



Módulo Minería de Datos Diplomado

**Por
Elizabeth León Guzmán, Ph.D.
Profesora
Ingeniería de Sistemas
Grupo de Investigación MIDAS**

Este documento se desarrolló a partir de otras fuentes que se encuentran

citadas tanto dentro del contenido como en los espacios reservados para la bibliografía.

Si usted es autor de los documentos que se tomaron como bibliografía y

considera que las referencias a su trabajo no están adecuadamente descritas, por favor comuníquese con la profesora Elizabeth León Perdomo a través del correo electrónico: eleonguz@unal.edu.co.

Agenda

- 1.**Datos
- 2.**Preprocesamiento
- 3.**Análisis Exploratorio

Datos



Atributos

- Atributo es una propiedad o característica de un objeto
Ejemplos: color de ojos de una persona, temperatura, etc
- Atributo es también conocido como variable, campo, típico, o característica
- Una colección de atributos describen un objeto
Objeto también se conoce como registro, punto, caso de la muestra, entidad o

Objetos

Atributos

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

Atributos

- Valores son números o símbolos asignados a un atributo

Mismo atributo puede asignarsele diferentes medidas

Ejemplo: altura se puede medir en pies o metros

- Los diferentes atributos se pueden asignar a un mismo conjunto de valores (dominio)

Ejemplo: valores de los atributos de identidad y la edad son números enteros, pero las propiedades de los valores de los atributos pueden ser diferentes:

ID no tiene límite

edad tiene un valor máximo y mínimo

Tipos de Atributos

- **Nominal**

Ejemplos: números de identificación, color de ojos, códigos postales

- **Ordinal**

Ejemplos: las clasificaciones (por ejemplo, el sabor de las patatas fritas en una escala de 1-10), los grados, la altura en {alto, bajo a medio,}

- **Intervalo**

Ejemplos: las fechas del calendario, las temperaturas en grados Celsius o Fahrenheit.

- **Radio (Proporción)**

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

Propiedades de los valores de los atributos

El tipo de un atributo depende de las siguientes propiedades:

- Distinción: = !=
- Orden: <>
- Suma: + -
- Multiplicación: * /

Nominal: distinción

Ordinal: claridad y orden

Intervalo: distinción, orden y adición

Radio: las 4 propiedades

Atributo Tipo	descripción	Ejemplos	Operación
Nominal	Los valores de un atributo nominal son sólo nombres diferentes. Los atributos nominales proporcionan información sólo lo suficiente para distinguir un objeto de otro. (=, !=)	códigos postales, números de identificación de empleados, color de ojos, el sexo: {hombre, mujer}	moda, la correlación de la entropía, la contingencia
Ordinal	Los valores de un atributo ordinal proporcionan información para ordenar objetos. (<,>)	Edades (niño, adoslecente, adulto, mayor) notas, números de la calle	Mediana, percentiles, rango de correlación
Interval	Para los atributos de intervalo, las diferencias entre los valores son significativas. Una unidad de medida existe. {+, -}	las fechas del calendario, la temperatura en grados Celsius o Fahrenheit	media, desviación estándar, la correlación de Pearson, prueba de t y F
Ratio	Para las variables de relación, tanto las diferencias y las relaciones son significativas. (*, /)	temperatura en grados Kelvin, las cantidades monetarias, cuenta, edad, masa, longitud, la corriente eléctrica	media geométrica, media armónica, la variación porcentual

Atributos discretos y continuos

□ Discreto

- Tiene sólo un conjunto finito o infinito numerable de valores
- Ejemplos: códigos postales, cuentas, o el conjunto de las palabras en una colección de documentos
- A menudo representado como variables enteras.
- Nota: Los atributos binarios son un caso especial de los atributos discretos

□ Continuo

- Tiene los números reales como valores de atributos
- Ejemplos: temperatura, altura o peso.
- Prácticamente, los valores reales sólo se puede medir y representar mediante un número finito de dígitos.
- Los atributos continuos se suelen representar como variables de punto flotante.

