CURSO ESPECIALIZACIÓN IA Y BIG DATA

MÓDULO: SISTEMAS DE APRENDIZAJE AUTOMÁTICO

U2: PREPARACIÓN DE DATOS

L'FSE inverteix en el teu futur

Fons Social Europeu - FSE

UNIÓ EUROPEA

GENERALITAT VALENCIANA

Conselleria d'Educació, Cultura i Esport

Índice

1.- Introducción

2.- Integración de datos

Los datos a integrar dependen de los objetivos del análisis predictivo, que, a su vez, dependen de los objetivos de negocio.

Ejemplo de análisis:

- Objetivo de negocio: “Reducir colas del supermercado”
- Objetivo de negocio refinado: “Asignar recursos más ajustados en cajas según la afluencia de clientes”.
- Objetivo de minería de datos: “Predecir con antelación la afluencia de clientes de una tienda en cualquier tramo del día”.

2.- Integración de datos

Como resultado de esta fase de definición del problema o de establecimiento de los objetivos del análisis predictivo:

- Lista de objetivos de minería de datos y su correspondencia con los objetivos de negocio.
- Detección de qué datos, tanto internos como externos pueden ser necesarios para cada uno de ellos. **Es necesario RECOPILAR e INTEGRAR los datos**.
- Priorización de la obtención de estos datos según la facilidad de llevarlo a cabo y el impacto en el objetivo de negocio esperado.

.

Dos aproximaciones en la recogida de datos:

- Recoger exclusivamente los datos necesarios para el objetivo u objetivos de análisis en cuestión.
- Crear un repositorio de datos para permitir este y otros análisis posteriores.

2.- Integración de datos

Las primeras fases del KDD (Knowledge Kiscovery in Databases) determinan que las fases sucesivas sean capaces de extraer conocimiento válido y útil a partir de la información original.

Generalmente, la información que se quiere investigar sobre un cierto dominio de la organización se encuentra:

- En bases de datos y otras fuentes muy diversas tanto internas como externas, de diferentes departamentos y de distintas organizaciones ⇒ diversos formatos y estructuras.
- Muchas de estas fuentes son las que se utilizan para el trabajo transaccional diario.

El análisis posterior será mucho más sencillo si la fuente es unificada, accesible (interna) y desconectada del trabajo transaccional.

Fuentes de datos

Generadas por máquinas:

- Datos de sensores: RFID tags, medidores inteligentes, dispositivos médicos, GPSs.
- Logs web: Captura de la actividad de servidores, aplicaciones, redes,…
- Datos de puntos de venta: Códigos de barras, códigos QR,…
- Datos financieros: Muchos datos financieros actualmente salen automáticos a partir de los almacenes, o de fuentes externas.

Generadas por humanos:

- Datos introducidos en máquinas.
- Datos del flujo de clicks (por ejemplo, para conocer el comportamiento de un cliente cuando compra).
- Datos relativos a juegos: Movimientos que hacemos en un juego. Esta idea puede ser trasladada fuera de los juegos.

Fuentes de datos

Ejemplos de datos no estructurados generados por máquinas:

- Imágenes satélite
- Datos científicos: Imágenes sísmicas, datos atmosféricos, datos físicos de fuentes potentes de energía,…
- Fotografías y de video: Seguridad, vigilancia, video de tráfico,…
- Datos de radares y sonares: De vehículos, meteorológicos, perfiles sísmicos oceanográficos,…

Ejemplos de datos no estructurados generados por personas:

- Textos en compañías
- Comentarios de redes sociales
- Datos de móviles
- Contenidos web: Youtube, Flickr, …

2.- Integración de datos

Problemas en la recogida de Información Interna

Información no siempre adecuada:

Datos insuficientes:

Ejemplos: la tarjeta de unos grandes almacenes pide la nacionalidad del cliente, pero no los hábitos alimenticios o el tipo de trabajo (sólo cuenta propia o ajena o parado). Estos factores no podrán entrar en el análisis

