MINERÍA DE DATOS

Datos

Elizabeth León Guzmán

Research Group on Data Mining – MIDAS
Universidad Nacional de Colombia, Bogotá D.C., Colombia

2021

Población y sus características

Población

- Conjunto de estudio cuyos elementos tienen propiedades en común
- Conjunto bien definido (es posible identificar cuales elementos pertenecen al conjunto y cuales no)
  Ejemplo: Personas inscritas en el curso de "Minería de Datos"

Características de la población: **Variables**

Se pueden medir. En estadística conocidas como Variables
Ejemplo: edad, altura, profesión, género

Población y sus características

Variable, individuo, valor (dato)

$$edad(x) = y$$

- La variable tiene rango
- La variable aleatoria es una función
- La posibilidad de que se tome un x y x tome un valor y, se le llama EVENTO

edad es la variable
x es el individuo u objeto
y es el valor o **dato**

Dato

Hecho individual acerca de algo de interés: númerico, alfanúmerico, etc.

Objetos de Datos

Objetos de Datos

Los objetos de datos se describen mediante una serie de **atributos o variables** que capturan las características básicas de un objeto. Un atributo es una propiedad o característica de un objeto que puede variar de un objeto a otro o de un tiempo dado a otro.

Nombre	Apellido	Edad	Género	Altura	teléfono
Juan	Díaz	25	m	1,70	3562819
María	Martínez	23	f	1,65	9873209
Ana	Ruiz	21	f	1,60	6734636

Conjunto de Datos

Objetos de datos

Otros nombres para un objeto de datos son:
registro, punto, vector, patrón, evento, caso, muestra, observación o entidad

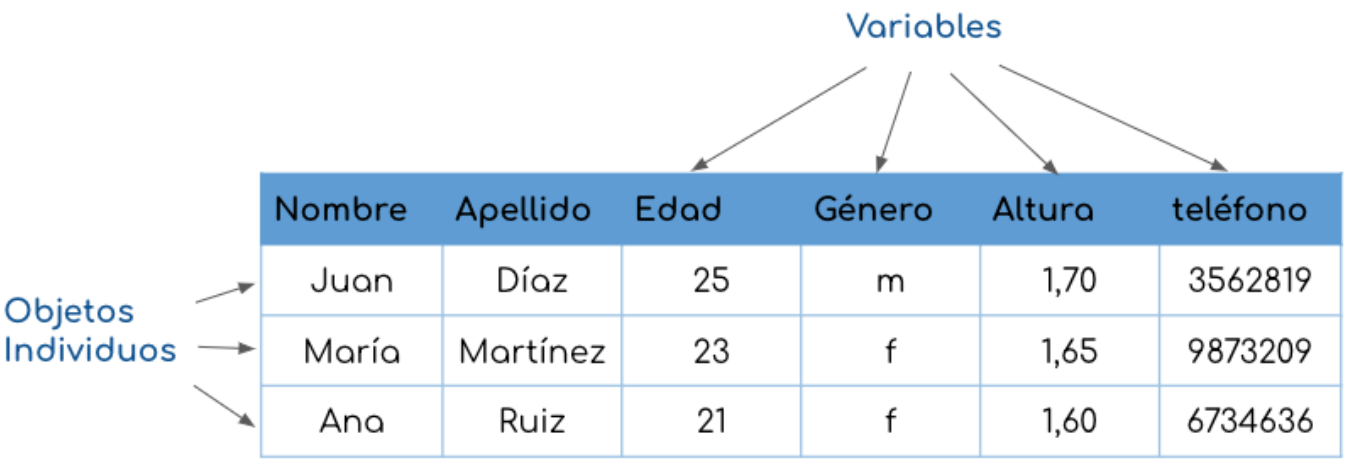
Atributos/Variables

Los atributos también se conocen con el nombre de: *variables, campos, típico, o característica*.

Conjunto de Datos

Conjuntos de datos

Un conjunto de datos puede verse como una colección de objetos de datos.



