Preprocesamiento de Datos

Minería de Datos

José T. Palma

Departamento de Ingeniería de la Información y las Comunicaciones
Universidad de Murcia

15 de febrero de 2021

Contenidos de la presentación

1. Introducción
2. Limpieza de datos
   - Datos ausentes
   - Datos con ruido
   - Datos inconsistentes y discrpancias
   - Variables con varianza cercana a cero
3. Transformaciones de datos
   - Normalización
   - Discretización: de variables numéricas a categóricas
   - De variables categóricas a numéricas
4. Conclusiones

Introducción

- El resultado de una técnica de Minería de Datos depende fuertemente de calidad y cantidad de los datos.
  - la aplicación de técnicas de minería de datos a datos de baja calidad generará conocimiento poco útil.
- Como ya sabemos los conjunto 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 construccion,...
- Un atributo tienen asociado un tipo que define el dominio de los valores que pueden tomar.

Limpieza de Datos

- 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 diferente causas:
  - **Datos incompletos**: pueden faltar algunos atributos de interés, o alguno 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:
    - Edad="42" y Fecha de Nacimiento= "12/07/2015"
    - Algunos objetos se avalúan en la escala "1,2,3" y otros en la escala"A,B,C".
  - **Errores intencionados** como forma de encubrir la falta de algunos datos
    - Encontrarnos la misma fecha de nacimiento para todos las personas, o gran parte de ellas.

Datos ausentes: Problemas

- Los **datos ausentes** pueden introducir varios problemas:
  - Pérdida de eficacia: se extraen menos patrones y, además, las conclusiones pueden ser estadísticamente menos concluyentes.
  - Complicaciones a la hora de analizar los datos, ya que muchas técnicas no están preparadas para gestionarlos.
    - y si pueden gestionarlos, puede que ignoren todo el objeto o el atributo.
  - 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 a los datos ausentes o a los datos completos.

Datos ausentes: Detección

- Si los datos proceden de una base de datos, generalmente los datos ausentes están representados como nulos.
- Pero en la mayoría de los casos puede resultar difícil detectarlos, es el caso de los *nulos camuflados*:
  - Las restricciones de integridad del sistema no nos permiten la introducción de nulos en campos con formato: direcciones, teléfonos, códigos postales o número de tarjetas de crédito, segundo apellido en extranjeros.
- Para el tratamiento de estos datos hay que conocer su causa:
  - Algunos valores faltantes expresan situaciones relevantes. La falta del teléfono puede implicar que la persona no quiere ser molestada.
  - Algunos datos realmente no existen.
  - Datos incompletos después de una combinación.

Datos ausentes: Soluciones I

- 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 en un problema de clasificación cuando la clase está ausente.
  - No es efectivo cuando el porcentaje de ausentes varía mucho entre atributos.
  - puede introducir sesgos en los datos.

Datos ausentes: Soluciones II

- Reemplazar el hueco por un valor.
  - Manualmente si no hay muchos o por una constante global.
  - Por un valor que preserve la media o la varianza para datos numéricos o la moda para nominales.
  - **Imputación**:
    - Usar el valor medio, de todos los valores de los atributos o sólo de los que pertencen a la misma clase
    - Usar el valor más probable
    - predecir el valor mediante alguna técnica predictiva (regresión o clasificación) como knn, árboles, regresión,...
  - Mediante técnicas específicas. Por ejemplo, determinación del sexo a partir del nombre o el código postal a partir de la dirección.

Datos ausentes: Soluciones III

- Aunque la imputación es la técnica más frecuente, hay que ser consciente de que:
  - se está perdiendo información, no sabremos que el dato era ausente.
  - puede que el dato que estamos introduciendo sea erróneo.
- En algunos casos, se puede crear un atributo adicional booleano que indique que el dato era ausente.

Datos con Ruido I

- 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*).

Datos con Ruido II

- **Regresión**. Se realiza un proceso de regresión para ajustar la función y sustituir los valores por los predichos por la función. Se pueden utilizar multitud de métodos diferentes.
- **Clustering**. El proceso de clustering o agrupamiento nos permite identificar los outliers.

Datos: Inconsistencias y Discrepancias

- Antes de proceder a resolver los problemas planteados por los datos ausentes y el ruido, se deben detectar las discrepancias en los datos.
- Las inconsistencias 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.

Datos: Detección de las Inconsistencias y Discrepancias

- 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 consecutividad 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.