Datos incompletos o de baja calidad:

Ejemplos: el formulario permite dejar algunos campos en blanco. Los datos de la tarjeta no se actualizan si el cliente cambia de domicilio, de trabajo, de pareja, …

2.- Integración de datos

Recogida de Información Externa:

Ejemplos:

- Demografías (censo), “páginas amarillas”, psicografías (perfiles por zonas), uso de Internet, información de otras organizaciones…
- Datos compartidos en una industria o área de negocio, organizaciones y colegios profesionales, catálogos, etc.
- Datos resumidos de áreas geográficas, distribución de la competencia, evolución de la economía, información de calendarios y climatológicas, programaciones televisivas deportivas, catástrofes,..
- Bases de datos externas compradas a otras compañías.

2.- Integración de datos

Objetivos:

Conseguir que datos sobre el mismo objeto unifiquen ➡ a través de identificadores externos comunes (por ejemplo DNI)

Datos sobre diferentes objetos permanezcan separados.

Errores clásicos:

Datos de diferentes objetos se unifican (más problemático)

2 o más fuentes de objetos iguales se dejan separadas

Problemas al unificar dos fuente:

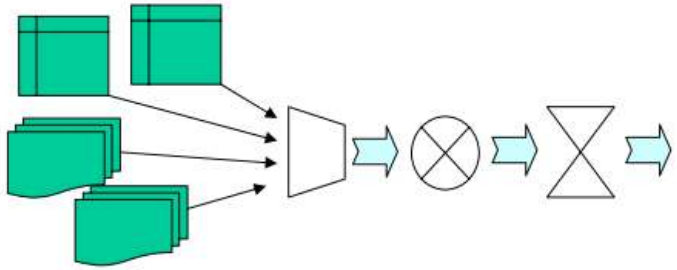
Pueden aparecer datos faltantes

Codificaciones diferentes (por ejemplo: varón y hombre)

3.- Preparación de datos

El objetivo de esta fase es crear un conjunto de datos más significativo y manejable en cuanto a tamaño.
Esta fase lleva mucho tiempo dentro del proceso de análisis predictivo (suele llevar el 50% del tiempo del proyecto)
Objetivo: Obtener la "VISTA MINABLE"

Vista Minable: conjunto de datos que incluyen todas las variables de interés para el problema concreto en el formato adecuado.



VISTA MINABLE

Idc	D-crédito (años)	C-crédito (euros)	Salario (euros)	Casa propia	Cuentas morosas	...	Devuelve-crédito
101	15	60.000	2.200	sí	2	...	no
102	2	30.000	3.500	sí	0	...	sí
103	9	9.000	1.700	sí	1	...	no
104	15	18.000	1.900	no	0	...	sí
105	10	24.000	2.100	no	0	...	no
...	...	...	...	...	...	...	...

3.1.- Comprensión y visualización de datos

Estas dos fases se alternan y complementan.

1r paso: Elaboración de un resumen estadístico de características de los atributos:

- Las herramientas de minería de datos nos proporcionan resúmenes.
- Pueden ayudarnos las técnicas de visualización.

ATRIBUTO	TIPO	# TOTAL	# NULS	# DIST	MEDIA	DESV.	MODA	MIN	MAX
Código postal	Nominal	10320	150	1672	-	-	"46003"	"01001"	"50312"
Sexo	Nominal	10320	23	6	-	-	"V"	"E"	"M"
Estado civil	Nominal	10320	317	8	-	-	Casado	"Casado"	"Viudo"
Edad	Numérico	10320	4	66	42,3	12,5	37	18	87
Total póliza p/a	Numérico	17523	1325	142	737,24€	327€	680€	375€	6200€
Asegurados	Numérico	17523	0	7	1,31	0,25	1	0	10
Matrícula	Nominal	16324	0	16324	-	-	-	"A-0003-BF"	"Z-9835-AF"
Modelo	Nominal	16324	1321	2429	-	-	"O. Astra"	"Audi A3"	"VW Polo"

3.1.- Comprensión y visualización de datos

Los tipos de datos se clasifican en:

- Numéricos o Cuantitativos: Son los atributos continuos. Por ejemplo, la edad.

