Pontificia Universidad JAVERIANA Cali

Aprendizaje Automático y Análisis de Datos

Conocimiento de los datos

Julián Gil González

julian.gil@javerianacali.edu.co (Periodo 2023-I)

3 de febrero de 2023

Agenda

Temas:

- Conocimiento de los datos.
- Introducción al procesamiento de los datos.

Objetivos del aprendizaje: Al final de esta clase los estudiantes estarán en la capacidad de:

- Explicar cuáles son los aspectos más importantes al momento de procesar una base de datos.

Pontificia Universidad JAVERIANA Cali

Table of Contents

Conociendo los datos I

- El éxito del aprendizaje de máquina depende de la cantidad y calidad de datos con los que se cuenta.
- De hecho el factor clave en los sistemas de aprendizaje profundo ha sido la disponibilidad de grandes cantidades de datos.
- En las aplicaciones reales, los datos provienen de diferentes fuentes y usualmente no han tenido ningún tipo de procesamiento.

Conociendo los datos II

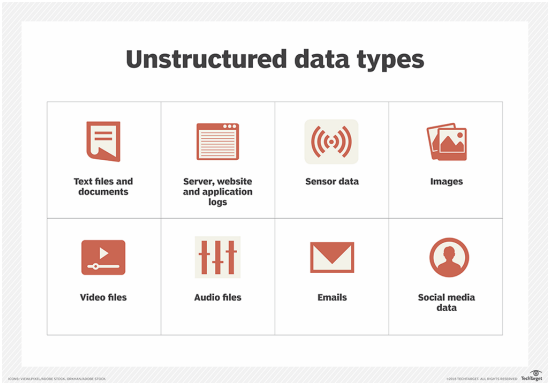
- En este sentido, es imperativo llevar a cabo etapas de conocimiento con el fin de determinar posibles anomalías en los datos.
- Dichas anomalías deben ser abordadas a través de etapas de procesamiento con el fin de tener una base de datos que favorezca al funcionamiento de los sistemas de aprendizaje de máquina.
- Se puede tener el mejor modelo de aprendizaje de máquina y aún así tener resultados pobres. La clave: el procesamiento de los datos.

Tipos de datos: Estructurados

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	populatio
0	-122.23	37.88	41.0	880.0	129.0	322.0
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0
2	-122.24	37.85	52.0	1467.0	190.0	496.0
3	-122.25	37.85	52.0	1274.0	235.0	558.0
4	-122.25	37.85	52.0	1627.0	280.0	565.0

Datos que se almacenan de forma ordenada, por ejemplo, a partir de tablas, bases de datos..

Datos No-estructurados [1]



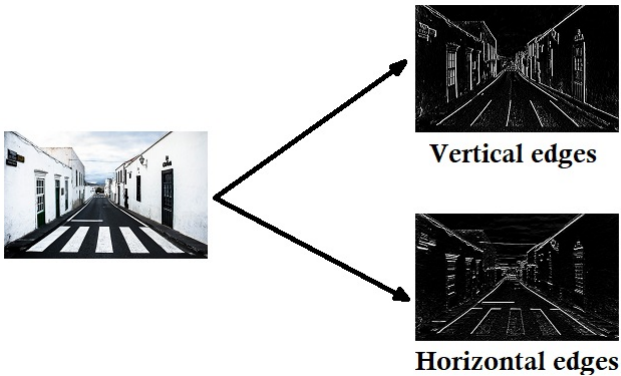
Datos cuya información no está organizada de acuerdo a una estructura particular.

[1] https://www.techtarget.com/searchbusinessanalytics/definition/unstructured-data

Datos estructurados vs datos no estructurados

- Los datos estructurados son idóneos para las técnicas clásicas de aprendizaje de máquina.
- En épocas previas al aprendizaje profundo, los ingenieros de visión por computador y procesamiento de lenguaje natural dedicaban jornadas para extraer características de los datos no estructurados.
- Los modelos de aprendizaje profundo tienen la capacidad de manipular datos no estructurados sin mayor procesamiento.

Ejemplo: Detección de bordes



- Kernel para bordes verticales

$$\begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}$$

- Kernel para bordes horizontales

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}$$

Diferentes kernel producen diferentes resultados. Solución: redes convolucionales.

Pontificia Universidad JAVERIANA Cali

Table of Contents

Tipos de atributos

- Se reconocen al menos dos tipos de atributos:
  - Los atributos numéricos.
  - Los atributos categóricos.
- Cada tipo de atributo tiene su propio procesamiento.
- Usualmente, los datos categóricos se representan a partir de cadenas de texto.

	ocean_proximity
17606	<1H OCEAN
18632	<1H OCEAN
14650	NEAR OCEAN
3230	INLAND
3555	<1H OCEAN
19480	INLAND
8879	<1H OCEAN
13685	INLAND
4937	<1H OCEAN
4861	<1H OCEAN