Documentación

Javier Andrés Causil Martínez

Universidad de Antioquia - Científico de datos

Portafolio

Causil

Índice general

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.

Capítulo 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 2

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"

2.1. ¿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.

2.2. 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.

2.3. 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,

2.4. 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 3

Proyectos

3.1. 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 4

Course Structure & Outline

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

4.1. 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

Capítulo 5

Ingeniero de datos

Es un profesional que tiene pasión por la automatización, confianza en el análisis y los datos. Sus habilidades se enfocan en la extracción, transformación y carga de datos (ETL). Automatización de procesos de carga de datos. Procesamiento de datos de diversas fuentes y en cómputo paralelo. Tiene conocimientos en programación con Python y bases solidas de ingeniería de software, manejo de datos estructurados y no estructurados, cómputo, almacenamiento y bases de datos en la nube. ETL con herramientas como SQL, Apache Spark, Airflow, Hadoop, AWS, Google Cloud, Azure, entre otros.

Capítulo 6

Big-Data

Qué es Map Reduce?

Paradigma de programación que permite trabajar en ambientes distribuidos usando una gran cantidad de servidores que conforman un clúster.

1. **Map:** recibe un conjunto de datos y lo transforma en un segundo conjunto de datos cuyos elementos se presentan en tuplas (clave - valor)
2. **filter:**
3. **Reduce:** Utiliza como entrada las tuplas generadas por el map y genera un conjunto de datos de salida reducido a partir de la combinación de los datos recibidos

¿Cómo trabaja MapReduce?

Tiene tres etapas Map(), Shuffle y Reduce().

Qué es un RDD

Un RDD (Dataset distribuido Resiliente)

Transformaciones: Las transformaciones se aplican a RDD y devuelven un RDD.

1. **Map:**
2. **filter:**
3. **Reduce:**
4. **flatMap:**
5. **mapPartitions:**
6. **mapPartitionsWithIndex:**
7. **sample:**
8. **union:**

9. **intersection:**
10. **distinct:**
11. **groupByKey:**

Acciones:

1.

https://spark.apache.org/docs/latest/api/python/index.html

Parte II

Metodologías

Capítulo 7

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

Bases de Datos

Capítulo 8

Relaciones

Objetos

Entidad rectangulos

Capítulo 9

MySQL

9.1. Funciones de control de flujo

Case: Esta expresión nos permite realizar la condición y devolver el primer valor que cumpla con dicha condición

Example 9.1.1.
Primer ejemplo

```
select
case 1
when 1 then 'uno'
when 2 then 'dos'
else 'otro n\'umero'
end as valor;

segundo ejemplo:

select idFactura, idProducto,

case
when cantidad > 2 then 'M\'as de dos productos vendidos'
when cantidad = 2 then 'Dos productos vendidos'
else 'Menos de dos productos vendidos'
end as cantidad
from detalle_factura;

Tercer ejemplo

select nombre,
case
when email IS NULL then 'No tiene email registrado'
else 'email'
end as email,
pais
from cliente;
```

Podemos ver que es una sentencia muy similar a switch de vuelve el primer caso que cumpla la condición.

IF :

Example 9.1.2.

```
Primer ejemplo
select if(1 < 2, true, false) as resultado;

segundo ejemplo

select
idProducto,
if(cantidad > 1, cantidad*precioUnitario, precioUnitario) as total
from detalle_factura;

Tercer ejemplo

select
nombre,
if(fechaIngreso < '2016-12-31', concat(idEmpleado, '-16'),
if(fechaIngreso < '2017-12-31', concat(idEmpleado, '-17'),
if(fechaIngreso < '2018-12-31', concat(idEmpleado, '-18'),
concat(idEmpleado,'-19')
  )
   )
   ) as codigo
from empleado;
```

IFNULL y NULLIF: IFNULL nos permite evaluar una primera expresión, si esta expresión es null, entonces devolverá el segundo valor pasado por parámetro y NULLIF :

Example 9.1.3.

```
Primer ejemplo:

select ifnull(null, 'texto') as resultado;

Segundo ejemplo:

En este ejemplo devuelve los contactos de la tabla cliente en la
columna nombre si tiene email nos da el email pero si este campo
es null nos devuelve el tel\'efono

select nombre, ifnull(email, telefono) as contacto
from cliente;

Tercer ejemplo:

select nombre,
ifnull( (select email from cliente where idCliente = '14'),
'No tiene eamil registrado' )
```

```
as email
from cliente
where idCliente = '14';

select
nullif(
(select precioUnitario from producto where idProducto = 1),
(select )
)
```

NULLIF:

9.2. Subqueries

Es una declaración select en otra declaración, los subqueries devuelven datos de la consulta principal, los subqueries puede ser utilizados para agregar, actualizar, eliminar, enviar datos.