Los tipos de conjuntos de datos

□ Registro

- Matriz de datos
- Datos del documentos (Espacio vectorial)
- Datos transaccionales

□ Gráfico

- World Wide Web
- Estructuras moleculares

□ Ordenado

- Datos espaciales
- Datos temporales
- Datos secuenciales
- Datos de secuencia genética

Características importantes de datos estructurados

Dimensionalidad

La maldición de la dimensionalidad

Escasez

Sólo cuenta con la presencia

Resolución

Patrones dependen de la escala

Conjunto de Datos: Registro

Colección de registros, cada uno de los cuales consta de un conjunto fijo de atributos

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

Conjunto de Datos: Matrix

- Si los objetos de datos tienen el mismo conjunto fijo de atributos numéricos, y después los objetos de datos se puede considerar como puntos en un espacio multidimensional, donde cada dimensión representa un atributo distinto

Tal conjunto de datos puede ser representado por una matriz m por n, donde hay m filas, una para cada objeto, y n columnas, una para cada atributo

- Deshacer cambios

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Conjunto de Datos: Documentos

**Cada documento se convierte en un “Vector de términos”,
cada término es un componente (atributo) del vector,
el valor de cada componente es el número de veces que el
término correspondiente se produce en el documento.**

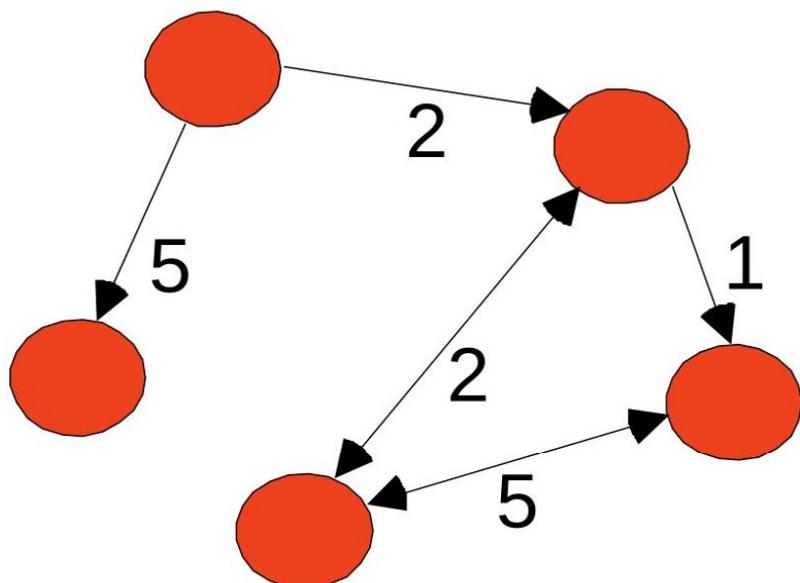
Conjunto de Datos: Transacción

Un tipo especial de datos de registro, donde cada registro (transacción) consiste en un conjunto de elementos.

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Conjunto de Datos: Grafos

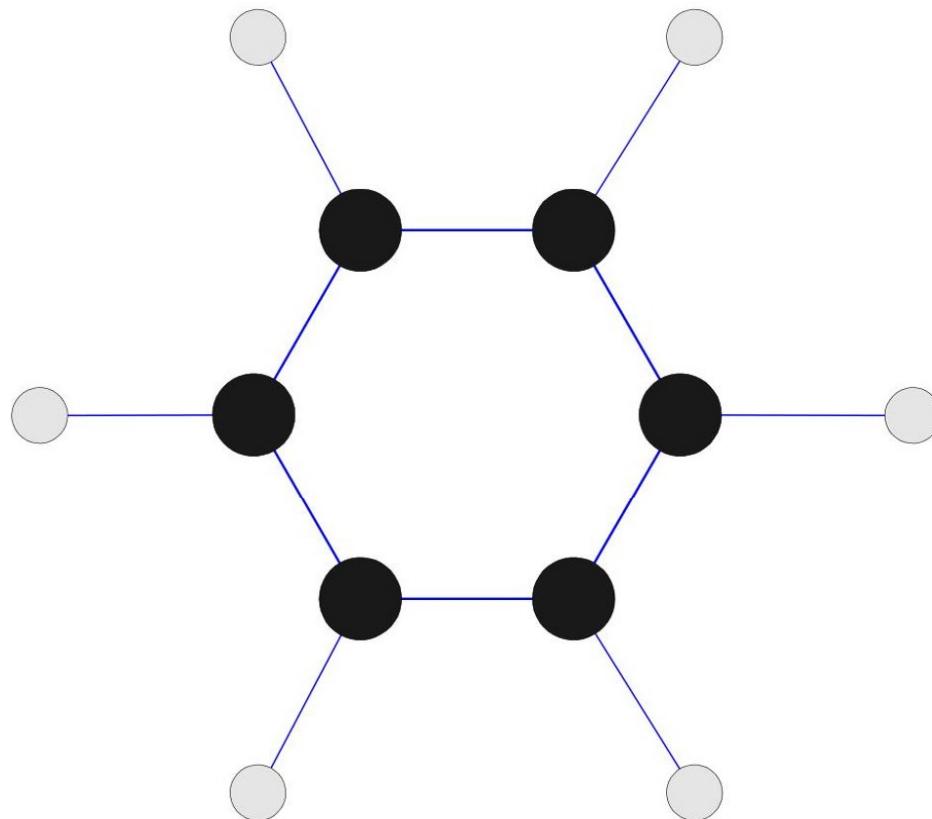
Ejemplos: gráfico genérico y enlaces HTML



```
<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
```

