

Documentación



Javier Andrés Causil Martínez

Universidad de Antioquia - Científico de datos

 Portafolio

 Causil

Índice general

I	Ciencia de datos	1
0.1.	Proyectos data Analysis	2
1.	Análisis de datos	3
1.1.	¿Qué tipo de información podemos analizar?	3
1.2.	Flujo de trabajo en ciencia de datos: fases, roles y oportunidades laborales . .	4
1.3.	Herramientas para cada etapa del análisis de datos	4
1.4.	Python en ciencia de Datos	4
2.	Proyectos	5
2.1.	Credit Card Fraud Detection	5
3.	Course Structure & Outline	6
3.1.	Flavors of Analytics	6
II	Metodologías	7
4.	CRISP-DM	8
III	Mathematics	9
5.	Probabilidad	11
5.1.	Técnicas de conteo	11
5.1.1.	Principio de la multiplicación	11

5.1.2. Principio de la adición	11
5.1.3. Permutaciones	12
5.1.4. Talleres	12
5.2. Permutaciones, combinaciones y variaciones	13
5.3. Conceptos básicos de probabilidad	13
5.3.1. Conceptos básicos	13
Bibliografía	14

Índice de figuras

©EUREKA INFINITY 2022

Eureka infinity es mi proyecto personal, su finalidad es publicar todo lo relacionado con la matemática que yo vaya produciendo relacionado con mi entorno, ya se han cursos, avances de proyectos personales, hasta hoy día tengo poca experiencia profesional, pero con el tiempo, la disciplina y la constancia en el estudio, estaré creciendo gracias a la comunidad.

Parte I

Ciencia de datos

¿Qué es data ciencia? Es el proceso de descubrir información valiosa de los datos.

¿cuál es su finalidad?

1. Tomar decisiones y crear estrategias de negocio.
2. Crear productos de software más inteligentes y funcionales.

¿De que trata este proceso?:

1. Obtención de los datos: a través de encuestas
2. Transformar y limpiar los datos.
3. Explorar, analizar y visualizar datos.
4. Usar modelos de machine learning*.
5. Integrar datos e IA a productos de software.

¿Qué es Data Science?

Data science o ciencia de datos es el proceso de descubrir información valiosa de los datos.

¿Cuál es su finalidad? Tomar decisiones y crear estrategias de negocio Crear productos de software más inteligentes y funcionales.

¿De qué trata este proceso?

Obtención de los datos Mediciones Encuestas Internet

Transformar y limpiar los datos Incompletos Formato Incorrecto

Explorar, analizar y visualizar datos Patrones o tendencias Insights Visualizaciones, gráficos o reportes

Usar modelos de machine learning

Machine learning o aprendizaje automático es una rama de inteligencia artificial. Su objetivo es que las computadoras aprendan. En machine learning, las computadoras observan grandes cantidades de datos y construyen un modelo capaz de generar predicciones para resolver problemas.

Integrar datos e IA a productos de software Ponerlos a disposición del usuario final.

La ciencia de datos es una intercepción de conocimiento entre (matemáticas y estadística), (ciencias computacionales) y conocimiento del dominio.

0.1. Proyectos data Analysis

Poner en practica lo mas rápido que se pueda, tener proyectos personales, en que gasto los dineros del mes, que productos consigo cada mes, encontrar anomalías, proyectos con kaggle.

Capítulo 1

Análisis de datos

¿Qué es ciencia de datos y big data? ¿Cómo afectan a mi negocio?

“¿Qué haces en tu trabajo (como científico de datos)? Mi trabajo es crear una solución matemática o estadística para un problema del negocio”

1.1. ¿Qué tipo de información podemos analizar?

Descubrir qué tipos de información existen, qué industrias los usan y qué tipo de acciones podemos tomar a partir de ellos.

Los principales datos que existen son:

Personas: Este tipo de datos lo extraemos de las personas, es decir lo generamos nosotros cuando le damos like a una foto de facebook, de preferencia, tipo de videos, de quien te gusta mas el contenido, subiendo una foto y etiquetando a un compañero.

Transacciones: las monetarias y las no monetarias, cualquier transacción que hago con una tarjeta de crédito o débito, cuando hacemos un pago electrónico o digital queda una huella, queda un registro de quien lo hizo, por que monto lo hizo y en que establecimiento lo hizo, es muy interesante por que las bancas digitales pueden hacerte recomendaciones sobre el tipo de comercio que te podía interesar.

No financieras: las compañías telefónicas identifican cual es tu patrón habitual, cuantas llamadas haces, a que personas llamas, cuanto duran tus llamadas, y a partir de esto te llaman para que no abandones el servicio.