Datos: Variables con varianza cercana a cero I

- En muchas situaciones podemos tener variables que tiene un sólo 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 zero 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.

Datos: Variables con varianza cercana a cero II

- Para detectarlas se utilizan dos métricas de forma conjunta:
  - el 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 objetos, que se aproximará a cero a medida que la granulariadad de la variable aumenta.
- Si el 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 I

- Las técnicas de transformación nos permiten preparar los datos de forma apropiada para poder aplicar las distintas técnicas de minería de datos.
- Básicamente la mayor parte de las técnicas de transformación de datos es aplicación sobreyectiva, es decir, a cada valor original le hace corresponder un valor transformado, pero varios valores originales pueden estar asociados a un mismo valor transformado.

Transformaciones de datos II

- Entre las técnicas de transformación de datos tenemos:
  - **Suavizado**: para eliminar el ruido tal y como acabamos de ver.
  - **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.
  - **Generalizacion:** de datos de bajo nivel o primitivos a datos de nivel mas alto. Para ello es necesario la existencia de jerarquías conceptuales que definan el nivel de abstracción de los conceptos.
  - **Creacción de atributos** a partir de lo ya existentes (algunas técnicas las veremos en el siguiente capíutlo).
  - **Normalización** que permite escalar los datos a un determinado rango, por ejemplo, $[0, 1]$ o $[-1, 1]$.

Transformaciones de datos: Normalización I

- 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 qué 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* y *normalización por escalado decimal*

Transformaciones de datos: Normalización II

- **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 $[min'_A, max'_B]$, mediante la transformación:

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

  Esta transformación mantiene las relación entre los datos originales.

Transformaciones de datos: Normalización III

- **Nomarlización por transformada z (z-score)**. En este caso, los valores de un variable A son normalizados en función de la su media $\overline{A}$ y su desviación típica, σ_A:

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

  - 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
    - desviación típica 1

Transformaciones de datos: Normalización IV

- Esto permite obtener:
  - 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 a la hora de comparar variables
- Las relaciones de correlación se mantienen.

