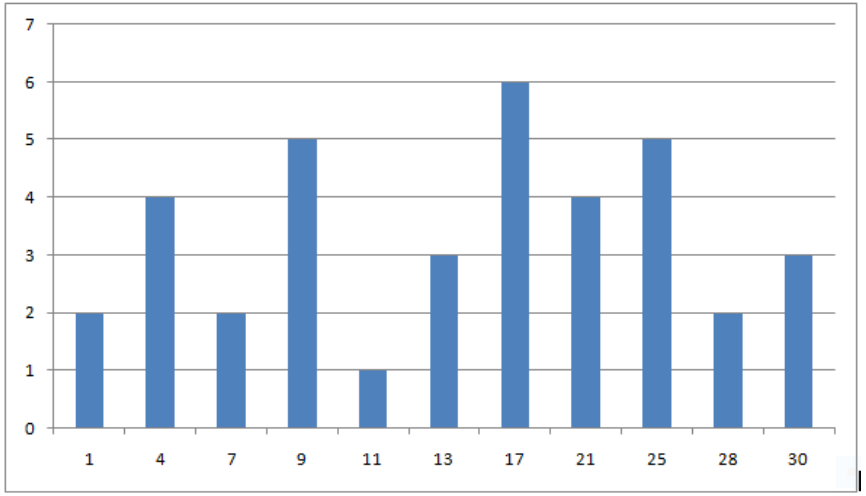


En este caso, se optó por poner los valores repetidos en el bin3, dejando desequilibrados los bin 4 y 5.

Discretización basada en histograma

- Un histograma, para un atributo concreto, muestra la frecuencia de cada uno de los posibles valores del atributo.
- De esta forma, un histograma agrupa en un mismo balde (bucket) pares valores-frecuencia.
- Podemos discretizar el rango de valores de un atributo agrupando baldes:
  - Intervalos de la misma longitud (equal-width).
  - Intervalos de la misma frecuencia (equal-depth).
  - Varianza óptima. Se consideran todas las posibilidades de agrupación de baldes y se selecciona la de menor varianza. En el cálculo de la varianza, los baldes están ponderados por la frecuencia del mismo.
  - Máxima diferencia. Los límites de los baldes (intervalos) se establecen entre los valores consecutivos con las $\beta - 1$ mayores distancias, siendo β el número de intervalos deseados.

Ejemplo: Los siguientes datos muestran el precio de los artículos más vendidos en orden de clasificación: 1,1,4,4,4,4,7,7,9,9,9,9,9,11,13,13,13,17,17,17,17,17,17,21,21,21,21,25,25,25,25,25,28,28,30,30,30.



Discretización basada en histograma

- Las particiones basadas en la varianza y la diferencia suelen ser más precisas y prácticas.
- Efectivos tanto para datos densos como dispersos.
- Efectivos tanto para datos uniformes como sesgados.
- Existen muchos criterios de particionado de un conjunto formado por n elementos:
  - Raíz cuadrada:

    número de intervalos = $\sqrt{n}$

    ancho = $\frac{\max(x)-\min(x)}{\sqrt{n}}$
  - Sturges:

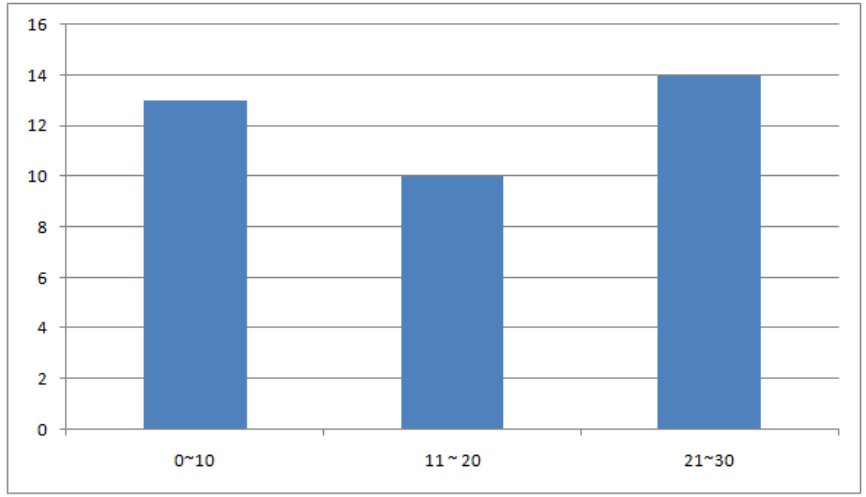
    número de intervalos = $\lceil 1 + log_2(n) \rceil$

