Procesamiento, análisis, modelado y síntesis de datasets

Definiciones básicas

Daniel Walther Berns

September 25, 2019

Dataset

- Un dataset es un conjunto de datos o valores medidos en experimentos científicos;

Dataset

- Un dataset es un conjunto de datos o valores medidos en experimentos científicos;
- Cada dataset tienen una cierta estructura propia, pero esta puede variar mucho entre diferentes datasets;

Dataset

- Un dataset es un conjunto de datos o valores medidos en experimentos científicos;
- Cada dataset tienen una cierta estructura propia, pero esta puede variar mucho entre diferentes datasets;
- Los datasets más simples pueden ser planillas en formato texto separado por comas;

Dataset

- Un dataset es un conjunto de datos o valores medidos en experimentos científicos;
- Cada dataset tienen una cierta estructura propia, pero esta puede variar mucho entre diferentes datasets;
- Los datasets más simples pueden ser planillas en formato texto separado por comas;
- Otros datasets más complejos pueden ser colecciones de imágenes o sonidos, almacenados en formato binario.

Ejemplo de dataset con forma de planilla

Tabla: Fisher's Iris Data

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.9	1.5	versicolor
5.8	2.7	5.1	1.9	virginica
6.3	2.9	5.6	1.8	virginica

Ejemplo de dataset con forma de planilla

Tabla: Fisher's Iris Data

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.9	1.5	versicolor
5.8	2.7	5.1	1.9	virginica
6.3	2.9	5.6	1.8	virginica

- Cada fila de la planilla corresponde a diferentes ejemplos de un objeto en estudio, un tipo de flor denominado Iris.

Ejemplo de dataset con forma de planilla

Tabla: Fisher's Iris Data

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.9	1.5	versicolor
5.8	2.7	5.1	1.9	virginica
6.3	2.9	5.6	1.8	virginica

- Cada fila de la planilla corresponde a diferentes ejemplos de un objeto en estudio, un tipo de flor denominado Iris.

Ejemplo de dataset con forma de planilla

Tabla: Fisher's Iris Data

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.9	1.5	versicolor
5.8	2.7	5.1	1.9	virginica
6.3	2.9	5.6	1.8	virginica

- Cada fila de la planilla corresponde a diferentes ejemplos de un objeto en estudio, un tipo de flor denominado Iris.
- Cada columna de la planilla corresponde a las distintas cualidades o atributos del objeto en estudio.

Aclaraciones

- La estructura de filas y columnas de la tabla de datos es perfecta para datos impresos en hojas de libros y revistas.

Aclaraciones

- La estructura de filas y columnas de la tabla de datos es perfecta para datos impresos en hojas de libros y revistas.
- Sin embargo, con la ayuda de la computadora y los equipos de mediciones automatizados es posible generar y emplear estructuras de datos más complejas y de mayor tamaño, como la hoja de cálculo (spreadsheet, en inglés).

Aclaraciones

- La estructura de filas y columnas de la tabla de datos es perfecta para datos impresos en hojas de libros y revistas.
- Sin embargo, con la ayuda de la computadora y los equipos de mediciones automatizados es posible generar y emplear estructuras de datos más complejas y de mayor tamaño, como la hoja de cálculo (spreadsheet, en inglés).
- la hoja de cálculo está formado por varias tablas de datos (columnas similares, tal vez diferente cantidad de filas).

Ejemplo de Secuencia (tabla con una columna)

Tabla: Secuencia: cantidad de clientes por hora en una caja de supermercado

Clientes por hora
10
30
29
31
...
11

Ejemplo de hoja de cálculo

Tabla: Comodoro Rivadavia

Hora	Temperatura	viento (km/h)	lluvia (ml/m^2)
10am	5.4	50	0
11am	8.1	52	0
12am	10.8	60	0
...			

Tabla: Rawson

Hora	Temperatura	viento (km/h)	lluvia (ml/m^2)
10am	7.9	20	0
11am	10.2	22	10
12am	13.2	10	0
...			

Notación de Secuencias, Tablas y Hojas de cálculo

- Secuencias, tablas y hojas de cálculo tienen una notación en común.

Notación de Secuencias, Tablas y Hojas de cálculo

- Secuencias, tablas y hojas de cálculo tienen una notación en común.
- $S(i)$ es una secuencia de nombre S e índice i.