Nombre	Apellido	Edad	Género	Altura	teléfono
Juan	Díaz	25	m	1,70	3562819
María	Martínez	23	f	1,65	9873209
Ana	Ruiz	21	f	1,60	6734636

La tabla completa es la Población o Muestreo

Conjuntos de Datos

Matriz

Nombre	Apellido	Edad	Género	Altura	teléfono
Juan	Díaz	25	m	1,70	3562819
María	Martínez	23	f	1,65	9873209
Ana	Ruiz	21	f	1,60	6734636

Atributos y escalas de medición

Escalas de medición

En el nivel más básico, los atributos no son números o símbolos. Sin embargo, para discutir y analizar con mayor precisión las características de los objetos, se les suele asignar números o símbolos. Para hacer esto de una manera bien definida, se necesita una **escala de medición**. Una escala de medición es una regla (función) que asocia un valor numérico o simbólico con un atributo de un objeto.

$$edad(x) = y$$

Atributos y escalas de medición

Medición

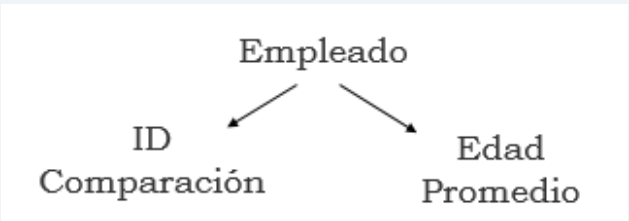
El proceso de medición es la aplicación de una escala de medición para asociar un valor con un atributo particular de un objeto específico. Por ejemplo:

- Usar una báscula para determinar el peso de alguien
- Clasificar a alguien por su género
- Contar el número de sillas en una habitación para saber si hay suficientes para los invitados a una reunión

Tipos de atributos

Los valores utilizados para representar un atributo pueden tener propiedades que no son propiedades del atributo en sí, y viceversa.

Ejemplo:



Aunque ambos atributos son **enteros**, cada uno tiene propiedades particulares que no son especificadas por el valor. A ambos atributos se les podría calcular el promedio, pero no tiene sentido calcular el promedio para ID del empleado. El tipo de dato entero no tiene restricción que evite calcular el promedio.

Propiedades de los valores de los atributos

El tipo de un atributo debe indicar qué propiedades del atributo se reflejan en los valores utilizados para medir. Una forma de especificar el tipo de un atributo es identificar las propiedades que debe describir el atributo. Estas son:

Distinción	$= \text{y} \neq$
Orden	$<, \leq y >, \geq$
Suma	$+ \text{y} -$
Multiplicación	$* \text{y} /$

Tan et al. (2005)

Tipos de atributos

Dadas las propiedades anteriores, se pueden definir cuatro tipos de atributos:

Atributo	**Descripción**	**Ejemplo**	**Operación**
Nominal	Proporcionan información para distinguir un objeto de otro. ($=$, $\neq$)	Códigos postales, números de identificación de empleados, color de ojos	Moda

Tan et al. (2005)

Tipos de atributos

Atributo	Descripción	Ejemplo	Operación
Ordinal	Proporcionan información para ordenar objetos. (<,>)	Edades (niño, adoslecente, adulto, mayor) notas, números de la calle	Mediana, percentiles, rango de correlación
Interval	Destaca diferencias entre valores +, -	Fechas calendario, Temperatura en grados Celsius o Fahrenheit	Media, desviación estándar, correlación de Pearson, prueba t y F

Diferentes tipos de atributos

Atributo	Descripción	Ejemplo	Operación
Proporción/Razón	Son importantes tanto las diferencias como las relaciones de proporción. (*, /)	Temperatura en grados Kelvin, cantidades monetarias, cuentas, edad, masa, longitud, corriente eléctrica	Media geométrica, media armónica, variación porcentual

Los atributos nominales y cardinales se conocen como **categóricos o atributos cualitativos**; los dos tipos restantes de atributos se denominan **atributos cuantitativos o numéricos**

Diferentes tipos de atributos

Ejemplo

- **Nominal**

  Ej: números de identificación, color de ojos, códigos postales
- **Ordinal**

  Ej: Existe orden: humedad {alta, media, baja}, altura {alto, bajo a medio}, edad.
- **Intervalo**

  Ej: las fechas del calendario, las temperaturas en grados Celsius o Fahrenheit.
- **Radio** (Proporción)

  Ej: temperatura en grados Kelvin, la duración, hora, recuentos

Descripción de atributos por número de valores

Una forma independiente de distinguir entre atributos es por el número de valores que pueden tomar:

DISCRETOS:

- Tiene sólo un conjunto finito o infinito numerable de valores.
- Suelen representarse usando variables enteras.
- Ejemplos: códigos postales, cuentas, o el conjunto de las palabras en una colección de documentos

Nota: Los atributos binarios son un caso especial de los atributos discretos. Ejemplo: verdadero o falso, sí o no, mujer u hombre.