Conjunto de Datos: Moleculas

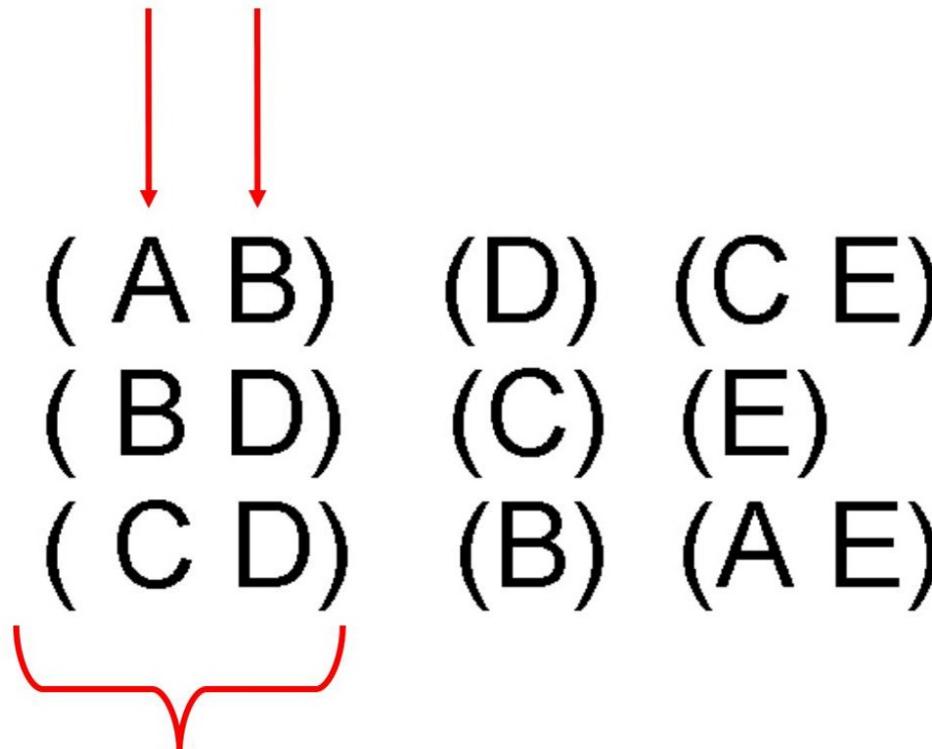
■ Molécula de benceno: C₆H₆



Conjunto de Datos: secuencia

□ Las secuencias de las operaciones

Items/Events



Conjunto de Datos: secuencias

I Los datos de la secuencia genómica

```
GGTTCCGCCTTCAGCCCCGCGCC  
CGCAGGGCCC GCCCGCGCCGTC  
GAGAAGGGCCC GCCTGGCGGGCG  
GGGGGAGGC GGGGCCGCCGAGC  
CCAACCGAGTCCGACCAGGTGCC  
CCCTCTGCTCGGCCTAGACCTGA  
GCTCATTAGGC GGCA GCGGACAG  
GCCAAGTAGAACACGCGAAGCGC  
TGGGCTGCCTGCTGCGACCAGGG
```