Notación de Secuencias, Tablas y Hojas de cálculo

- Secuencias, tablas y hojas de cálculo tienen una notación en común.
- $S(i)$ es una secuencia de nombre S e índice i.
- $T(i, j)$ es una tabla de datos de nombre T e índices i, j.

Notación de Secuencias, Tablas y Hojas de cálculo

- Secuencias, tablas y hojas de cálculo tienen una notación en común.
- $S(i)$ es una secuencia de nombre S e índice i.
- $T(i, j)$ es una tabla de datos de nombre T e índices i, j.
- $C(i, j, k)$ es una hoja de cálculo de nombre C e índices i, j, k.

Notación de Secuencias, Tablas y Hojas de cálculo

- Secuencias, tablas y hojas de cálculo tienen una notación en común.
- $S(i)$ es una secuencia de nombre S e índice i.
- $T(i,j)$ es una tabla de datos de nombre T e índices i, j.
- $C(i,j,k)$ es una hoja de cálculo de nombre C e índices i, j, k.
- Los índices i, j, k son números enteros que van desde 1 hasta un límite definido por la memoria de la computadora.

Interpretación de una hoja de cálculo

- Empleamos una hoja de cálculo para manejar tabla de datos distribuidos en el tiempo o en el espacio.

Interpretación de una hoja de cálculo

- Empleamos una hoja de cálculo para manejar tabla de datos distribuidos en el tiempo o en el espacio.
- Una secuencia temporal de tablas de datos puede compararse con una película.

Interpretación de una hoja de cálculo

- Empleamos una hoja de cálculo para manejar tabla de datos distribuidos en el tiempo o en el espacio.
- Una secuencia temporal de tablas de datos puede compararse con una película.
- Podemos considerar cubos de datos con más de tres índices. Por ejemplo, $C(i, j, k, p, q)$. Estos conjuntos de datos o datasets son de uso común en registro de datos con distribuciones espacio temporales. Por ejemplo, propiedades del agua de mar (temperatura, salinidad, PH, turbidez, contenido de bacterias por mililitro) en varios puntos del globo en distintos meses del año y medidos por diferentes grupos científicos.

Filas y columnas de la tabla de datos

- Las filas de una tabla de datos están asociadas a objetos que nos interesa observar.

Filas y columnas de la tabla de datos

- Las filas de una tabla de datos están asociadas a objetos que nos interesa observar.
- Las filas se conocen también como registros, puntos, casos, muestras, entidades, o instancias.

Filas y columnas de la tabla de datos

- Las filas de una tabla de datos están asociadas a objetos que nos interesa observar.
- Las filas se conocen también como registros, puntos, casos, muestras, entidades, o instancias.
- Las columnas de una tabla de datos están asociadas a algunos atributos de los objetos en observación.

Filas y columnas de la tabla de datos

- Las filas de una tabla de datos están asociadas a objetos que nos interesa observar.
- Las filas se conocen también como registros, puntos, casos, muestras, entidades, o instancias.
- Las columnas de una tabla de datos están asociadas a algunos atributos de los objetos en observación.
- Las columnas se conocen también como variables, campos, o características.

Tipos de atributos

Existen diferentes tipos de atributos

- Nominales: colores, nombres de las estaciones del año, códigos postales.

Tipos de atributos

Existen diferentes tipos de atributos

- Nominales: colores, nombres de las estaciones del año, códigos postales.
- Ordinales: rankings, calificaciones,

Tipos de atributos

Existen diferentes tipos de atributos

- Nominales: colores, nombres de las estaciones del año, códigos postales.
- Ordinales: rankings, calificaciones,
- Intervalos (variación respecto a un punto de referencia): fechas, temperaturas en Celsius o Fahrenheit.

Tipos de atributos

Existen diferentes tipos de atributos

- Nominales: colores, nombres de las estaciones del año, códigos postales.
- Ordinales: rankings, calificaciones,
- Intervalos (variación respecto a un punto de referencia): fechas, temperaturas en Celsius o Fahrenheit.
- Proporciones (número por unidad): temperatura en Kelvin, longitudes, tiempo, tensiones, corrientes.

Ejemplo

Tabla: Características físicas de personas

Género	Mano hábil	Edad	Peso	Tipo de sangre	Educación
M	*D*	46	80	A rh+	egb
F	*D*	30	50	B rh-	polimodal
F	*I*	28	48	0 rh-	universitaria

Operaciones con valores de atributos

Se pueden aplicar diferentes operaciones a los valores de los atributos.

