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. Combinatorics **11**

5.1. Four Basic Counting Principles 11

5.1.1. Principio de la adición 12

5.1.2. Principio de la multiplicación 12

5.1.3. Permutaciones 12

6. Probabilidad 14

6.1. Permutaciones, combinaciones y variaciones 14

6.2. Conceptos básicos de probabilidad 14

6.2.1. Conceptos básicos 14

6.2.2. Espacio de probabilidad de Laplace 15

6.2.3. Probabilidad condicional y eventos independientes 15

6.2.4. Probabilidad geométrica 16

6.3. Variables aleatorías y sus distribuciones 16

6.3.1. Definiciones y propiedades 16

Bibliografía 16

Índice de figuras

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 travéz 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 transación que hago con una tarjeta de crédito o débito, cuando hacemos un pago electrónico o dígital 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 finacieras: las compañias 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 advirtiendo de la información que van a recoger.

Machine 2 machine: Una conexion de una maquina a otra maquina, 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 conexion 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 anÃ¡lisis 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 crah course desaigned 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 ir can be helpful to categorize various roles or job titles based on difertent 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 whi skews towards programmind 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 syllogismus, 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 mÃ¡s 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 theorema, y éste del griego): Proposición que afirma una verdad susceptible de demostración.

Capítulo 5

Combinatorics

5.1. Four Basic Counting Principles

Definition 5.1.1. (Pairwise disjoint sets.) Let S be a set. A partition of S is a collection $S_1, S_2, \ldots, S_m$ of subsets of S such that each element of S is in exactly one of those subsets:

$$S_1 \cup S_2 \cup \cdots \cup S_m = S \quad \text{and} \quad S_i \cap S_j = \emptyset \quad \text{for all } i \neq j.$$

Thus, the sets $S_1, S_2, \ldots, S_m$ are pairwise disjoint sets, and their union is S. The subsets $S_1, S_2, \ldots, S_m$ are called the parts of the partition.

The numbers of objects of a set S is denoted by $|S|$ and is sometimes called the size of S.

Example 5.1.2.

1. $\Omega = \{T, H\}$ where T denotes tails and H denotes heads. The set Ω is a partition of the set of all possible outcomes of a single coin toss.

$$\Omega = \{T\} \cup \{H\}$$

Definition 5.1.3. (Addition Principle.) Suppose that a set S is partitioned into pairwise disjoint parts $S_1, S_2, \ldots, S_n$. The number the of objects in S can be determined by finding the number of objects un each of the parts, and adding the number so obtained:

$$|S| = |S_1| + |S_2| + \ldots + |S_n|. \tag{5.1}$$

If the sets $S_1, S_2, \ldots, S_m$ are allowed to overlap, then a more profound principle, the inclusion-exclusion principle, is needed.

Example 5.1.4.

1.

Definition 5.1.5. (Multiplication Principle.) Let S be a set of ordered pairs (a, b) of objects, where the first object a comer from a set of size p , and for each choice of object a there are q choices for object b. Then the size of S is $p \times q$:

$$|S| = p \times q. \tag{5.2}$$

The mulplication principle is actually a consequence of the addition principle.

A second useful formulation of the multiplication principle is as follows: If a first task has p outcomes and, no matter what the outcome of the first task, a second task has q outcomes, then the two tasks performed consecutively have $p \times q$ outcomes.

5.1.1. 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 5.1.6.

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 elije 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 experimiento consiste en lanzar un dado y seguidamente lanzar una moneda. nuestro conjunto de resultados sería

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

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

5.1.2. Principio de la multiplicación

Definition 5.1.7. (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 $n_1 \cdot n_2 \cdot \ldots \cdot n_k$ maneras.

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 5.1.8.

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 lso conceptos de permutación y combinación:

Definition 5.1.9. Permuatción

Capítulo 6

Probabilidad

The enunciated mathematics in this section are reference of the books [1], [2].

6.1. Permutaciones, combinaciones y variaciones

6.2. Conceptos básicos de probabilidad

6.2.1. Conceptos básicos

Espacio de probabilidad

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

Definition 6.2.1. (Experimiento 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 llamada espacio aleatorio

Definition 6.2.2. (Espacio muestral) El conjunto Ω de todos los posibles resultados de un experimiento aleatorio es llamado espacio muestral. Un elemento $w \in \Omega$ es llamado un resultado o punto muestral.

Example 6.2.3. (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. **Experimeinto:** Lanzar un dado ordianrio 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 6.2.4. (Espacio muestral discreto) Un espacio muestral Ω, es llamado discreto si es finito o contable.

Definition 6.2.5. (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 $A_1, A_2, \ldots \in \mathcal{F}$, entonces $\bigcup i = 1^\infty A_i \in \mathcal{F}$.

6.2.2. Espacio de probabilidad de Laplace

Definition 6.2.6. (Espacio de probabilidad de Laplace) Un espacio de probabilidad $(\Omega, \mathfrak{J}, P)$ con Ω finito, $\mathfrak{J} = \mathcal{P}(\Omega)$ y $P(w) = 1/|\Omega|$ para todo $w \in \Omega$ es llamado un espacio de probabilidad de Laplace. La medida de probabilidad P es llamada uniforme o la distribución clásica sobre Ω.