Example 9.2.1.

```
 Ejemplo n\'umero 1:
 Consiste en traer cuyos empleados tengan mayor salario al promedio:

 select idEmpleado, nombre, salario
 from empleado
 where salario > (select avg(salario) from empleado);

Ejemplo 2: Seleccionamos los empleados que no pertenezcan al departamento general:

select nombre, apelllido, idDepartamento
from empleado
where idDepartamento NOT IN (select idDepartamento
                             from departamento
                             where nombre like "%general%"
                              );
Ejemplo 3: facturas de los clientes que pertenezacan a Canada o Brasil:

select idCliente, idFactura
from factura
where idCliente IN( select idCliente
from cliente
where pais = 'canada' or pais = 'Brasil'
 );
```

Subconsultas:

Example 9.2.2.

```
select *
from factura
where idCliente = (select idCliente form cliente where nombre = 'Jordi');
```

```
select *
from producto
where precioUnitario <=
(select avg(precioUnitario) from producto where idCategoria = 5)
and idCategoria = 5;
```

comparando subconsultas

Subconsultas:

Example 9.2.3.

```
select idProducto, nombre
from producto
where idProducto = ANY (select idProducto from detalle_factura);
```

1. Introducción y fundamentos del aprendizaje automático.
   - Introducción y fundamentos del aprendizaje automático
   - Introducción, Definiciones, Sklearn Script básico de una simulación en ML
   - Regresión lineal y regresión logística + Taller
   - Taller con dataset grande limpieza de datos + train/test con métrica de scorebásica para regresión y para clasificación
2. Clasificación y selección de modelos
   - Paramétrico vs No paramétrico: K-nn vs Gaussian. Taller sobre los modelos, fronteras de decisión.
   - Selección de modelos, overfitting y regularización.
   - Taller con dataser real selección de modelos"k-fold, k-folds estratificado, k-fold por grupos, Boostrapping.
3. Árboles de decisión y máquinas de vectores de soporte
   - Árboles, Baggind + Random Forest.
   - Máquinas de Vectores de Soporte, One vs All, All vs All.
   - Tallér práctico comparación de modelos de la semana.
4. Boosting y selección de características
   - Boosting: AdaBoost, Gradient Boosting.
   - Selección de características e importancia de variables, PCA, LDA.
   - Taller de aplicación de las técnicas de la semana.

9.3. Unidad 1

9.3.1. Introducción y fundamentos del aprendizaje automático

Metodologías

dasd

1. KDD: Knowledge Discovery in Databases process.

   a) Selección: Selección e integración de los datos objetivo provenientes de fuentes multiples y heterogéneas.

   b) Procesamiento:

   - Eliminación de ruido y datos aislados o outliers.
   - Uso del conocimiento previo para eliminar las inconsistencias y los duplicados.
   - Escogencia y uso de estrategias para manejar la información faltante en los datasets.

   c) Transformación: Conversiń de los atributos

   - Preparación de los datos para el análisis.
   - Uso de transformaciones de atributos como: numerización, discretización, etc.
   - El resultado es conjunto de filas y columnas denominado vista minable.

   d) Minería de datos:

   - Análisis de los patrones o relaciones a descubrir.
   - Se comprende de 3 pasos:
     - Selección de la tarea.
     - Selección del algoritmo(s).
     - Aplicación/Entrenamiento del algoritmo.

   e) Implementación/Evaluación:

   - Implementación, interpretación o difución del modelo.

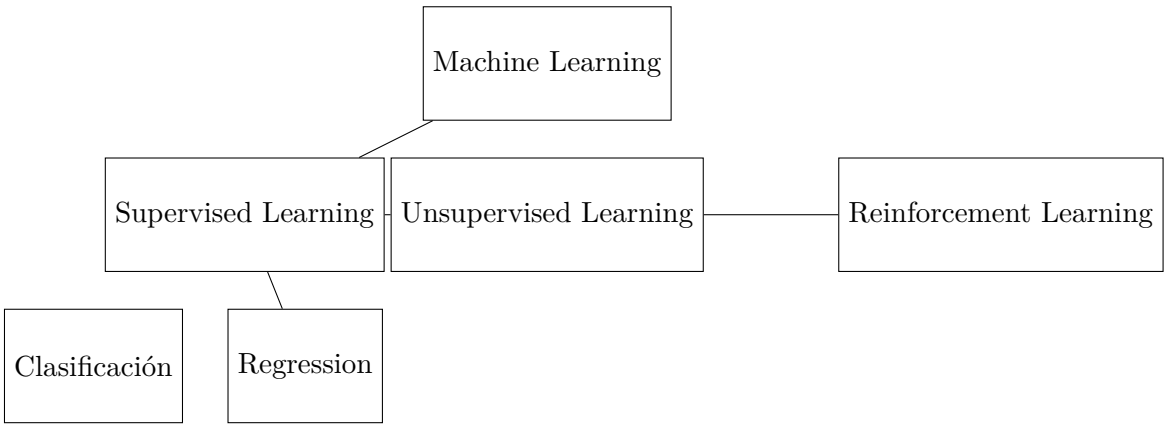
   f) Actualización y monitorización:

   - Consiste en ir revalidando el modelo con cierta frecencia sobre nuevos datos, con el objetivo de detectar si el modelo requiere una actualización.

2. CRISP DM

   - Entendimiento del negocio.
   - Entendimiento de los datos.
   - Preparación de los datos.
   - Modelado.
   - Evaluación.
   - Despliegue.

jlksjads


Machine Learning
Supervised Learning
Unsupervised Learning
Reinforcement Learning
Clasificación
Regression

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.