

COMP 6315 Minería de Datos

CLASE 2: TIPO de Datos
(basada en notas del Prof V. Kumar)

Dr. Edgar Acuna
Departamento de Matematicas

Universidad de Puerto Rico- Mayaguez
academic.uprm.edu/eacuna/comp6315.html

Tipos de Datos

- **Datos estructurados:** Son datos faciles de organizar y usualmente estan disponible en bases de datos. Solo cuenta por un 20% de los datos disponibles actualmente. Datos de sensores, datos de llamadas telefonicas, datos de tarjetas de credito, datos de cajeros automaticos.
- **Datos No estructurados:** No estan organizados en un formato de estructura predefinida. No estan almacenados en una base de datos relacional. Puede ser textual o non-textual, y generado por humano o maquina. Generalmente son datos provenientes de las redes sociales: Twitter, Facebook, LinkedIn, etc. Tambien mensajes de texto por e-mail, imagenes, videos, files de audio, etc.
- Puede ser almacenado dentro de una base de datos no-relacional como NoSQL.

Tipos de datos (cont)

También hay **datos semi-estructurados**, en donde se mantienen tags para identificar elementos de datos por separado, que permite después hacer agrupamientos y jerarquias. Los mensajes Email pueden ser considerados como datos semi-estructurados. La metadata asociada a los email's permite clasificar y hacer búsqueda de los mensajes.

Ejemplos de datos semi-estructurados incluyen documentos en Markup Language XML, en JSON (JavaScript Object Notation) y bases de datos NoSQL (MongoDB, Hive, Cassandra, Couch DB, etc). Estas bases de datos son usadas mayormente en Big Data cuando se transmite en tiempo real datos de aplicaciones web.

En minería de datos se trabaja mayormente con datos estructurados. Data Science trabaja con cualquier tipo de datos. El proceso de convertir datos no estructurados en datos capaces de ser analizados es llamado “Wrangling”.

Que es un conjunto estructurado de datos?

- Es una coleccion de objetos con sus respectivo atributos.
- Un atributo es una propiedad o caracteristica de un objeto.
 - Ejemplos: color de ojos de una persona, peso, salario annual, etc.
 - Un atributo tambien es conocido como variable, caracteristica, o “feature”.
- Una coleccion de atributos describe un objeto.
 - Un objeto es tambien llamado registro, caso, muestra, o instancia.

Atributos

<i>Tid</i>	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

Objetos

Atributos

- Los valores de un atributo son los numeros o simbolos asignados a un atributo.
- Segun la escala de medicion hay cuatro tipos distintos de atributos: nominal, ordinal, de intervalo y de razon.

Propiedades de los valores de atributos

- Los tipos de atributos dependen de si poseen una de las siguientes propiedades.
 - Distincion: $= \neq$
 - Orden: $< >$
 - Adicion: $+ -$
 - Multiplicacion: $* /$
- Atributo Nominal: distincion
- Atributo ordinal: distincion y orden
- Atributo de intervalo: distincion, orden y adicion
- Atributo de razon: posee las 4 propiedades

Tipo de Atributo	Descripcion	Ejemplos	Operaciones
Nominal	Los valores de estos atributos son solamente nombres distintos. O sea los atributos nominales solo dan informacion para distinguir un objeto de otro ($=$, \neq)	Codigos postales, numeros de identificacion, color de ojos, sexo: { <i>male</i> , <i>female</i> }	moda, entropia, medidas de asociacion, pruebas de χ^2 .
Ordinal	Los valores de estos atributos dan suficiente informacion para ordenar objetos. ($<$, $>$)	Nivel de educacion, nivel de empleo, notas, numero de calles	mediana, percentiles, medidas estadisticas noparametricas
De intervalo	Para este tipo de atributos la diferencia entre sus valores tienen significado. Es decir existe una unidad de medicion. ($+$, $-$)	Temperatura, fechas, etc.	Media, desviacion estandar, medidas estadisticas parametricas
De razon	Las diferencias y las razones entre sus valores tienen significado. ($*$, $/$)	Cantidades monetarias, peso, edad, longitud, corriente electrica.	Media geometrica, media armonica, variacion porcentual.