6.2.3. Probabilidad condicional y eventos independientes

Definition 6.2.7. (Probabilidad condicional) Sea $(\Omega, \mathfrak{J}, P)$ un espacio de probabilidad. Si $A, B \in \mathfrak{J}$ con $P(A) > 0$, entonces la probabilidad de el evento B bajo la condición A es definida de la siguiente manera:

$$P(B|A) \colon = \frac{P(A \cap B)}{P(A)} \tag{6.1}$$

Theorem 6.2.8. *(Medida de probabilidad condicional) Sea $(\Omega, \mathfrak{J}, P)$ un espacio de probabilidad, y sea $A \in \mathfrak{J}$ con $P(A) > 0$. Entonces:*

1. *$P(\cdot\,|\,A)$ es una medida de probabilidad sobre Ω centrada en A, esto es, $P(A|A) = 1$.*

2. *Si $A \cap B = \emptyset$, entonces $P(B|A) = 0$.*

3. *$P(B \cap C|A) = P(B|A \cap C)P(C|A)$ si $P(A \cap C) > 0$.*

4. *Si $A_1, A_2, \ldots, A_n \in \mathfrak{J}$ con $P(A_1 \cap A_2 \cap \cdots \cap A_{n-1}) > 0$, entonces*

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}).$$

Theorem 6.2.9. (Teorema de la probabilidad total) *Sea $A_1, A_2, A_3, A_4, \ldots,$ particiones finitas o countables de Ω, esto es, $A_i \cap A_j = \emptyset$ para todo $i \neq j$ y $\bigcup_{i=1}^{\infty} A_i = \Omega$; tal que $P(A_i) > 0$ para todo $A_i \in \mathfrak{J}$. Entonces, para cada $B \in \mathfrak{J}$:*

$$P(B) = \sum_i P(B|A_i)P(A_i) \tag{6.2}$$

Corollary 6.2.10. (Regla de Bayes) *Sean $A_1, A_2, \ldots$ un partición finita o contable de Ω con $P(A_i) > 0$ para todo i; entonces, para cada $B \in \mathfrak{J}$ con $P(B) > 0$:*

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_j P(B|A_j)P(A_j)} \quad \textit{para todo } i. \tag{6.3}$$

Definition 6.2.11. (Distribuciones A priori y A Posteriori)
Sea $A_1, A_2, \cdots$ una partición finita o contable de Ω con $P(A_i) > 0$ para todo i. Si B es un elemento de $\mathfrak{J}$ con $P(B) > 0$, entonces $(P(A_n))_n$ es llamada la distribución "a priori", esto es, antes de que suceda B, y $(P(A_n|B))_n$ es llamada la distribución "a posteriori", esto es, depués que suceda B.

Definition 6.2.12. (Eventos independientes) Dos eventos A y B se dicen ser independientes si y sólo si:

$$P(A \cap B) = P(A)P(B).$$

Si no se cumple la condición anterior los eventos serían dependientes.

Definition 6.2.13. (Familia independiente) Una familia de eventos $\{A_i \,|\, i \in I\}$ se dice ser independiente si

$$P\left(\bigcap_{i \in J} P(A_i)\right) = \prod_{i \in J} P(A_i)$$

para cada subconjunto finito $J \neq \emptyset$ de I. Estos eventos son mutuamente independientes.

Definition 6.2.14. (Eventos independientes por parejas)
Una familia de eventos $\{A_i \,|\, i \in I\}$ se dice ser independiente por parejas (2×2) si:

$$P(A_i \cap A_j) = P(A_i)P(A_j) \text{ para todo } i \neq j.$$

Independencia de pares no implica independencia de familia o eventos mutuamente independientes.

6.2.4. Probabilidad geométrica

Sea $(\Omega, \mathfrak{J})$ un espacio medible, y asuma m cómo una medida geométrica, tal como longitud, área o volumen, definidas sobre $(\Omega, \mathfrak{J})$. Definimos la probabilidad de un evento A como:

$$P(A) = \frac{m(A)}{m(\Omega)} \tag{6.4}$$

6.3. Variables aleatorías y sus distribuciones

6.3.1. Definiciones y propiedades

Definition 6.3.1. (Variable aleatoria) Sea $(\Omega, \mathfrak{J}, P)$ un espacio de probabilidad. Una variable aleatoria es una función $X\colon \Omega \to \mathbb{R}$ tal que, para toda $A \in B$, $X^{-1}(A) \in \mathfrak{J}$, donde B es la σ-álgebra de Borel.sobre $\mathbb{R}$.

Bibliografía

[1] C.B. Liliana, A. Arunachalam, D. Selvamuthu *Introduction to probability and stochastic processes with applications*, 1ª ed., New Jersey: John Wiley & Sons, Inc, 2012.

[2] R.A. Brualdi. *Introductory Combinatorics*, 5ª ed., Republic of China: Prentice-Hall, 2009.

[3] "Uso de Jupyter Notebook en un entorno virtual", 2022. [En linea]. Disponible en: `https://es.acervolima.com/uso-de-jupyter-notebook-en-un-entorno-virtual/`. [Accedido: 21-jun-2022]

[4] "Create a Next.js App", 2022.[En linea]. Disponible en: `https://nextjs.org/learn/basics/create-nextjs-app`. [Accedido: 15-jun-2022]

[5] M.P. Do Carmo *Differential Geometry Of Curves & Surfaces*, 1ª ed., USA: Prentice-Hall, 1976.

[6] L. Leithold, , 2007. Editorial Harla. Méjico. *Cálculo con Geometría Analítica.*, 1ª ed., USA: Prentice-Hall, 1976.