Navegación web: Estas son las famosas cookies, ellas están advirtiéndote de la información que van a recoger.

Machine 2 machine: Una conexión de una máquina a otra máquina, la usan las plataformas de transporte, google maps y para hacer la locación entre dispositivos.

Biométricos: Cada vez son mas habituales y únicas, huellas digitales, reconocimiento facial.

1.2. Flujo de trabajo en ciencia de datos: fases, roles y oportunidades laborales

Roles en datos:

Ingeniero de datos: crear bases de datos Hacer que la empresa, hace la conexión de los dispositivos y las bases de datos,

Analista business intelligence: A partir de la información que ha creado el ingeniero de datos va extraer la data, crear cuadros de control, crear dashboard, monitoreo, va automatizar estos procedimientos para que cualquier persona de la empresa pueda interpretarla, las herramientas mas utilizadas son SQL y Excel. No necesariamente sabe Python.

Data Scientist: Sabe hacer el rol del analista, sabe extraer la información y sabe predecir, con las herramientas de estadística, nos guía a donde vamos.

Data Translator: Nos ayuda a proyectar el equipo, nos ayuda a comunicar con los otros equipos del negocio.

1.3. Herramientas para cada etapa del análisis de datos

El primero es el rol del analista y del ingeniero estas son las personas que crean bases de datos y utilizan SQL, se sintetiza la información de la base de datos.

El científico de datos son herramientas predictivas, son R y Python, R es mas estadístico análisis descriptivo,

1.4. Python en ciencia de Datos

Por que numpy para el análisis de datos. Tenemos tres cosas a destacar

1. Un poderoso objeto array multidimensional.
2. Funciones matem

Crear un virtual environments ejecutamos la siguiente linea de comando

```
python3 -m venv my_env  
source bin/activate
```

Capítulo 2

Proyectos

2.1. Credit Card Fraud Detection

Anonymized credit card transactions labeled as fraudulent or genuine

About Dataset

Context

The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172 % of all transactions.

Content

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are "Time" and "Amount". Feature "Time" contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature "Amount" is the transaction Amount, this feature can be used for example-dependant cost-sensitive learning. Feature "Class" is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Given the class imbalance ratio, we recommend measuring the accuracy using the Area Under the Precision-Recall Curve (AUPRC). Confusion matrix accuracy is not meaningful for unbalanced classification.

Capítulo 3

Course Structure & Outline

Hey everyone chris dutton here and welcome to thinking like an analyst this is a crash course designed for

3.1. Flavors of Analytics

All right let's take a minute and talk about the various roles or flavors of analytics because the thing is this field is very broad and it's very diverse so it can be helpful to categorize various roles or job titles based on different types of skills so you may have seen venn diagrams out there that look something like this this version is adapted from learn.co we've got business intelligence skills they're in the blue bubble programming coding skills gray and math and stats skills in yellow now by visualizing skills in this way you can map various roles to the diagram based on the overlaps so for instance someone who skews towards programming and math might fall into the machine learning bucket people with and bi skills might fall into data engineering bi and math and stats maybe we call those people advanced analysts and those who use all three types of skills relatively equally might fall into the data science category and while this can be helpful to an extent remember that in reality this is fluid it's flexible and it's often somewhat subjective as well you know you can be a bi analyst who loves stats or programming or a machine learning engineer with exceptional business intelligence skills or literally any other combination of these categories the key is that this diagram this entire diagram represents the broader world of data analytics and well many types of roles fall under this analytics umbrella they're all aligned towards the same ultimate goal using data to make smart decisions. Now think about it that applies whether you're a bi analyst a statistician a data scientist or an everyday excel jockey and the difference is they come down to things one the types of problems that you're trying to solve and two the types of tools that you're using to solve them so in the next lesson we'll dive into one of the most common comparisons out there business intelligence versus data science but for now what's important to specific role you play we're all playing for the same team

Parte II

Metodologías

Capítulo 4

CRISP-DM

CRISP-DM, que son las siglas de Cross-Industry Standard Process for Data Mining, es un método probado para orientar sus trabajos de minería de datos.

Parte III

Mathematics

Demostración: (del latín *demostratio-onem*) Prueba de una verdad por medio del raciocinio, partiendo de principios evidentes. En ciencia, en general se puede extender a la comprobación experimental de un principio o teoría.

Proposición (del latín *propositio-onem*): Enunciación de una verdad ya demostrada o que se ha de demostrar.

Silogismo (del latín *sylogismus*, y éste del griego): Argumento formado por tres proposiciones: premisa mayor, premisa menor y conclusión.

Enunciado (del latín *enuntiatio-onem*) Expresión o manifestación de una idea.