- Nominales o Cualitativos: Son los atributos discretos. Divididos en:

    - Nominales: Asociados con etiquetas o nombres. Por ejemplo, el trabajo de la paciente.
    - Ordinales: Nominales ordenados. Por ejemplo, Nivel de estudios.
    - Intervalos: Valores ordenados medidos por intervalos iguales. Por ejemplo edad por rango, donde los intervalos van de 5 en 5 años.
    - Ratios: Medidas donde el punto origen está definido en si mismo.

3.1.- Comprensión y visualización de datos

Atributos Nominales: Debemos analizar con detalle cada uno de los atributos. Podemos detectar:

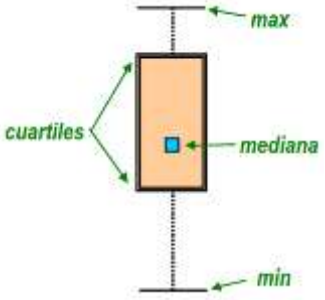
- Valores redundantes. Ej: Hombre,Varón
- Valores despreciables. Ej: agrupar valores como *otros*

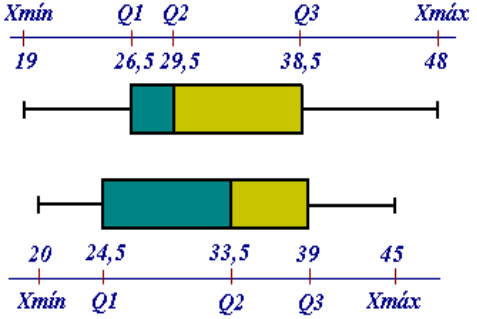
Atributos Numéricos: Debemos analizar con detalle cada uno de los atributos. Podemos detectar:

- Valores anómalos
- Distribuciones en los datos

Podemos usar herramientas gráficas

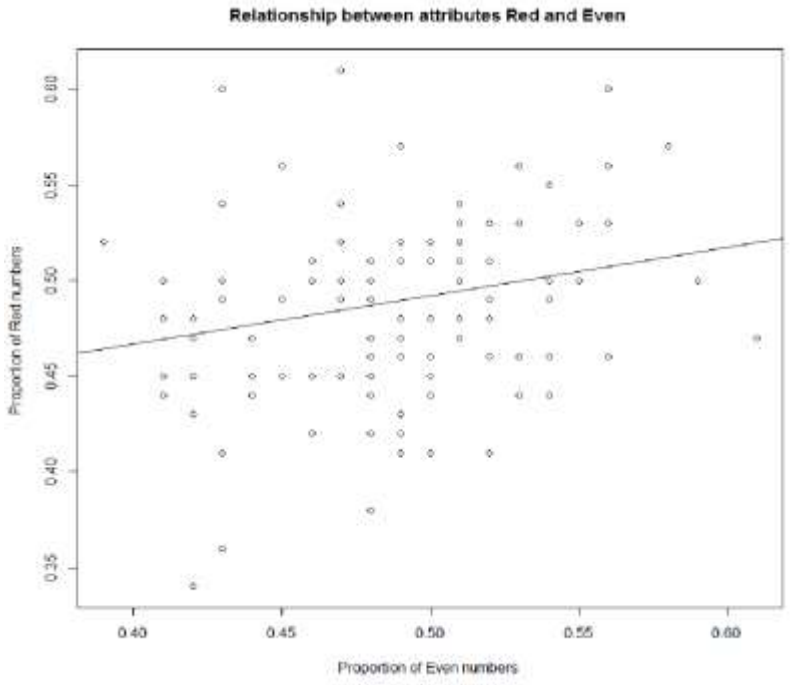
- Los diagramas de caja (***box plot***)
- Gráficas de dispersión