    ancho = $\frac{\max(x)-\min(x)}{\lceil 1+log_2(n) \rceil}$
  - Rice:

    número de intervalos = $\left\lceil 2\sqrt[3]{n} \right\rceil$

    ancho = $\frac{\max(x)-\min(x)}{\left\lceil 2\sqrt[3]{n} \right\rceil}$

Ejemplo: Particionado de los datos anteriores en baldes de igual longitud de 10:



Discretización basada en la entropía

- Técnica descendente y supervisada, que utiliza el concepto de ganancia de información.
- Parecida a la utilizada en la generación de árboles de decisión.
- Sea un conjunto de datos D, para discretizar un atributo cualquier A:
  1. Cada posible valor de A es considerado como límite de un intervalo o punto de ruptura.
     - Cada posible punto de ruptura particiona los datos en dos subconjuntos uno, D_1, con aquellos datos que satisfacen A ≤ punto de ruptura y otro, D_2, con los que satisfacen que A > punto de ruptura.
  2. De todos los posibles puntos de rupturas cogemos aquel que produzca una partición con la ganancia mínima de información, es decir, min(I(A,D)).

$$I(A,D) = \sum_{D_i} \frac{|D_i|}{|D|} E(D_i)$$

con:

$$E(S) = -\sum_{i=1}^{n} p_i log_2(p_i)$$

donde E determina la entropía, n es el número de clases y p_i la probabilidad de que una instancia pertenezca a la clase i.

  3. El proceso se va repitiendo de forma recursiva en cada una de las particiones obtenidas hasta que se alcance un criterio de parada, por ejemplo, que la ganancia de información alcance un umbral o que se alcance el número de intervalos deseados.

Discretización basada en la entropía

Ejemplo: En un entorno controlado, se midieron en determinados momentos la temperatura de unas juntas y si en ese momento la junta fallaba o no

Paso 1: Calcular las entropías para la clase

Fallo en la junta	
Y	N
7	17

$$E(Fallo) = E(7,17) = E(0.29,0.71) =$$
$$-0.29 \times log_2(0.29) - 0.71 \times log_2(0.71) = 0.871$$

Paso 2: Calcular las entropías para la clase dado un bin

		Fallo en la junta	
		Y	N
Temperatura	<= 60	3	0
	> 60	4	17

$$E(Fallo, Temperatura) =$$
$$P(\leq 60) \times E(3,0) + P(> 60) \times E(4,17) = 0.615$$

Paso 3: Calcular la ganancia dado un bin

$$Ganancia\ (Fallo, Temperatura) = E(Fallo) - E(Fallo, Temperatura) = 0.256$$

Se puede comprobar que el mejor intervalo para la temperatura es (<=60, >60), ya que devuelve la mejor ganancia si lo comparamos con otras configuraciones:

- Intervalo (<=70, >70) – Ganancia = 0.101
- Intervalo (<=75, >75) – Ganancia = 0.148

Temperatura	Fallo
53	Y
56	Y
57	Y
63	N
66	N
67	N
67	N
67	N
68	N
69	N
70	N
70	Y
70	Y
70	Y
72	N
73	N
75	N
75	Y
76	N
76	N
78	N
79	N
80	N
81	N

Discretización basada en el test de la χ^2

- Técnica ascendente y supervisada.
- La idea básica consiste en ir fusionando intervalos adyacentes que presenten una distribución de clases parecida.

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- Sea un conjunto de datos D, para discretizar un atributo cualquier A:
  1. Cada posible valor de A es considerado como un intervalo diferente.
  2. Se calcular el estadístico χ^2 en cada par de intervalos adyacentes.
  3. Aquellos pares de intervalos con los valores χ^2 más pequeños (distribución de clases similar) se fusionan.
  4. El proceso continúa de forma recursiva hasta que se alcanza algún criterio de parada:
     - Cuando el valor χ^2 para todos los pares de intervalos adyacentes sobrepasa un determinado umbral dependiente del nivel de significancia, normalmente entre 0,1 y 0,01.
     - Cuando se alcance el número de intervalos deseados.
     - Cuando la frecuencia relativa de las distintas clases en cada intervalo presenta diferencias mayores a un determinado umbral.

