Tipos y orígenes de datos

Autor
Martín Vladimir Alonso Sierra Galvis

Gestión de Datos, Maestría en Ciencia de Datos
Pontificia Universidad Javeriana Cali

Versión 1.0
Santiago de Cali, febrero de 2023

Tabla de contenido

Tipos de datos

Consideraciones

Antes de revisar las diferentes fuentes desde las que podemos obtener datos, es conveniente recordar que los datos pueden ser clasificados en diferentes tipos de acuerdo a diferentes características que posean. La importancia de los tipos de datos radica en que, si los conocemos, sabremos exactamente cómo tratar y procesar correctamente los datos con los que estemos trabajando.

Según su dominio

Anteriormente, en la lectura La Gestión de Datos en la era de la Ciencia de Datos, vimos que los datos se clasifican en dos categorías principales: datos cuantitativos y datos cualitativos. Este tipo de clasificación es lo que se conoce como clasificación por dominio. El dominio de un dato define cómo es el dato y qué tipo de operaciones se pueden realizar con él. Por ejemplo, si un dato pertenece al dominio cuantitativo, posiblemente se puedan ejecutar operaciones aritméticas con el mismo. Por otro lado, si un dato pertenece al dominio cualitativo, entonces puede ser un texto al que se le pueda aplicar operaciones como búsquedas de subcadenas de texto. Pero ¡ojo! Aquí también es importante que tengamos en cuenta, además de cómo es el dato, qué está representando. Es posible que un dato numérico pertenezca en realidad al dominio cualitativo. Lo podemos reconocer porque lo que representa el dato es una categoría. El ejemplo más común son los números que representan estratos sociales. No tiene ningún sentido aplicar operaciones de suma, resta, multiplicación o división a un estrato social. Es por estas posibles confusiones que existen otras subcategorías, denominadas NOIR.

Nominal

La subcategoría Nominal clasifica los datos en varias categorías diferentes donde no está implícito ningún criterio de clasificación. En palabras más simples, los datos nominales no representan un orden específico o un ranking. Ejemplos: datos que representen géneros o estados maritales; números que representen colores. Esta subcategoría pertenece al dominio cualitativo.

Ordinal

Al contrario que la subcategoría Nominal, la Ordinal cobija a todos aquellos datos que representen un orden o ranking. El ejemplo que más se ajusta a esta subcategoría es el que pusimos anteriormente de los estratos. Esta subcategoría pertenece a la categoría cualitativa.

Intérvalo

La subcategoría Intérvalo agrupa todos aquellos datos que representan un orden y entre ellos hay una distancia o medida significativa pero no tienen un punto cero. Ejemplo: datos que representen años. Esta subcategoría pertenece a la categoría cuantitativa.

Racional

La última categoría, la Racional, contiene todos los datos que representan un orden y entre ellos hay una distancia o medida significativa, y que además tengan un punto cero. Ejemplo: datos que representen salarios. Esta subcategoría pertenece a la categoría cuantitativa.

Además de las subcategorías NOIR, podemos nombrar como subcategorías de la categoría cuantitativa los datos **discretos** y los datos **continuos**. Los datos discretos son aquellos datos numéricos que no pueden tener un valor entre dos consecutivos. Por otro lado los datos continuos son datos numéricos que pueden tener cualquier valor en un intérvalo.

Según su organización

Todas las categorías y subcategorías de dominio son muy utilizadas al momento de realizar tareas como análisis exploratorio de datos, preparación y limpieza de datos, e incluso análisis estadístico. Ahora bien, cuando estamos accediendo a los datos en una fuente, debemos también identificar qué tipo de organización tienen, pues esto puede definir como debemos accederlos y qué herramientas necesitamos para dicho acceso.

Estructurados

Los datos estructurados son aquellos que se ajustan a un modelo lógico, normalmente constituído por tablas. Estas, a su vez, están conformadas por filas y columnas. Este tipo de organización permite agrupar datos de

diferentes tablas para formar relaciones. Los datos estructurados tienen una organización a priori, son fáciles de almacenar, buscar y analizar, pero su estructura es sumamente rígida, lo que hace que muchas veces no sean muy flexibles. Son comúnmente gestionados usando el lenguaje de consultas estructurado SQL. Ejemplos: datos guardados en hojas de cálculo o en bases de datos relacionales.

No estructurados

La mayor parte de los datos con los que interactuamos son datos no estructurados. Estos son datos que no están asociados a ningún modelo lógico de datos por lo que no tienen una organización a priori. Estos datos, al no poseer una organización definida, son más difíciles de administrar, buscar y analizar, aunque son extremadamente flexibles. Los datos estructurados fueron mayormente evitados por las empresas hasta el desarrollo de la Inteligencia Artificial y los algoritmos de Machine Learning. Ahora son considerados sumamente valiosos. Ejemplos: datos en un archivo de texto, contenido en redes sociales, imágenes de satélites, audios o videos.