Transformaciones de datos: Ejemplo de escalado I

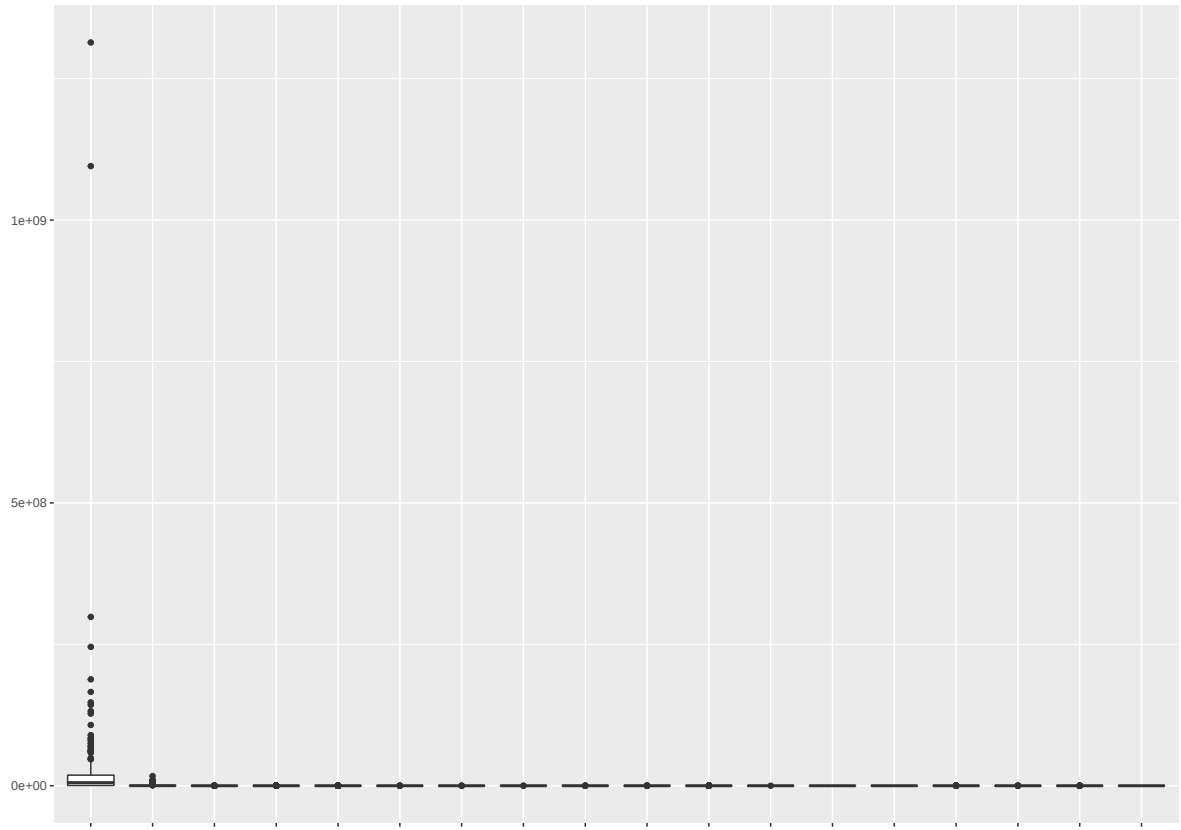


Figura: Datos no escalados

Transformaciones de datos: Ejemplo de escalado I

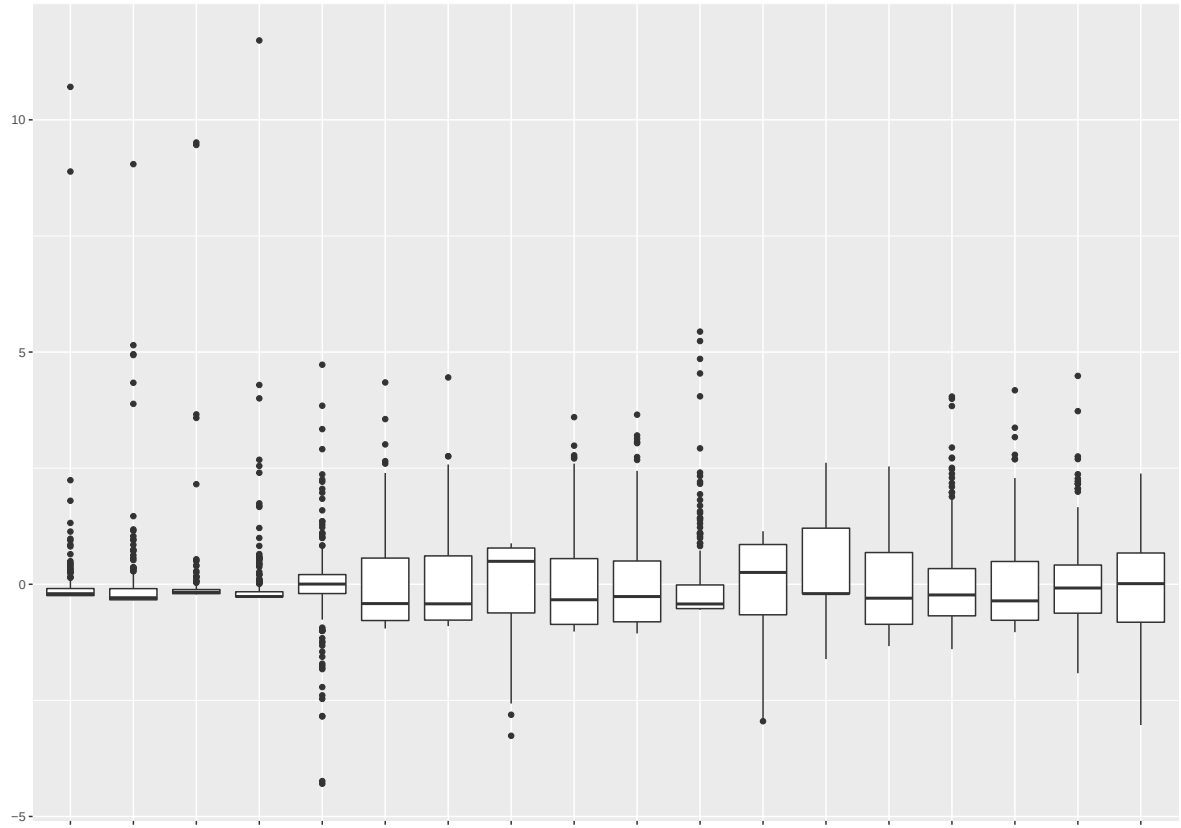


Figura: Datos Escalados

Transformaciones de datos: Normalización V I

- **Normalización por escalado decimal**. Este tipo de normalización 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 máx $|v'| < 1$

Transformaciones de datos: Discretización I

- 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?
  - Algunas técnicas de minería de datos sólo aceptan atributos discretos.
  - Cuando existen ciertos umbrales significativos.
  - La integración de escales diferentes.
  - Cuando la interpretación de la escala no sea lineal.