Discretización basada en el test de la χ^2

Ejemplo:

	Juega a la consola (A)	No juega a la consola (B)	Suma (filas)
Le gustan las ciencias (S)	25 (9)	20 (36)	45
No le gustan las ciencias (N)	5 (21)	100 (84)	105
Suma (columnas)	30	120	150

$$Exp(S,A) = 45 * \frac{30}{150} = 9;\ Exp(S,B) = 45 * \frac{120}{150} = 36$$

$$Exp(N,A) = 105 * \frac{30}{150} = 21; \mathrm{Exp(N,B)} = 105 * \frac{120}{150} = 84$$

$$\chi^2 = \frac{(25-9)^2}{9} + \frac{(5-21)^2}{21} + \frac{(20-36)^2}{36} + \frac{(100-84)^2}{84} = 50.793$$

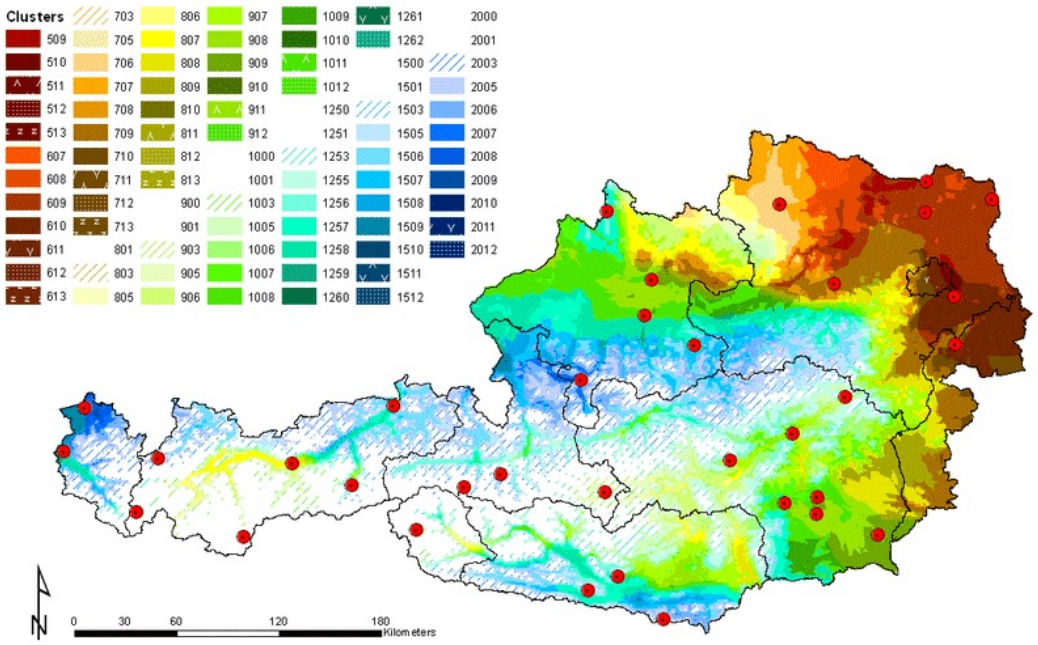
Prueba de independencia de dos factores: cuanto mayor sea el valor χ^2, más probable es que los factores estén relacionados.

Esto muestra que las personas a las que les gusta la ciencia ficción y juegan a la consola está correlacionados en el mismo grupo.