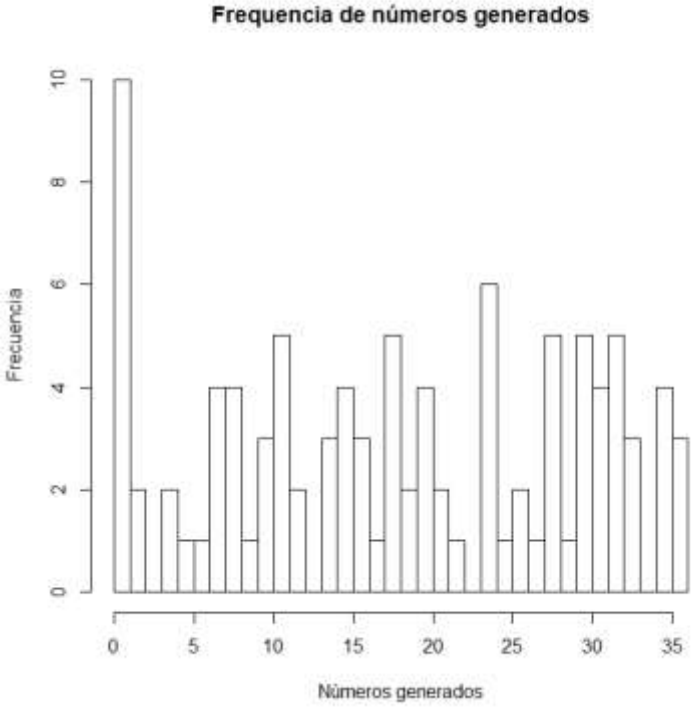


Comprensión y visualización de datos

Gráficas de dispersión



Histograma



Comprensión y visualización de datos

Estudio de correlaciones, asociaciones y dependencias funcionales

$$Cor(\bar{x},\bar{y}) = \frac{Cov(\bar{x},\bar{y})}{\sigma_x \cdot \sigma_y} \qquad Cov(\bar{x},\bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

Análisis estadísticos

Tipos:

- Estudio de la distribución de los datos.
- Estimación de densidad.
- Detección datos anómalos.
- Análisis de dispersión

Comprensión y visualización de datos

Podemos descubrir patrones predictivos utilizando lo que se llama análisis predictivo visual, cuyos objetivos son:

- **Visualización previa**: Aprovechar la capacidad humana para ver patrones, anomalías y tendencias a partir de imágenes y así poder comprender mejor los dalos datos.

- **Visualización posterior**: ayudar al usuario a comprender más rápidamente patrones descubiertos por un sistema KDD.

3.2.- Limpieza de datos (data cleansing)

Se llevan a cabo actividades de detección, eliminación o corrección de instancias corrompidas o inapropiadas en los juegos de datos.

A nivel de valores de atributos se gestionan los valores ausentes, los erróneos y los inconsistentes. Un ejemplo podrían ser los valores fuera de rango (outliers).

El proceso de integración de datos puede ser una de las principales fuentes de incoherencias en los datos. Fruto de la fusión de juegos de datos distintos, se pueden generar inconsistencias que deben ser detectadas y subsanadas.

3.2.- Limpieza de datos (data cleansing)

Se deben eliminar el mayor número posible de datos erróneos o inconsistentes (limpieza)

Muy relacionado con la Calidad de Datos.

Muchas de las operaciones de limpieza requieren el uso de técnicas de minería de datos que permiten entre otras cosas:

- Detectar valores anómalos como son los outliers (datos inconsistentes) y eliminarlos
- Detección de registros de baja calidad (inapropiados, incompletos o erróneos) y su eliminación o relleno.
- Eliminar el ruido

3.2.- Limpieza de datos (data cleansing)

Datos faltantes:

Es importante examinar las razones tras datos faltantes y actuar en consecuencia.