Hipótesis (del latín *hypothesis*, y éste del griego): Suposición de una idea, con el objetivo de deducir de ella alguna consecuencia.

Lema (del latín *lemma*, y éste del griego): Proposición cuya demostración antecede a un teorema.

Axioma (del latín *axioma*, y éste del griego): Principio, verdad, sentencia clara y evidente, que no necesita demostración.

Conclusión (del latín *conclusio-onem*): Proposición que se deduce o es consecuencia de las premisas.

Conjetura (del latín *conjectura*): Opinión fundada en probabilidades, indicios o apariencias.

Corolario (del latín *corollarium*) Proposición que por sí sola se deduce de lo ya demostrado.

Premisa (del latín *praemissa*): Cualquiera de las dos proposiciones lógicas de un silogismo, de donde se deduce la conclusión. En un silogismo, a la proposición más general se denomina mayor y a la otra menor.

Principio (del latín *principium*): Argumento considerado como origen, fundamentación o razón primera de un razonamiento.

Tesis (del latín *thesis*, y éste del griego): Conclusión o proposición que se mantiene con razonamientos.

Postulado (del latín *postulatus*): Proposición cuya verdad se admite sin pruebas y que es necesaria para servir de base en ulteriores razonamientos.

Propiedad (del latín *propietas*): Atributo o cualidad de una persona o cosa.

Teorema (del latín *theoremata*, y éste del griego): Proposición que afirma una verdad susceptible de demostración.

Capítulo 5

Probabilidad

5.1. Técnicas de conteo

5.1.1. Principio de la multiplicación

Definition 5.1.1. (Principio de la multiplicación) Si un experimento consiste de n etapas o sucesos y si la primera etapa o suceso puede ser realizada de n_1 maneras, la segunda etapa de n_2 maneras, la k -ésima etapa de n_k maneras, entonces el experimento completo puede ser realizado de $n_1 \cdot n_2 \cdot \dots \cdot n_k$ maneras.

5.1.2. Principio de la adición

Si un suceso S_1 ocurre de a maneras y un suceso S_2 ocurre de b maneras entonces puede decirse que el número total de maneras en que puede ocurrir S_1 o el suceso S_2 es $a + b$.

Example 5.1.2.

- 1 **Ejemplo 1:** ¿Cuántos resultados diferentes se pueden obtener lanzando un dado o una moneda? Indetificamos que el experimento consiste en elegir que objeto lanzar, suponga:

$$S_1: = \text{ lanzar el dado } \text{ y } S_2: = \text{ lanzar la moneda.}$$

Si se elige el dado, entonces el número de resultados posibles es $\{1, 2, 3, 4, 5, 6\}$ o si elige la moneda el número de resultados posibles es $\{C, S\}$. Así el número total de resultados posibles para el suceso S_1 o S_2 es:

$$6 + 2 = 8$$

- 2 **Ejemplo 2:** ¿Cuántos resultados diferentes se pueden obtener lanzando un dado y una moneda? Considerando los mismo sucesos S_1 y S_2 del ejemplo anterior, el experimento consiste en lanzar un dado y seguidamente lanzar una moneda. nuestro conjunto de resultados sería

$$\{(1, C), (1, S), (2, C), (2, S), \dots, (6, C), (6, S)\}.$$

Así la cantidad de sucesos es $6 \cdot 2 = 12$.

5.1.3. Permutaciones

¿De cuántas formas se pueden ordenar u organizar los elementos de un conjunto o de un subconjunto? Para responder tal pregunta, realicemos las siguientes definiciones.

Definition 5.1.3.

1. Ordenar: significa poner una cantidad de elementos en un orden específico que obedece a una regla lógica.
2. Organizar: implica poner una cantidad de elementos de forma que no tienen un orden específico o dicho orden no es importante dentro de un determinado contexto.
3. Conjunto:
4. Subconjunto:
5. Orden: corresponde a la aplicación de un criterio específico que le confiere cierta jerarquía a los elementos que forman parte de un conjunto.

Definimos los conceptos de permutación y combinación:

Definition 5.1.4. Permutación

5.1.4. Talleres

1. Se selecciona al azar un vehículo en cierta ciudad. Si todas las letras de la placa del vehículo son diferentes, ¿Cuántos autos tienen la misma característica? ¿Cuántos autos tienen placas con todos sus dígitos impares?

Solución: Suceso:

S_1 : = formar una placa de auto eligiendo una letra del abcdario sin repetir y elegir un número del 1 al 10.