- Distinción: igualdad, desigualdad.

Operaciones con valores de atributos

Se pueden aplicar diferentes operaciones a los valores de los atributos.

- Distinción: igualdad, desigualdad.
- Ordenamiento: menor, mayor.

Operaciones con valores de atributos

Se pueden aplicar diferentes operaciones a los valores de los atributos.

- Distinción: igualdad, desigualdad.
- Ordenamiento: menor, mayor.
- Adición.

Operaciones con valores de atributos

Se pueden aplicar diferentes operaciones a los valores de los atributos.

- Distinción: igualdad, desigualdad.
- Ordenamiento: menor, mayor.
- Adición.
- Multiplicación.

Operaciones por tipo de atributos

- Atributos nominales: distinción.

Operaciones por tipo de atributos

- Atributos nominales: distinción.
- Atributos ordinales, distinción, ordenamiento.

Operaciones por tipo de atributos

- Atributos nominales: distinción.
- Atributos ordinales, distinción, ordenamiento.
- Atributos numéricos, que además se dividen en
  - Atributos de intervalos: distinción, ordenamiento, adición.
  - Atributos de proporciones: distinción, ordenamiento, adición, multiplicación.

Atributos nominales

- Un atributo nominal toma valores pertenecientes a un conjunto discreto, numerable y limitado de nombres o identificadores, con la información necesaria para distinguir un objeto de otro.

Atributos nominales

- Un atributo nominal toma valores pertenecientes a un conjunto discreto, numerable y limitado de nombres o identificadores, con la información necesaria para distinguir un objeto de otro.
- Por ejemplo, códigos postales, números de identificación, color de ojos, género $\{femenino, masculino\}$.

Atributos nominales

- Un atributo nominal toma valores pertenecientes a un conjunto discreto, numerable y limitado de nombres o identificadores, con la información necesaria para distinguir un objeto de otro.
- Por ejemplo, códigos postales, números de identificación, color de ojos, género $\{femenino, masculino\}$.
- Operaciones: modo, cálculo de entropía, correlación de contingencia, test chi cuadrado.

Atributos ordinales

- Un atributo ordinal toma valores para los que existe una relación de orden.

Atributos ordinales

- Un atributo ordinal toma valores para los que existe una relación de orden.
- Un atributo ordinal puede tomar valores numéricos pero la diferencia de valores no tiene importancia mas allá de su ordenamiento.

Atributos ordinales

- Un atributo ordinal toma valores para los que existe una relación de orden.
- Un atributo ordinal puede tomar valores numéricos pero la diferencia de valores no tiene importancia mas allá de su ordenamiento.
- Ejemplos:
  - bajo, mediano, alto;
  - malo, regular, bueno, muy bueno, excelente;
  - calificaciones;
  - números de direcciones.

Atributos ordinales

- Un atributo ordinal toma valores para los que existe una relación de orden.
- Un atributo ordinal puede tomar valores numéricos pero la diferencia de valores no tiene importancia mas allá de su ordenamiento.
- Ejemplos:
  - bajo, mediano, alto;
  - malo, regular, bueno, muy bueno, excelente;
  - calificaciones;
  - números de direcciones.
- Operaciones: mediana, histogramas.

Atributos numéricos

- Un atributo numérico es una cantidad mensurable, representada como número entero (punto fijo) o real con cantidad limitada de decimales (punto flotante).

Atributos numéricos

- Un atributo numérico es una cantidad mensurable, representada como número entero (punto fijo) o real con cantidad limitada de decimales (punto flotante).
- Un atributo numérico puede ser de dos tipos: intervalo o proporción.

Atributos numéricos

- Un atributo numérico es una cantidad mensurable, representada como número entero (punto fijo) o real con cantidad limitada de decimales (punto flotante).
- Un atributo numérico puede ser de dos tipos: intervalo o proporción.
- Un atributo numérico del tipo intervalo toma valores cuyas diferencias son interpretables, pueden ser sumados o restados pero no tienen un punto de referencia.

Atributos numéricos

- Un atributo numérico es una cantidad mensurable, representada como número entero (punto fijo) o real con cantidad limitada de decimales (punto flotante).
- Un atributo numérico puede ser de dos tipos: intervalo o proporción.
- Un atributo numérico del tipo intervalo toma valores cuyas diferencias son interpretables, pueden ser sumados o restados pero no tienen un punto de referencia.
- Un atributo numérico del tipo proporción es un valor numérico con un punto cero o referencia.

