

Fundamentos de Aprendizaje de Máquina

Diplomado en Desarrollo de Aplicaciones con Inteligencia Artificial

Erick López

Pontificia Universidad Católica del Perú

03 de Agosto del 2019

eelopezo@pucp.edu.pe

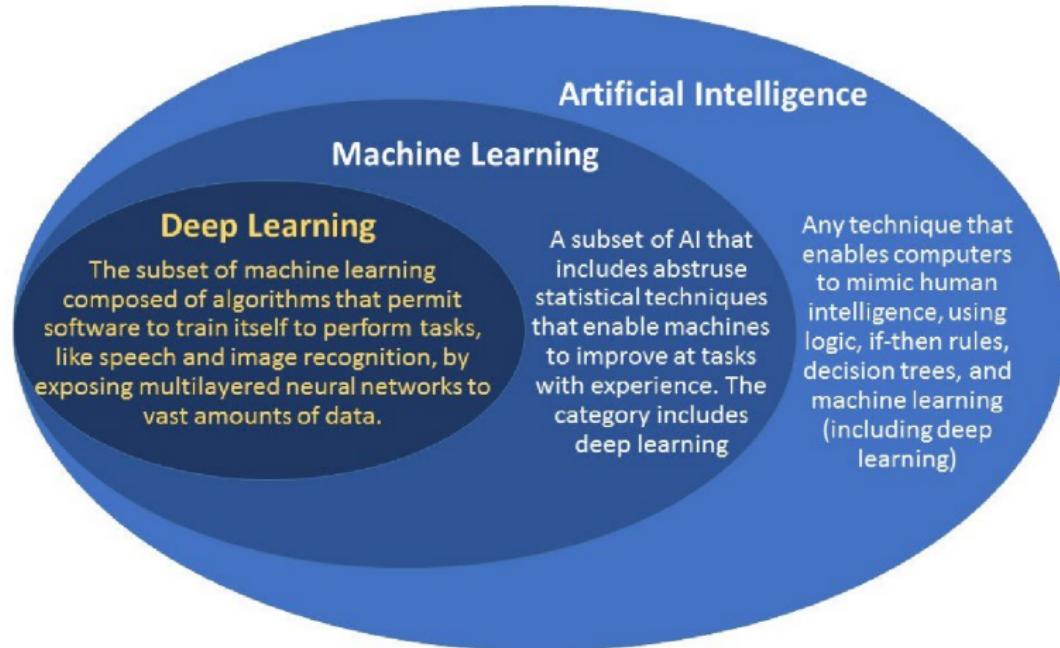
Temas de interés:

- análisis predictivo,
- pronóstico de series temporales,
- ciencia de datos,
- estadística computacional,
- reconocimiento de patrones,
- machine learning,
- inteligencia artificial,
- computación cuántica

¿Qué es el Aprendizaje de Máquina? (machine learning)