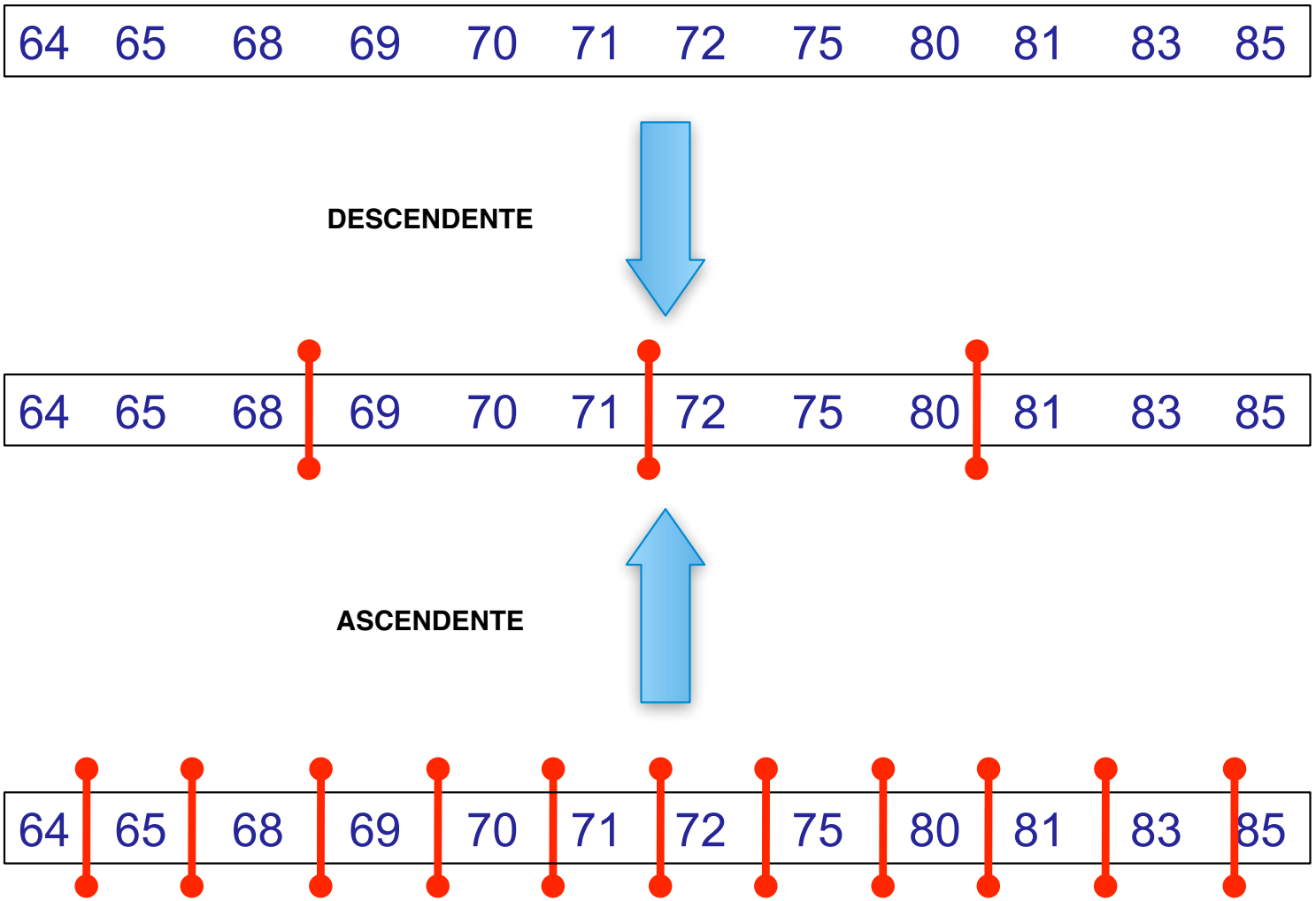
Transformaciones de datos: Discretización I

- 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 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 todos las instancias.
    - Los métodos **locales** utilizan diferentes puntos de corte a diferentes conjunto de instancias.