Vamos a los problemas de la vida real

- Los datos no tienen forma de planilla (por ejemplo, son imágenes, videos o sonidos almacenados en formato binario).

Vamos a los problemas de la vida real

- Los datos no tienen forma de planilla (por ejemplo, son imágenes, videos o sonidos almacenados en formato binario).
- Los equipos de laboratorio registran y entregan datos en formatos propietarios (sin acceso público, abierto y gratuito)

Vamos a los problemas de la vida real

- Los datos no tienen forma de planilla (por ejemplo, son imágenes, videos o sonidos almacenados en formato binario).
- Los equipos de laboratorio registran y entregan datos en formatos propietarios (sin acceso público, abierto y gratuito)
- Falta información sobre formatos de datos.

Vamos a los problemas de la vida real

- Los datos no tienen forma de planilla (por ejemplo, son imágenes, videos o sonidos almacenados en formato binario).
- Los equipos de laboratorio registran y entregan datos en formatos propietarios (sin acceso público, abierto y gratuito)
- Falta información sobre formatos de datos.
- Las computadoras conectadas a los equipos de medición se vuelven obsoletas.

Vamos a los problemas de la vida real

- Los datos no tienen forma de planilla (por ejemplo, son imágenes, videos o sonidos almacenados en formato binario).
- Los equipos de laboratorio registran y entregan datos en formatos propietarios (sin acceso público, abierto y gratuito)
- Falta información sobre formatos de datos.
- Las computadoras conectadas a los equipos de medición se vuelven obsoletas.
- Los datasets son masivos: la cantidad de datos es mayor a la que una persona puede procesar o analizar.

Formatos propietarios

- Si se dice que un formato de archivo es propietario, significa que existe una limitación legal que impide desarrollar un programa para operar con los archivos con dicho formato.

Formatos propietarios

- Si se dice que un formato de archivo es propietario, significa que existe una limitación legal que impide desarrollar un programa para operar con los archivos con dicho formato.
- Adicionalmente, el formato propietario puede ser secreto.

Formatos propietarios

- Si se dice que un formato de archivo es propietario, significa que existe una limitación legal que impide desarrollar un programa para operar con los archivos con dicho formato.
- Adicionalmente, el formato propietario puede ser secreto.
- El objetivo de los formatos propietarios es la protección de intereses comerciales.

¿Qué podemos hacer con un dataset?

- Exploración de datasets

¿Qué podemos hacer con un dataset?

- Exploración de datasets
- Modelado

¿Qué podemos hacer con un dataset?

- Exploración de datasets
- Modelado
- Análisis

¿Qué podemos hacer con un dataset?

- Exploración de datasets
- Modelado
- Análisis
- Procesamiento

¿Qué podemos hacer con un dataset?

- Exploración de datasets
- Modelado
- Análisis
- Procesamiento
- Síntesis

Software para operaciones con datasets

- Es posible escribir pequeños programas en (Matlab/Octave/Python)

Software para operaciones con datasets

- Es posible escribir pequeños programas en (Matlab/Octave/Python)
- con el objetivo de automatizar tareas de cálculo, graficación, procesamiento, análisis, modelado y generación de reportes.

Software para operaciones con datasets

- Es posible escribir pequeños programas en (Matlab/Octave/Python)
- con el objetivo de automatizar tareas de cálculo, graficación, procesamiento, análisis, modelado y generación de reportes.
- La clave es automatizar trabajo, o sea delegar trabajo en la computadora, para dedicar nuestro tiempo a la interpretación de resultados.

Ejemplo

- Cuando obtenemos un resultado, es necesario comprobar su corrección.

Ejemplo

- Cuando obtenemos un resultado, es necesario comprobar su corrección.
- Por ejemplo, si tenemos un sistema de ecuaciones $A \cdot x = b$, entonces el vector $c = A^{-1} \cdot b$ debería cumplir con la verificación $A \cdot c - b = 0$

Ejemplo

- Cuando obtenemos un resultado, es necesario comprobar su corrección.
- Por ejemplo, si tenemos un sistema de ecuaciones $A \cdot x = b$, entonces el vector $c = A^{-1} \cdot b$ debería cumplir con la verificación $A \cdot c - b = 0$
- Es conveniente que para cada cálculo definamos una verificación, para detectar los casos límite donde alguna hipótesis no se cumpla.