Atributos Continuos y discretos

- Atributos discretos
 - Tiene un numero finito o infinito contable de valores.
 - Ejemplos: numero de carros vendidos por dia, numero de hijos en una familia, numero de veces que aparece una palabra en una coleccion de documentos.
 - Se representan a menudo por numeros enteros.
 - Los atributos binarios son un caso especial de atributos discretos Ejemplo: Pasar, Fracasar.
- Atributos continuos
 - Sus valores que asumen son numeros reales
 - Ejemplos: temperatura, altura, peso,etc.
 - En la practica los numeros reales pueden ser representados por un numero finito de digitos.

Tipos de conjuntos de datos

- **Datos de Registro**
 - Matrices de datos
 - Datos de documentos
 - Datos de transacciones
- **Datos de graficas**
 - Redes
 - Estructuras moleculares
- **Ordenados**
 - Datos espaciales
 - Datos temporales
 - Datos de secuencias geneticas

Datos de Registro

- Datos que consisten de una coleccion de registros, cada uno de los cuales consiste de un conjunto fijo de atributos.

<i>Tid</i>	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

Matriz de datos

- Si todas las instancias tienen el mismo numero de atributos, entonces se puede considerar que las instancias son puntos en un espacio multidimensional, donde cada dimension representa un distinto atributo.
- Este conjunto de datos puede ser representado por una matriz con m filas y n columnas una por cada atributo.

Sexo	edad	Altura	Peso	Colest
1	56	1.70	100	238
0	45	1.60	60	178

Datos de documentos

- Cada documento es como un vector de “terminos”
 - El valor de cada componente (atributo) del vector es el numero de veces que el termino correpondiente aparece en el documento.

	team	coach	pla y	ball	score	game	wi n	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

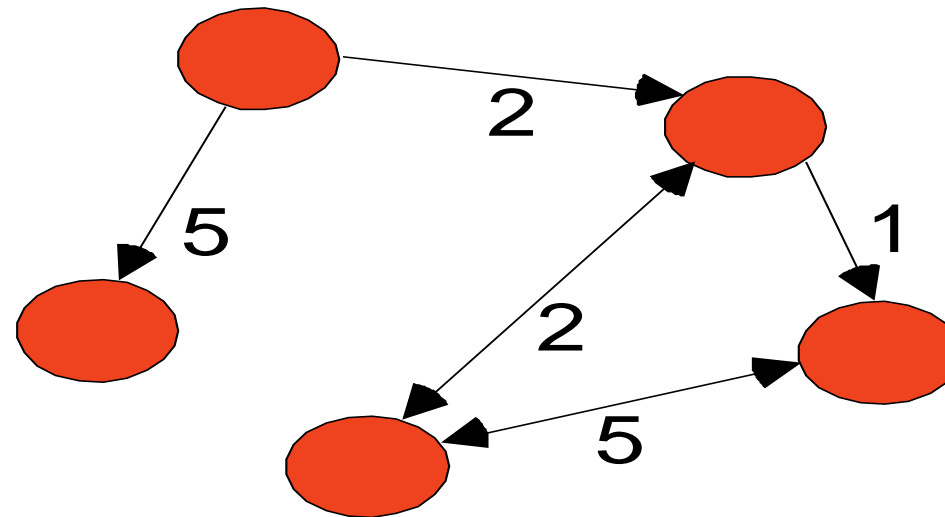
Datos de transacciones

- Cada record (transaccion) envuelve un conjunto de items.
- Por ejemplo, el conjunto de articulos comprados por un cliente cada vez que va a un supermercado constituye una transaccion, mien tras que los productos comprados en cada compra son los items.

<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

Datos de grafos

- Ejemplo



Datos ordenados

- Datos de secuencia genetica

**GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG**

Bases

A=Adenina

C=Citosina

G=Guanina

T=Tianina

Datos ordenados

- Datos tempo-espaciales

Temperatura
mensual de
tierra y oceano