Descripción de atributos por número de valores

CONTINUOS:

- Es aquel cuyos valores son números reales.
- Suelen representarse como variables de punto flotante.
- Prácticamente, los valores reales sólo se puede medir y representar mediante un número finito de dígitos.
- Ejemplos: temperatura, altura o peso.

Descripción de atributos por número de valores

- Cualquiera de los tipos de escala de medición (nominal, ordinal, intervalo o proporción) se puede combinar con cualquiera de los tipos en función del número de valores de atributo (binario, discreto y continuo). Sin embargo, algunas combinaciones ocurren con poca frecuencia o no tienen mucho sentido.
- Generalmente, los atributos nominales y ordinales son binarios o discretos, mientras que los atributos de intervalo y proporción son continuos.

Características de los conjuntos de datos

Existen tres características que aplican a muchos conjuntos de datos y tienen un impacto significativo en el técnicas de minería de datos que se utilizan:

- Dimensionalidad
- Dispersión
- Resolución

Características de los conjuntos de datos

1. DIMENSIONALIDAD

- Se define como el número de atributos que poseen los objetos en el conjunto de datos.
- Las dificultades asociadas con el análisis de datos de alta dimensión a veces se denominan *la maldición de la dimensionalidad*.
- Debido a esto, una motivación importante en el preprocesamiento de los datos es la reducción de la dimensionalidad.

Características de los conjuntos de datos

Ejemplo de Alta Dimensionalidad: Matriz de frecuencia de términos en documentos

La dimensionalidad es tan grande como el vocabulario.

	equipo	entrenador	balón	partido	ganador	perdedor	estadio	jugador	perdedor	...	barra
Documento 1	3	0	5	1	1	1	0	0	1	...	0
Documento 2	0	0	0	0	3	4	2	0	4	...	1
Documento 3	1	1	0	1	2	2	0	4	2	...	0
Documento 4	2	2	4	3	0	0	4	3	0	...	0

Características de los conjuntos de datos

2. DISPERSIÓN

- Se refiere a la propiedad de los conjuntos de datos cuando los valores de variables toman demasiados ceros 0.
- Para algunos conjuntos de datos la mayoría de los atributos de un objeto tienen valores de 0; en muchos casos, menos del 1% de las entradas son distintas de cero.
- Cuando los datos son dispersos, hay una ventaja en el almacenamiento, porque generalmente solo los valores distintos de cero son almacenados y manipulados. Esto resulta en ahorros significativos con respecto al tiempo de cómputo y almacenamiento.

Características de los conjuntos de datos

Ejemplo de dispersión: frecuencia de términos

La presencia de términos en documentos cortos puede ser 0, cuando el vocubulario es muy grande

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

1 si la obra contiene la palabra, en otro caso 0

Ejemplo de Shakespeare, tomado de IR de Manning

Características de los conjuntos de datos

3. RESOLUCIÓN

- Las variables pueden tomar valores diferentes de acuerdo con el nivel de resolución (o granularidad de los datos)
- Es posible obtener datos a diferentes niveles de resolución y, a menudo, las propiedades de los datos son diferentes a diferentes resoluciones.
- Los patrones en los datos también dependen del nivel de resolución. Si la resolución es demasiado *fina*, un patrón puede no ser visible o puede esconderse detrás del ruido; si la resolución es demasiado *gruesa*, el patrón puede desaparecer.

Características de los conjuntos de datos

Ejemplo de resolución: Variaciones en la *presión atmosférica*

- En una escala de horas reflejan el movimiento de tormentas y otros sistemas climáticos.
- En una escala de meses, tales fenómenos no son detectables.