Sabemos que las letras del abcdario son 27, aplicando el principio de la multiplicación, el número de placas posibles es:

$$27 \cdot 26 \cdot 25 \cdot 10 \cdot 10 \cdot 10 = 17550000$$

$${}_{27}P_3 \cdot {}_{10}C_3 =$$

2. **Ejemplo 2:** Se lanzan 20 monedas no cargadas, ¿Cuántos posibles resultados tienen solo tres caras?

Solución:

$$\binom{20}{3} = \frac{20!}{(20-3)!3!} = \frac{20 \cdot 19 \cdot 18 \cdot 17!}{17!3!} = 20 \cdot 19 \cdot 3 = 1140$$

3. Se seleccionan al azar tres personas de un grupo por 10 obreros, 4 pintores y 6 carpinteros. ¿Cuántos grupos diferentes conformados por un obrero, un pintor y un carpintero se pueden formar?

Solución:

$$10 \cdot 4 \cdot 6 = 240$$

4. El pedido de una computadora personal digital puede especificar uno de cinco tamaños de memoria, cualquier tipo de monitos de tres posibles, cualquier tamaño de disco duro de entre cuatro posibles, y puede incluir o no una tableta para lápiz electrónico. ¿Cuántos sistemas distintos pueden ordenarse?

Solución:

$$5 \cdot 3 \cdot 4 \cdot 2 = 120.$$

El total de sistemas que pueden formarse son 120.

5. Un proceso de manufactura está formado por 10 operaciones, las cuales pueden efectuarse en cualquier orden. ¿Cuántas secuencias de producción distintas son posibles?

Solución:

$$10! = 3628800$$

6. Un proceso de manufactura está formado por 10 operaciones. Sin embargo, cinco de ellas deben terminarse antes de que pueda darse inicio a las otras cinco. Dentro de cada conjunto de cinco, las operaciones pueden efectuarse en cualquier orden. ¿Cuál es el número de secuencias de operaciones distintas posible?

Solución:

$$P_5^{10} + =$$

5.2. Permutaciones, combinaciones y variaciones

5.3. Conceptos básicos de probabilidad

5.3.1. Conceptos básicos

Espacio de probabilidad

En esta sección desarrollamos la noción de medida de probabilidad y se presentan sus propiedades básicas.

Definition 5.3.1. (Experimento aleatorio) Un experimento es aleatorio si su resultado no puede ser determinado de antemano.

Si el conjunto de los posibles resultados de un experimento aleatorio es conocido, a este conjunto se le llama espacio aleatorio

Definition 5.3.2. (Espacio muestral) El conjunto Ω de todos los posibles resultados de un experimento aleatorio es llamado espacio muestral. Un elemento $w \in \Omega$ es llamado un resultado o punto muestral.

Example 5.3.3. (Ejemplos de espacios muestrales)

1. **Experimento:** Lanzar una moneda. Los posibles resultados en este caso son “cara” y “sello”. Esto es

$$\Omega = \{C, S\}$$

2. Experimento: Lanzar un dado ordinario tres veces consecutivas. En este caso los posibles resultados son tripletas de la forma (i, j, k) donde $i, j, k \in \{1, 2, 3, 4, 5, 6\}$. Esto es

$$\Omega = \{(i, j, k) : i, j, k \in \{1, 2, 3, 4, 5, 6\}\}$$

3. Experimento:

Definition 5.3.4. (Espacio muestral discreto) Un espacio muestral Ω , es llamado discreto si es finito o contable.

Definition 5.3.5. (σ - Álgebra) Sea $\Omega \neq \emptyset$. Una colección \mathcal{F} de subconjuntos de Ω es llamada σ -álgebra si satisface las siguientes propiedades:

1. $\Omega \in \mathcal{F}$.
2. Si $A \in \mathcal{F}$, entonces $A^c \in \mathcal{F}$.
3. Si $A_1, A_2, \dots \in \mathcal{F}$, entonces $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Bibliografía

- [1] S.C. Brenner, L.R. Scott *The Mathematical Theory of Finite Element Methods*, 3^a ed., USA: Springer, 2008.
- [2] “Uso de Jupyter Notebook en un entorno virtual”, 2022. [En línea]. Disponible en: <https://es.acervolima.com/uso-de-jupyter-notebook-en-un-entorno-virtual/>. [Accedido: 21-jun-2022]
- [3] “Create a Next.js App”, 2022.[En línea]. Disponible en: <https://nextjs.org/learn/basics/create-nextjs-app>. [Accedido: 15-jun-2022]
- [4] M.P. Do Carmo *Differential Geometry Of Curves & Surfaces*, 1^a ed., USA: Prentice-Hall, 1976.
- [5] L. Leithold, , 2007. Editorial Harla. México. *Cálculo con Geometría Analítica.*, 1^a ed., USA: Prentice-Hall, 1976.