Razones sobre datos faltantes**:**

- Algunos valores faltantes expresan características relevantes. Por ejemplo, la falta de teléfono puede representar en muchos casos un deseo de que no se moleste a la persona en cuestión, o un cambio de domicilio reciente.
- Valores no existentes: muchos valores faltantes existen en la realidad, pero otros no. P. ej.
  - El cliente que se acaba de dar de alta no tiene consumo medio de los últimos 12 meses.
  - Si la paciente no indica la edad de su primera revisión, puede indicar que esta es su primera revisión.
- Datos incompletos: si los datos vienen de fuentes diferentes, al combinarlos se suele hacer la unión externa (outer join) y no la interna (inner join), con lo que muchos datos faltantes representan que esas tuplas vienen de una de las dos fuentes y no existe en la otra.

3.2.- Limpieza de datos (data cleansing)

Acciones ante datos faltantes:

Ignorar: algunos algoritmos son robustos a datos faltantes (p.ej. árboles).
Filtrar (eliminar o reemplazar) la columna: solución extrema, pero a veces existe otra columna dependiente con datos de mayor calidad.
Filtrar la fila: claramente sesga los datos, porque muchas veces las causas de un dato faltante están relacionadas con casos o tipos especiales.
Reemplazar el valor: por medias. A veces se puede predecir a partir de otros datos, utilizando cualquier técnica de ML.
Segmentar: Se segmentan las tuplas por los valores que tienen disponibles. Se obtienen modelos diferentes para cada segmento y luego se combinan.
Modificar la política de calidad de datos y esperar hasta que los datos faltantes estén disponibles.

3.2.- Limpieza de datos (data cleansing)

Opciones para reemplazar un valor faltante:

- Manualmente
- Automáticamente por un valor que preserve la media o la varianza o por el valor de moda.
- Predecirlo:
  - Por clasificación o regresión
  - Hay algoritmos usados para rellenar valores faltantes

3.2.- Limpieza de datos (data cleansing)

Valores erróneos:

Ejemplos: Palabras mal escritas; sensores estropeados o mal calibrados, o incluso datos duplicados.

Para detectarlos podemos usar técnicas de:

- Clustering
- Clasificación
- Regresión

No detectarlos a tiempo puede ser problemático si luego se normaliza.

3.2.- Limpieza de datos (data cleansing)

Acciones ante datos anómalos (outliers):

Ignorar: algunos algoritmos son robustos a datos anómalos (p.ej. Árboles de decisión)
Filtrar (eliminar o reemplazar) la columna: solución extrema, pero a veces existe otra columna dependiente con datos de mayor calidad.
Filtrar la fila: claramente sesga los datos, porque muchas veces las causas de un dato erróneo están relacionadas con casos o tipos especiales.
Reemplazar el valor:

- Por el valor 'nulo' si el algoritmo lo trata bien
- Por máximos o mínimos, dependiendo por donde es el outlier
- Por medias.
- A veces se puede predecir a partir de otros datos, utilizando cualquier técnica de Inteligencia Artificial.

Discretizar: transformar un valor continuo en uno discreto (p.ej. muy alto, alto, medio, bajo, muy bajo) hace que los outliers caigan en 'muy alto' o 'muy bajo' sin mayores problemas.

3.2.- Limpieza de datos (data cleansing)

Inconsistencias:

Un problema grave que afecta a varios métodos de aprendizaje predictivo son los registros inconsistentes, es decir, dos o más registros con los mismo valores en los atributos, pero diferente valor en el atributo clase.

Algunas técnicas no soportan las inconsistencias en los datos. Por lo que se deben eliminar unificando (siempre que se pueda) los registros en una única clase.

Gracias