Conjunto de Datos:

■ Espacio-temporales de datos

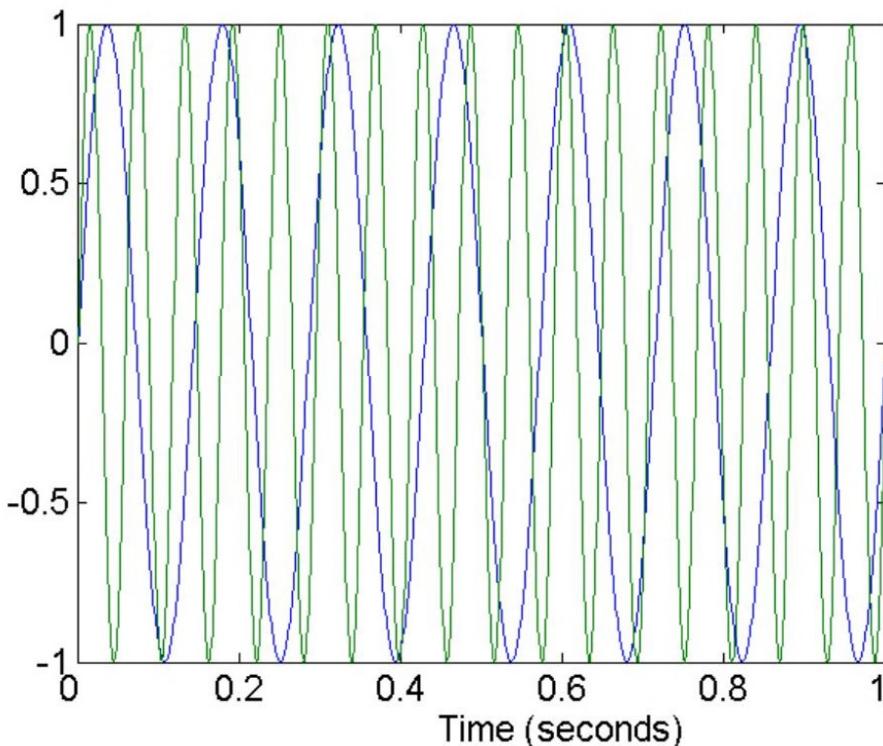
**Temperatura
media mensual
de la tierra y el
mar**

Calidad de los datos

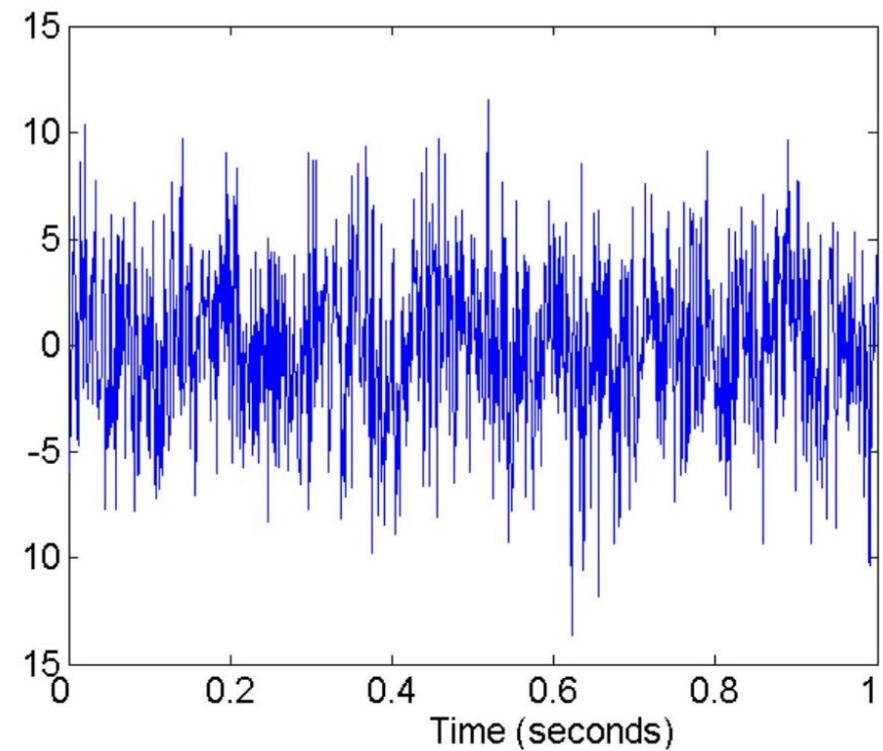
- ¿Qué tipos de problemas de calidad de datos?
- ¿Cómo podemos detectar problemas con los datos?
- ¿Qué podemos hacer acerca de estos problemas?
- Ejemplos de problemas de calidad de datos:
El ruido y los valores atípicos
los valores perdidos
duplicar los datos