Semiestructurados

Los datos semiestructurados están a medio camino entre los estructurados y los no estructurados. Aunque no están sujetos a ningún modelo en particular, poseen ciertas características que permiten organizarlos, aunque de una forma menos rígida que, por ejemplo, los datos asociados a un modelo de tablas. Esto les otorga una alta flexibilidad y, en la actualidad, existen herramientas que permiten gestionarlos con relativa facilidad. Ejemplo: datos en formato JSON.

Según su nivel

Además de las anteriores categorías existe una tercera: según su nivel. En palabras más claras, se clasifican según lo que describen y a que grado de descripción llegan. Para entender esta clasificación, se puede pensar en los datos como si fueran piezas de LEGO o muñecas rusas. En esta categoría existen tres tipos principales denominados Microdato, Macrodato y Metadato.

Microdato

Los microdatos son datos que se refieren o describen a un objeto en particular. El ejemplo más común son los datos que se pueden recopilar de una persona en un formulario: nombre, apellido, edad, email, ciudad, dirección, estado civil, etc.

Macrodato

Por el contrario, los macrodatos son datos que se refieren o describen a agrupaciones de objetos. Los macrodatos se construyen a partir de microdatos y son muy utilizados para responder preguntas de análisis. Por ejemplo: colección de personas casadas en determinada ciudad o país.

Metadato

Los metadatos son datos que se encuentran en el nivel más alto de descripción pues su rol principal es describir o explicar otros datos, incluídos los microdatos y los macrodatos. Los metadatos pueden traer ciertas ventajas como facilitar la búsqueda, el análisis, la estandarización y la integración de datos de diferentes formatos y fuentes. Además puede ayudar a brindar mayor seguridad a los datos y mejoras en procesos de cambio y generación de informes. Uno de los ejemplos más conocidos de metadatos son aquellos datos que se agregan a una fotografía digital. Normalmente cuando tomamos fotos con nuestros dispositivos móviles, el mismo dispositivo se encarga de anexar a la foto datos como la hora a la que fue tomada o la ubicación donde se tomó.

Los orígenes de datos

Consideraciones

Simplemente a modo de aclaración: cuando aquí mencionamos el término orígenes de datos, no estamos haciendo alusión a la época en particular cuando se comenzó a utilizar el mismo. Por orígenes de datos nos referimos a los medios, sean físicos o electrónicos, que son capaces de generar datos o en los que podemos encontrar datos para trabajar con ellos.

Archivos

Un archivo es una secuencia de datos almacenados en un medio persistente del dispositivo, específicamente el disco duro. Los archivos son utilizados por un software específico. Cuentan siempre con un nombre, una extensión que define su formato y, por ende, los software que lo pueden utilizar, y una ubicación en el sistema de archivos del sistema operativo. Los archivos son una de las fuentes de datos más comunes y numerosas, a la vez que complejas, pues la variedad de datos que proveen los archivos es inmensa.

Texto

Los archivos de texto son el tipo más común de archivos. Como su nombre lo indica son archivos que solo admiten datos de tipo texto, por lo que se consideran archivos restrictivos. Incluso si escribimos números en uno de estos archivos, dichos números se interpretarán como texto. La ventaja de estos archivos es que son muy fáciles de leer y entender mediante cualquier aplicación de edición de texto. Otra ventaja es que un error en su contenido no implica que el archivo se vuelva corrupto, por lo que se dice que son resistentes a errores. Son de caracter **no estructurado**, pues los datos que se colocan en un archivo de texto no deben cumplir con una organización específica.

JSON

Los archivos de tipo JSON, o **J**ava**S**cript **O**bject **N**otation por sus siglás en inglés, son un tipo especial de archivos de texto en el que los datos se organizan de manera jerárquica, mediante parejas clave-valor. Lo interesante de este tipo de archivos es que son capaces de distinguir los distintos tipos de datos que albergan. Son sencillos de entender y se pueden

abrir con editores de texto o aplicaciones especializadas para leer JSON. Aunque son resistentes a errores, lo cierto es que el formato debe seguir una sintaxis definida, al igual que la organización de los datos, razón por la que se consideran archivos **semiestructurados**. En la actualidad son muy utilizados en el intercambio de datos entre dispositivos conectados, como por ejemplo, un cliente y un servidor. A continuación se muestra una imagen de este tipo de archivo.

CSV

Los archivos CSV son un tipo especial de archivos de texto donde los datos se organizan separándolos por medio de comas. De ahí su nombre **C**omma **S**eparated **V**alue. Lo que se trata de hacer en un archivo CSV es organizar los datos de una manera similar a cómo se organizarían en una tabla. Incluso estos archivos pueden ser exportados o importados en aplicaciones de hoja de cálculo como Excel o LibreOffice Calc, mismas que son reconocidas por trabajar los datos en formato tabular. Los archivos CSV son muy utilizados como datasets sobre los que se realizan procesos de limpieza, transformación y preparación de datos. Se pueden presentar errores al cargar los datos si no se sigue correctamente la separación por comas. Pertenecen a la categoría de datos **semiestructurados**. Si tomamos como ejemplo los datos de la sección JSON, estos se organizarían de la siguiente forma en un CSV

John, Doe, john.doe@gmail.com, 35

Multimedia

Los archivos multimedia hacen parte de una categoría de archivos denominada archivos binarios. La principal característica de los archivos binarios es que solamente los pueden interpretar aquellos softwares especializados para los que fueron creados. Si en algún momento tratamos

de abrir un archivo binario con un editor de texto, solamente veremos un conjunto de símbolos y no lograremos entender nada. A pesar de esto, los archivos binarios y en especial los archivos multimedia contienen gran cantidad de datos que con el procesamiento adecuado, se pueden convertir en información muy interesante. La desventaja es que este tipo archivos no es resistente a errores, pues si se llega a modificar de manera inadecuada, es muy posible que se vuelva corrupto. Son considerados **no estructurados**.