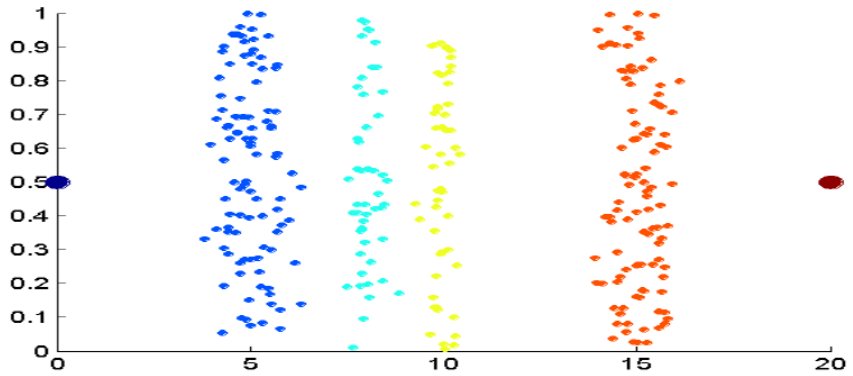
Análisis clúster

- Podemos utilizar un algoritmo de agrupamiento para discretizar un atributo numérico.
- Solo haría falta asociar una categoría a cada grupo o clúster.
- Pueden generar discretizaciones de alta calidad:
  - Tienen en cuenta la distribución del atributo a discretizar, y la distancia entre los datos.
  - Técnicas de clustering jerárquico nos permiten obtener una jerarquía conceptual.

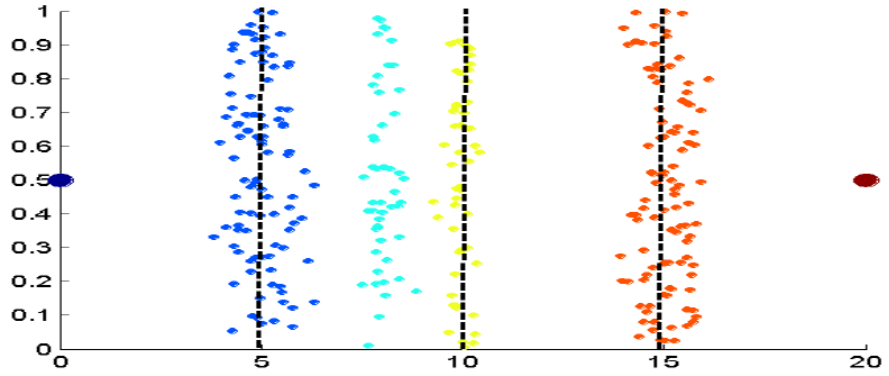
Ejemplo: Agrupamientos por temperaturas y precipitaciones



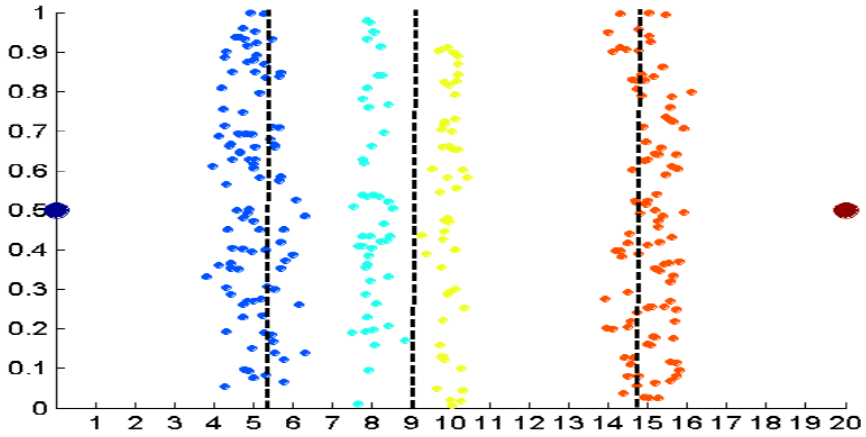
Comparativa



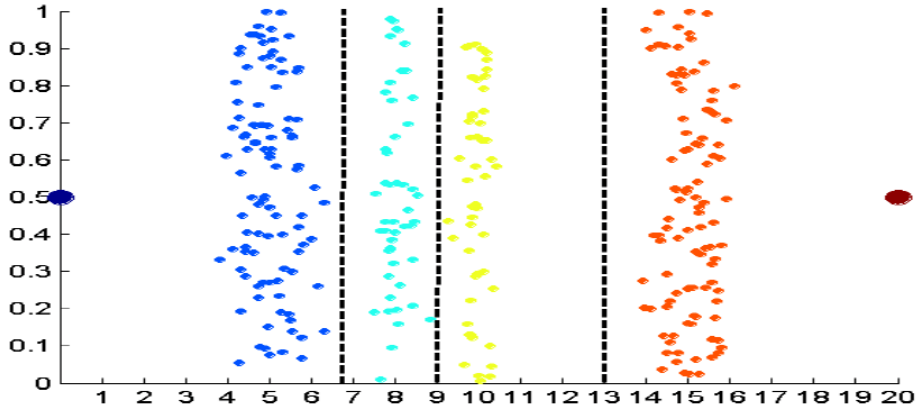
Datos originales



Intervalos de igual longitud



Intervalos de igual frecuencia



Intervalos a través de un método de agrupamiento (k-medias)