Tipos de conjuntos de datos

- Datos de registro
  - Matriz de datos
  - Matriz de datos escasos
  - Datos transaccionales
- Datos basados en grafos.
  - Datos con relaciones entre objetos
  - Datos con objetos que son grafos
- Datos ordenados
  - Datos de secuencia
  - Datos temporales
  - Datos espaciales
  - Datos de series de tiempo

Datos de Registro

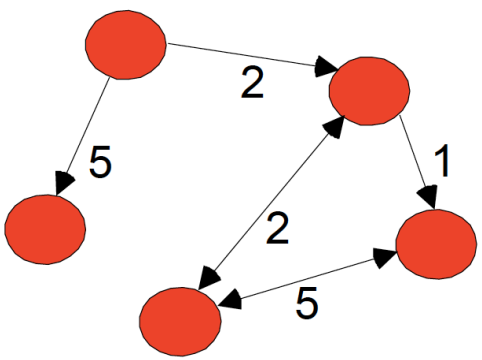
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

	equipo	entrenador	balón	partido	ganador	perdedor	estadio	jugador	perdedor
Documento 1	3	0	5	1	1	1	0	0	1
Documento 2	0	0	0	0	3	4	2	0	4
Documento 3	1	1	0	1	2	2	0	4	2
Documento 4	2	2	4	3	0	0	4	3	0

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

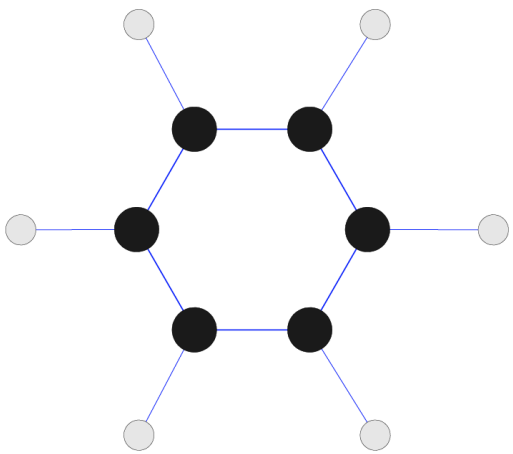
Tan et al. (2005)

Datos basados en grafos



Tan et al. (2005)