Dispositivos

Al principio de la era del internet y durante muchos años, los datos que se transmitían a través de la red eran datos generados por los seres humanos. Esta situación ha cambiado en los últimos años con el advenimiento del Internet de las Cosas. El Internet de las Cosas es un término que expone el hecho de que, en la actualidad, la mayor parte de los dispositivos que utilizamos están conectados a internet, generando y transmitiendo datos. El ejemplo más conocido es la domótica o uso de dispositivos para implementar hogares inteligentes. Lo interesante del Internet de las Cosas es que ahora son los dispositivos los encargados de generar datos; nosotros como humanos simplemente los almacenamos y los gestionamos. Pero, con esto, aparecen nuevos retos como la gestión del Big Data debido al volumen de datos generados por dichos dispositivos; o la mejora en sistemas de seguridad. Esto último debido a que, en muchas ocasiones, los datos que se generan y transmiten pueden ser sensibles o privados, además del control que estos dispositivos han comenzado a tener sobre tareas y elementos cotidianos como, por ejemplo, la administración de las cerraduras de las puertas de una casa. Los datos generados por estos dispositivos son considerados **no estructurados**.

Bases de Datos

El término Bases de Datos hace referencia a colecciones de información que se encuentra almacenada electrónicamente en un dispositivo. La información se guarda y se consulta sistemáticamente, lo que quiere decir, que dichas acciones deben seguir unas reglas específicas. Es por esto que para poder interactuar con una base de datos es necesario un software especial denominado Sistema Gestor de Bases de Datos. Aunque la base de datos más popular y utilizada es la Base de Datos Relacional, actualmente existen

muchos tipos para fines diversos. Entre las más famosas podemos encontrar las Bases de Datos Orientadas a Documentos, las Bases de Datos Clave-Valor, las Bases de Datos de Familia de Columnas, y las Bases de Datos de Grafos. Las bases de datos son la fuente más conocida de datos **estructurados** y **semiestructurados**.

Repositorios de datasets

Como científicos de datos será muy común trabajar con datasets, conjuntos de datos que, como ya vimos, están organizados en formato CSV. Lo interesante de los datasets es que pueden ser una fuente de datos bastante interesante para alimentar y entrenar algoritmos de Machine Learning. Eso sí, en la mayoría de los casos, tendremos que hacer un proceso minucioso de limpieza y transformación de datos. Estos datasets se pueden encontrar en repositorios en la web. Un claro ejemplo de ello es la aplicación Kaggle, de la que se pueden descargar muchos datasets de manera gratuita. Los datasets son considerados fuentes con datos **semiestructurados**.

Referencias

1. Bagdasarian, B. *Talking Data Part 2: Macro Data vs Micro Data*. Marzo de 2018. [En línea]: disponible en https://blog.hubspot.com/customers/talking-data-part-2-macro-micro-data. [Accedido en 2023].
2. Computerphile. *Data Analysis 1: What is Data? - Computerphile*. Julio de 2019. [En línea]: disponible en https://www.youtube.com/watch?v=SEeQgNdJ6AQ&t=411s. [Accedido en 2023].
3. D. J. Hand. *Microdata, Macrodata and Metadata*. Computational Statistics, Volume 2, pp 325-340. Conference paper, versión Preview. [En línea]: disponible en https://link.springer.com/chapter/10.1007/978-3-642-48678-4_41. [Accedido en 2023].
4. DW Español. *Qué es el internet de las cosas*. Octubre de 2019. [En línea]: disponible en https://www.youtube.com/watch?v=fWrY571yHYI. [Accedido en 2023].
5. Forbes. *What's The Difference Between Structured, Semi-Structured and Unstructured Data?* [En línea]: disponible en https://www.forbes.com/sites/bernardmarr/2019/10/18/whats-the-difference-between-structured-semi-structured-and-unstructured-data/?sh=7e4fa7372b4d. [Accedido en 2023].
6. Geeks For Geeks. *What is Data?* Abril de 2022. [En línea]: disponible en https://www.geeksforgeeks.org/what-is-data/. [Accedido en 2023].
7. Power Data. *¿Qué son los metadatos y cuál es su utilidad?* Marzo de 2016. [En línea]: disponible en https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/que-son-los-metadatos-y-cual-es-su-utilidad. [Accedido en 2023].
8. Universidad Técnica Federico Santa María. *Archivos*. [En línea]: disponible en http://progra.usm.cl/apunte/materia/archivos.html. [Accedido en 2021].