Transformaciones de datos: Discretización II

- **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.

Transformaciones de datos: Discretización III



Transformaciones de datos: Discretización IV

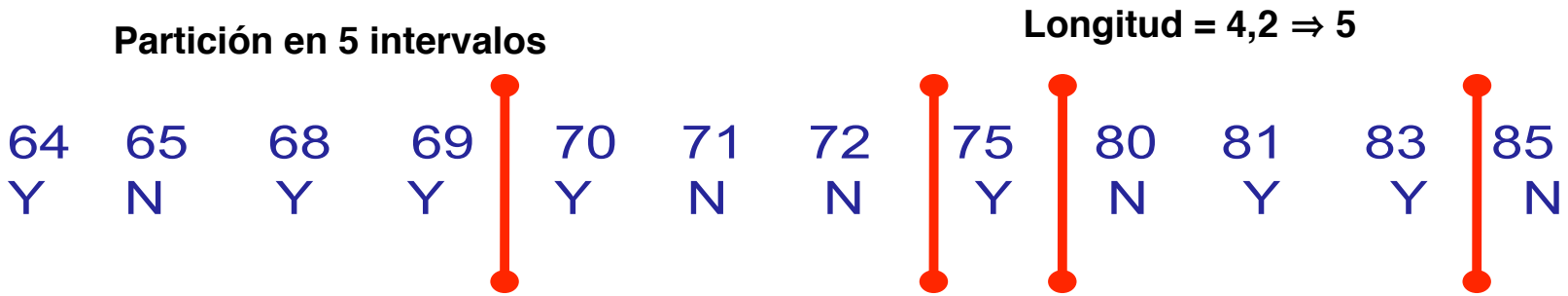
Técnicas más comunes. Entre las técnicas más utilizadas vamos a analizar:

- **Binning** (descendente, no supervisada). Que ya hemos 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 cluster** (ascendente o descendente, no supervisado).

Discretización: Binning I

- **Binning con intervalos de la misma longitud (equal-width)**. Se divide el rango de valores en intervalos de la misma longitud.
  - Para determinar la longitud de los intervalos $w = (V_{max} - V_{min})/N$.
  - donde N es el número de intervalos y V_{max} y V_{min} el valor máximo y mínimo que toma el atributo y 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.

Discretización: Binning II



$$\text{bin}_1 = [64-, 69]$$
$$\text{bin}_2 = (69, 75]$$
$$\text{bin}_3 = (75, 80]$$
$$\text{bin}_4 = (80, 85]$$
$$\text{bin}_5 = (85, 90+]$$

Discretización: Binning III

- **Binning por intervalos de la misma amplitud (equal-depth, frecuency)**. 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.

Discretización: Binning IV

Partición en 5 intervalos **Número de elementos = 2,4 ⇒ 2**

Discretización: Basada en histograma I

- Un **Histograma**, para un atributo concreto, nos 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 posiblidades 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.

Discretización: Basada en histograma II

- **Máxima diferencia**. Los límites de los baldes (intervalos) se establece entre los valores consecutivos con la $\beta - 1$ mayores distancias, siendo β el número de intervalos deseados.

- Las particiones basadas en la varianza y la diferencia suelen ser las más precisas y prácticas.
- Los histogramas también son muy efectivos tanto para datos densos como dispersos.
- También son efectivos tanto para datos uniforme como altamente sesgados.

Discretización: Basada en histograma III

- Existen muchos criterios, entre los que podemos destacar:
  - **Raiz Cuadrada:**

$$n\_intervalos = \sqrt{n} \; ; \; ancho = \frac{\text{máx}(x) - \text{mín}(x)}{\sqrt{n}}$$

  - **Sturges:**

$$n\_intervalos = \lceil 1 + \log_2 n \rceil \; ; \; ancho = \frac{\text{máx}(x) - \text{mín}(x)}{\lceil 1 + \log_2 n \rceil}$$

  - **Rice:**

$$n\_intervalos = \lceil 2\sqrt[3]{n} \rceil \; ; \; ancho = \frac{\text{máx}(x) - \text{mín}(x)}{\lceil 2\sqrt[3]{n} \rceil}$$