Ruido

- El ruido se refiere a la modificación de los valores originales
Ejemplos: la distorsión de la voz de una persona cuando se habla por un teléfono pobre y "nieve" en la pantalla de



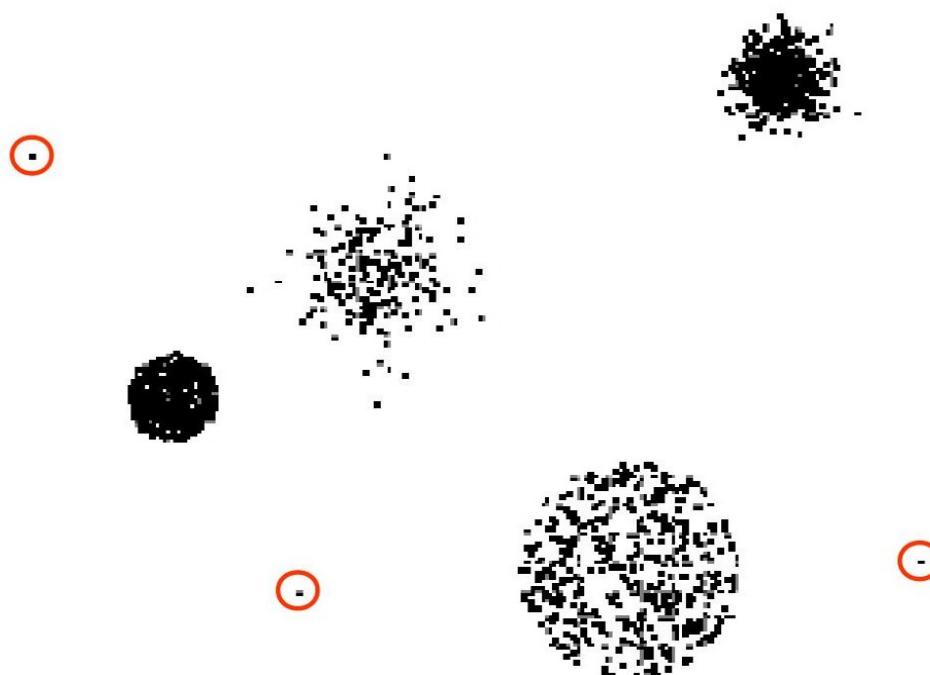
Two Sine Waves



Two Sine Waves + Noise

Valores atípicos “Outliers”

- Los valores extremos son objetos con características que son considerablemente diferentes que la mayoría de los otros objetos en el conjunto de datos



Valores perdidos

- La información no se recoge
(Ejemplo, las personas se pueden negar a dar su edad y peso)
- Los atributos no pueden ser aplicables a todos los casos
(Ejemplo, el ingreso anual no es aplicable a los niños)
- Manejo de los valores perdidos
 - Eliminar los objeto
 - Estimar los valores perdidos
 - Ignorar el valor perdido durante el análisis
 - Reemplazar con posibles valores (ponderados por sus probabilidades)

Datos Duplicados

- Conjunto de datos pueden incluir objetos de datos que son duplicados,
 - procedentes de fuentes heterogeneas
 - Ejemplos:
 - La misma persona con múltiples direcciones de email
 - Limpieza
 - Proceso de lidiar con los problemas de datos duplicados

Datos Duplicados

- Conjunto de datos pueden incluir objetos de datos que son duplicados,
 - procedentes de fuentes heterogeneas
 - Ejemplos:
 - La misma persona con múltiples direcciones de email
 - Limpieza
 - Proceso de lidiar con los problemas de datos duplicados

Bibliografia

Introduction to Data Mining. Tan, Steinbach, Kumar. 2006