Ejemplo

- Cuando obtenemos un resultado, es necesario comprobar su corrección.
- Por ejemplo, si tenemos un sistema de ecuaciones $A \cdot x = b$, entonces el vector $c = A^{-1} \cdot b$ debería cumplir con la verificación $A \cdot c - b = 0$
- Es conveniente que para cada cálculo definamos una verificación, para detectar los casos límite donde alguna hipótesis no se cumpla.
- Por ejemplo, para poder invertir una matriz y resolver un sistema de ecuaciones lineales, la matriz debe ser bien condicionada.

Verificaciones automatizadas

- Lo ideal es automatizar las verificaciones.

Verificaciones automatizadas

- Lo ideal es automatizar las verificaciones.
- Si la verificación se pasa satisfactoriamente, nuestro programa debe escribir 'todo bien'.

Verificaciones automatizadas

- Lo ideal es automatizar las verificaciones.
- Si la verificación se pasa satisfactoriamente, nuestro programa debe escribir 'todo bien'.
- Si la verificación falla, nuestro programa debe escribir debe escribir 'hay un error en X, linea N', donde 'X' es el programa que falló, y 'N' es el número de línea.

Cualidades del Software para operar con Datasets

- Cuando nos planteamos la posibilidad de desarrollar nuestro propio software para operar con datasets, nuestra intención es poder comunicar los resultados a otras personas.

Cualidades del Software para operar con Datasets

- Cuando nos planteamos la posibilidad de desarrollar nuestro propio software para operar con datasets, nuestra intención es poder comunicar los resultados a otras personas.
- Además, queremos obtener resultados reproducibles no solamente por nosotros, sino también por cualquier persona que desee hacerlo.

Cualidades del Software para operar con Datasets

- Cuando nos planteamos la posibilidad de desarrollar nuestro propio software para operar con datasets, nuestra intención es poder comunicar los resultados a otras personas.
- Además, queremos obtener resultados reproducibles no solamente por nosotros, sino también por cualquier persona que desee hacerlo.
- Para lograr estas cualidades, necesitamos organizar, documentar y publicar tanto el software como los datasets.

Cualidades del Software para operar con Datasets

- Cuando nos planteamos la posibilidad de desarrollar nuestro propio software para operar con datasets, nuestra intención es poder comunicar los resultados a otras personas.
- Además, queremos obtener resultados reproducibles no solamente por nosotros, sino también por cualquier persona que desee hacerlo.
- Para lograr estas cualidades, necesitamos organizar, documentar y publicar tanto el software como los datasets.
- El objetivo de este curso es mostrar como desarrollar nuestro propio software para operar con datasets, con la organización y la documentación adecuadas para su publicación y operación por otras personas.

Que vamos a ver

- Github: para guardar programas.

Que vamos a ver

- Github: para guardar programas.
- Google Colab: para aprender python en la nube.

Que vamos a ver

- Github: para guardar programas.
- Google Colab: para aprender python en la nube.
- Anaconda: para instalar python en nuestras computadoras (para Windows, Mac y Linux)

Que vamos a ver

- Github: para guardar programas.
- Google Colab: para aprender python en la nube.
- Anaconda: para instalar python en nuestras computadoras (para Windows, Mac y Linux)
- Numpy, Pandas y Matplotlib: para leer archivos (csv, json y excel), realizar cálculos y generar gráficos.

Que vamos a ver

- Github: para guardar programas.
- Google Colab: para aprender python en la nube.
- Anaconda: para instalar python en nuestras computadoras (para Windows, Mac y Linux)
- Numpy, Pandas y Matplotlib: para leer archivos (csv, json y excel), realizar cálculos y generar gráficos.
- Generación de reportes escritos en Latex y Markdown.

Tareas

- Aseguresé de tener una cuenta de gmail disponible, para disponer de acceso a Google Drive y Google Colab.

Tareas

- Aseguresé de tener una cuenta de gmail disponible, para disponer de acceso a Google Drive y Google Colab.
- Busque los archivos con datos de experimentos de los que disponga. Escriba un reporte breve explicando como los organiza, si hace copia de protección, y como hace para generar gráficos y resultados para publicaciones. También describa que resultados desearía obtener y cuáles son sus datos disponibles.