Discretización: Basada en histograma IV

- **Scott:**

$$n\_intervalos = \frac{\text{máx}(x) - \text{mín}(x)}{\frac{3{,}5\sigma}{\sqrt[3]{n}}} \; ; \; ancho = \frac{3{,}5\sigma}{\sqrt[3]{n}}$$

- **Freedman-Diaconis:**

$$n\_intervalos = \frac{\text{máx}(x) - \text{mín}(x)}{\frac{2 \cdot IQR(x)}{\sqrt[3]{n}}} \; ; \; ancho = \frac{2 \cdot IQR(x)}{\sqrt[3]{n}}$$

Discretización: Basada en la entropía I

- Como ya hemos mencionado es una técnica descendente y supervisada, que utiliza el concepto de ganancia de información.
- La técnica es muy 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.
     - Es decir, cada posible punto de ruptura particiona los datos en dos subconjuntos uno, D_1, con aquellos datos que satisfacen $A \leq$ punto de ruptura y otro, D_2, con los que satisfacen que $A >$ punto de ruptura

Discretización: Basada en la entropía II

2. De todos los posibles puntos de rupturas cogemos aquel que produzca una partición con la *ganancia mímina de información*, es decir, mín$(I(A, D))$.

$$I(A, D) = \frac{|D_1|}{|D|} Ent(D_1) + \frac{|D_2|}{|D|} Ent(D_2), \text{ con } Ent(S) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

3. donde m es el número de clases y p_i la probabilidad de que una instanica pertenezca a la clase i.
4. 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,
   - que se alcance el número de intervalos deseados.

Discretización: Fusión basada en el test χ^2 I

- Como ya hemos mencionado es una técnica ascendente y supervisada.
- La idea básica consiste en ir fusionando intervalos adyacentes que presenten una distribución de clases parecida.
- 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 continua de forma recursiva hasta que se alcanza algún criterio de parada:

Discretización: Fusión basada en el test χ^2 II

- 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.
- se alcance el número de intervalos deseados.
- cuando la frecuencia relativa de las distintas clases en cada intervalos presenta diferencias mayores a un determinado umbral.

Discretización: Analisis de clusters

- Podemos utilizar un algoritmo de clustering para discretizar un atributo númerico.
- Sólo haría falta asociar una categoría a cada grupo o cluster.
- 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.

De variables categóricas a numéricas

- **Variable categórica:** 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 sólo admiten variables numéricas
- **Técnicas**:
  - Codificación ordinal
  - Codificaicón 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 $\{5, 4, 3, 2, 1, 0\}$
- !! Cuidado con hacer esta transformación en la variable a predecir !!
  - Podemos estar prediciendo valores entre las categorías, p.e. 4.5, que puede no tener sentido
  - En la mayoría de los casos la variable a predecir se pude (y debe) mantener como categórica.

Codificación One-Hot

- Se aplica a las variables categóricas nominales.
  - No existe una relación entre las categorías.
  - La anterior codificación no tiene sentido aplicarla porque estaríamos imponiendo el orden.
- Procedimiento:
  - Se crea una nueva variable binaria para cada categoría.
  - Cada nueva variable tomará el valor 1 si está presenta la categoría, 0 en caso contrario

Codificación One-Hot: Ejemplo

- Nacionalidad $=\{Alemana, Francesa, Italiana, Portuguesa\}$

id	Nacionalidad
i_1	Alemana
i_2	Portuguesa
i_3	Italiana
i_4	Francesa

id	Nac_Ale.	Nac_Fra.	Nac_Ita	Nac_Por.
i_1	1	0	0	0
i_2	0	0	0	1
i_3	0	0	1	0
i_4	0	1	0	0

- ¿Qué problema plantea esta codificación?

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[1]
  - 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 variables creadas.
- Este tipo de codificación es el ideal para el caso de dos categorías

[1] Ver a partir de la 82 del libro An Introduction to Statistical Learning para ver las implicaciones de las variables *dummy* en una regresión lineal.

Codificación One-Hot: Ejemplo

- Nacionalidad $=\{Alemana, Francesa, Italiana, Portuguesa\}$

id	Nacionalidad
i_1	Alemana
i_2	Portuguesa
i_3	Italiana
i_4	Francesa

id	Nac_Fra.	Nac_Ita	Nac_Por.
i_1	0	0	0
i_2	0	0	1
i_3	0	1	0
i_4	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.
- Algunas estrategias para tratar con datos desbalanceados.

Referencias


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.