Variables categóricas a numéricas

- Variable cuyo dominio lo forman un número finito etiquetas/categorías.
  - Nominales: etiquetas/categorías no relacionadas
  - Ordinales: etiquetas/categorías ordenadas.
- Existen técnicas que pueden manipular datos categóricos y otras que solo admiten variables numéricas
- Principales técnicas de transformación:
  - Codificación ordinal
  - Codificación One-Hot
  - Codificación por variables dummy

Codificación ordinal

- Se aplica a las variables categóricas ordinales.
- La idea es mantener el orden de las categorías asignando un número entero a cada categoría.
- Por ejemplo, las calificaciones {A, B, C, D, E, F} se podrían codificar como {5, 4, 3, 2, 1, 0}.
- Cuando se aplica a la variable a predecir:
  - Podemos estar prediciendo valores entre las categorías, p.ej. 4.5, que pueden no tener sentido.
  - En la mayoría de los casos la variable a predecir se puede (y debe) mantener como categórica.

Codificación One-Hot

- Se aplica a las variables categóricas nominales.
- Cuando no existe una relación entre las categorías, la codificación ordinal carece de sentido, ya que al aplicarla estaríamos imponiendo un orden.
- Procedimiento One-Hot:
  - Se crea una nueva variable binaria para cada categoría.
  - Cada nueva variable tomará el valor 1 si ésta presenta la categoría, 0 en caso contrario.

Ejemplo: Nacionalidad = {Alemana, Francesa, Italiana, Portuguesa}

id	Nacionalidad
i1	Alemana
i2	Portuguesa
i3	Italiana
i4	Francesa

id	Nac_Alemana	Nac_Francesa	Nac_Italiana	Nac_Portuguesa
i1	1	0	0	0
i2	0	0	0	1
i3	0	0	1	0
i4	0	1	0	0

Codificación por variables dummy

- La codificación One-Hot tiene el problema de introducir información redundante.
  - Conocer el valor asignado a tres categorías permite inferir el valor asociado a la otra categoría
  - Esto introduce un problema de multicolinearidad.
  - Problemático en redes neuronales o técnicas de regresión sin regularización.
- Solución:
  - Para N categorías se crean N-1 variables.
  - La categoría excluida se codifica mediante un 0 en el resto de las variables creadas.

Ejemplo: Nacionalidad = {Alemana, Francesa, Italiana, Portuguesa}

id	Nacionalidad
i1	Alemana
i2	Portuguesa
i3	Italiana
i4	Francesa

id	Nac_Francesa	Nac_Italiana	Nac_Portuguesa
i1	0	0	0
i2	0	0	1
i3	0	1	0
i4	1	0	0

Conclusiones

- En este capítulo hemos analizado la importancia del procesamiento de datos previo a la aplicación de cualquier técnica de minería de datos.
  - Bien debido a unos datos de baja calidad.
  - O bien debido a que la técnica utilizada lo requiere.
- Las técnicas de limpieza de datos nos permiten tratar con datos ausentes, con ruido e inconsistentes.
- Las técnicas de transformación de datos nos permiten transformar los datos de entrada para realizar cambios de escala o discretizar variables continuas.

Bibliografía

- Jiawei Han, Micheline Kamber, and Jian Pei. Data mining: concepts and techniques: concepts and techniques. Elsevier, 2011.
- José Hernández Orallo, Ma José Ramírez Quintana, and César Ferri Ramírez. Introducción a la Minería de Datos. Pearson Prentice Hall, 2004.
- Basilio Sierra Araujo. Aprendizaje automático: conceptos básicos y avanzados: aspectos prácticos utilizando el software Weka. Pearson Prentice Hall Madrid, 2006.