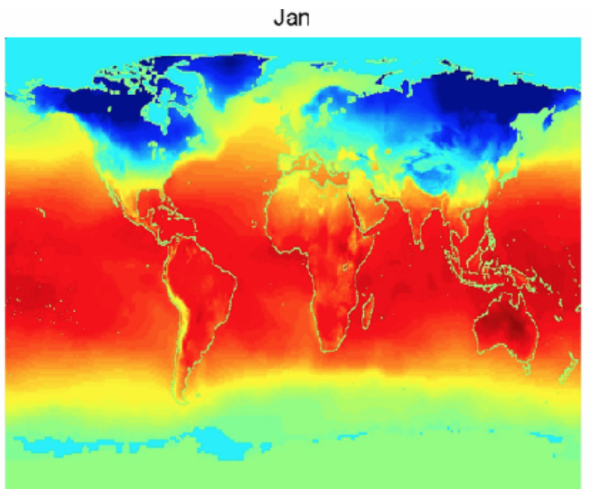
```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```



Datos ordenados

GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

Tan et al. (2005)



Calidad de los Datos

Problemas que afectan la calidad de datos:

- Errores humanos
- Limitaciones en los dispositivos de medida
- Errores en el proceso de colectar los datos: Valores o objetos enteros sin datos, datos duplicados, inconsistencias en los datos.

Ejemplo de inconsistencia de datos:

una persona puede tener 2 m de altura y un peso de 2 Kg

Calidad de los Datos

Error de medición

Problema que resulta del proceso de medición. El valor medido difiere del valor real. Para atributos continuos la diferencia entra los dos valores (real y medido) se le llama **error**

Error de la collección de datos

Se refiere a errores como omitir objetos o valores de atributos, o incluir inapropiadamente

Calidad de los Datos

Datos con poca calidad afectan considerablemente el trabajo en el preprocesamiento de datos.

> ***"The most important point is that poor data quality is an unfolding disaster"***
> Thomas C. Redman, DM Review, August 200

Ejemplo:

Modelo de predicción de asignación de crédito a personas construido con datos de mala calidad, puede:

- Negar créditos a personas potencialmente buenos clientes.
- Asignar créditos a personas que no necesariamente vayan a pagar el crédito.

Calidad de los Datos

Existen técnicas que permiten detectar y corregir errores. Por ejemplo: un error por teclado del computador (mal digitalización) puede ser corregido con programas que detectan estos errores y luego los corrigen (puede ser con intervención humana).

El Proceso para aplicar Minería de Datos trata:

- La detección y corrección de la calidad de los datos: *"Data Cleaning"*
- Usar algoritmos que puedan tolerar una calidad de datos *pobre*

Calidad de los Datos

La Calidad de los datos esta relacionada con el Proceso de medición y colección de los datos. Problemas que se generan en la medición y colección de los datos:

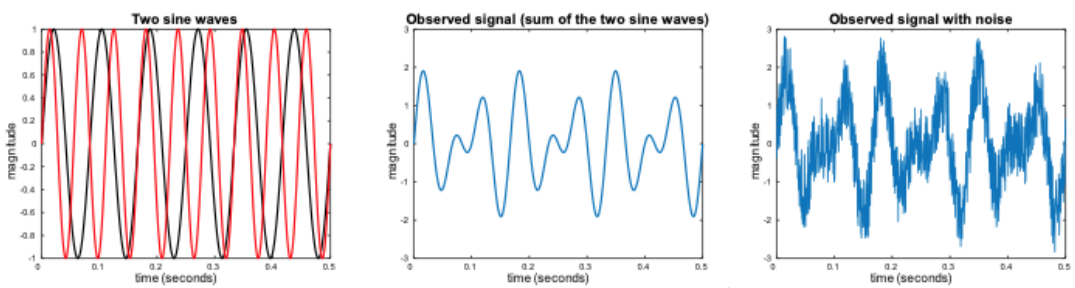
- Ruido y *"outliers"*
- Valores pérdidos
- Datos duplicados
- Datos incorrectos (no confiables)

Calidad de los Datos

RUIDO

Componente aleatorio de un error de medida. Puede ser:

- Para Objetos
  Objeto extraño (adición de un objeto falso o "spurious")
- Para dimensiones (atributos)
  Modificación de valores originales



Source: Tan et al. (2005)

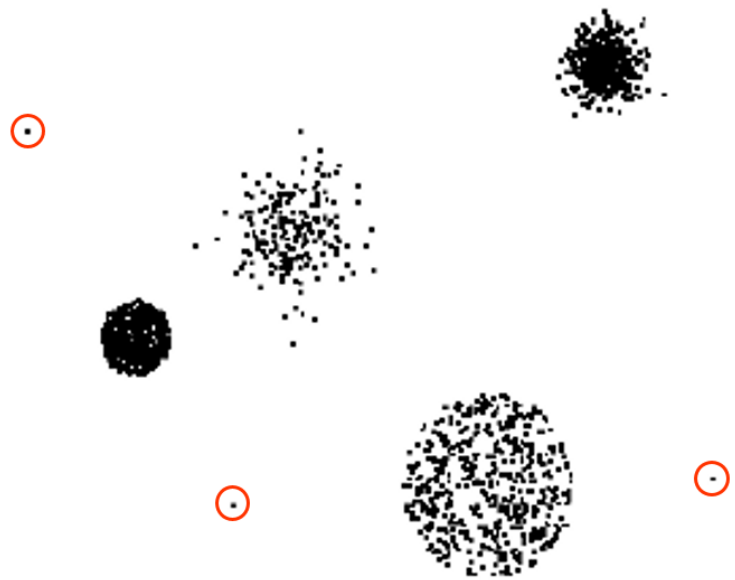
Calidad de los Datos

"OUTLIERS" O VALORES ATÍPICOS

Objetos que poseen caraterísticas completamente diferentes que la mayoría de objetos del conjunto de datos

Detección de "Outliers"

- Ruido que interfiere con el análisis.
- Objetivo del análisis: Detección de Fraude, Detección de Intrusos.



Source: Tan et al. (2005)

Calidad de los Datos

VALORES PERDIDOS O FALTANTES

- No recolectados (personas se pueden negar a darlos)
- Atributos que no se aplican a todos los objetos (impuestos no son aplicables para niños)

Existen varia estrategías para lidear con datos faltantes, cada uno es adecuado en ciertas circunstancias.

Calidad de los Datos

DATOS DUPLICADOS

Es necesario tener cuidado al identificar datos duplicados:

- Si hay dos o más objetos representando el mismo objeto, pero los valores en un atributo difieren.
- Si hay datos que son siilares pero no son el mismo (dos personas con el mismo nombre)

Una vez detectado que el objeto tiene objetos redundantes, se pueden eliminar los objetos dejando solamente uno.

Calidad de los Datos

VALORES INCONSISTENTES

- Algunas veces fácil de detectar (valores negativos)
- Necesario consultar fuentes externas

Una vez detectado el error puede ser corregido. Puede requerir información adicional yo generar información redundante.

References I

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining, (first edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.