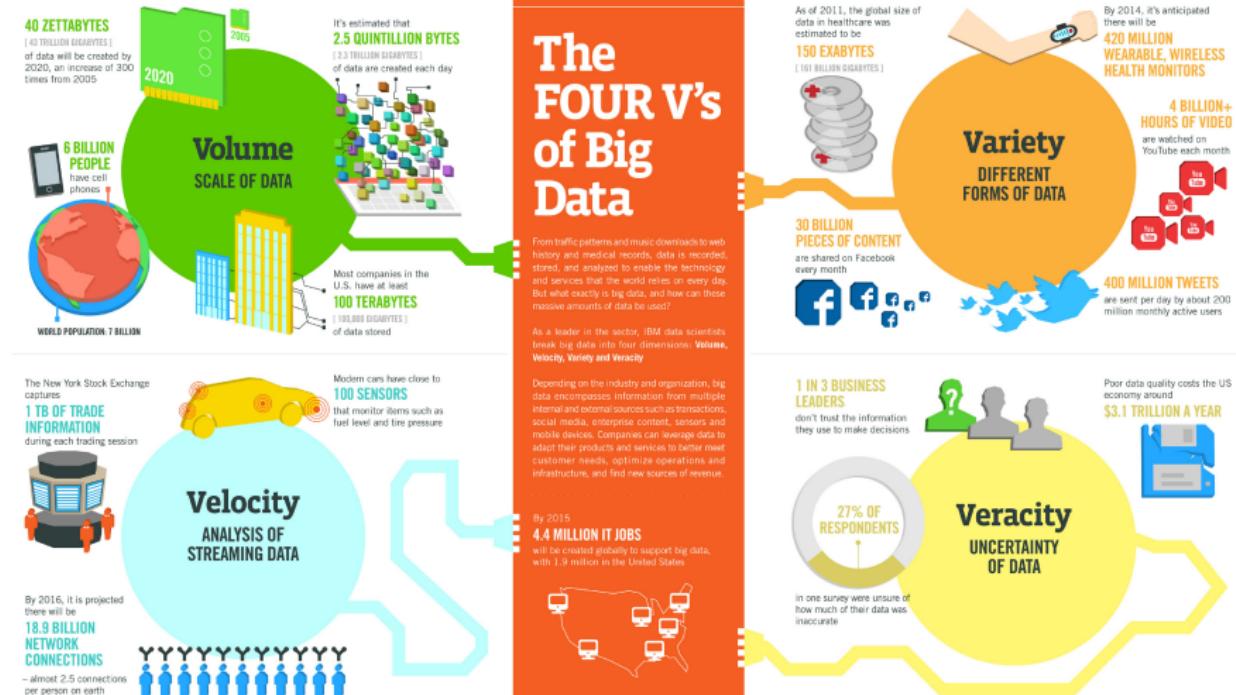
Machine Learning vs Data Science

Machine Learning	Data Science
Desarrolla nuevos modelos (individuales)	Deriva conocimiento a partir de big data, de forma eficiente e inteligente.
Demuestra propiedades matemáticas de los modelos	Comprende propiedades empíricas de los modelos.
Mejora/Valida sobre algunos pequeños conjuntos de datos relativamente limpios	Desarrolla/Usa herramientas que puedan manejar conjuntos de datos masivos.
Publica un paper	Toma acciones!



(https://www.researchgate.net/post/What_is_the_difference_between_artificial_intelligence_AI_and_machine_learning_ML)

Las 4 V's de Big Data



Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, NEPTEC, Gartner

(<https://www.ibmbigdatahub.com/infographic/four-vs-big-data>)



Data mining techniques

Classification

Clustering

Regression

Outer

Sequential
Patterns

Prediction

Association
Rules

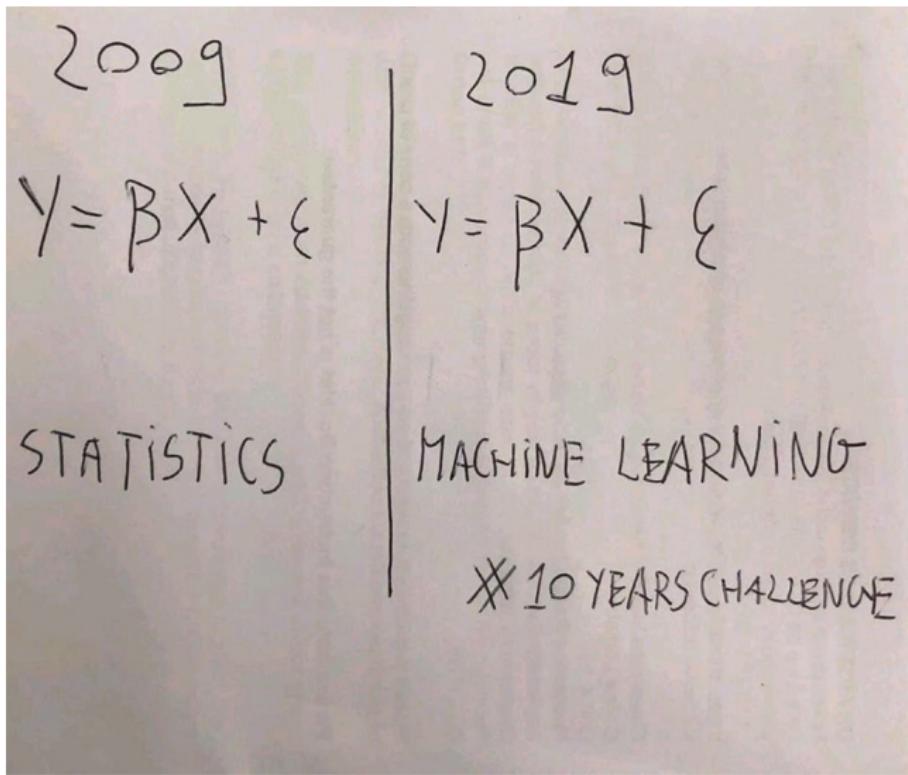
(<https://sforce.co/2OfLbYI>, <https://www.guru99.com/data-mining-tutorial.html>)

¿Qué es el Machine Learning?

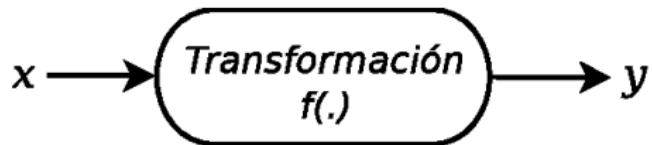
... o dicho de otra forma ...

¿Qué es un modelo de Machine Learning?

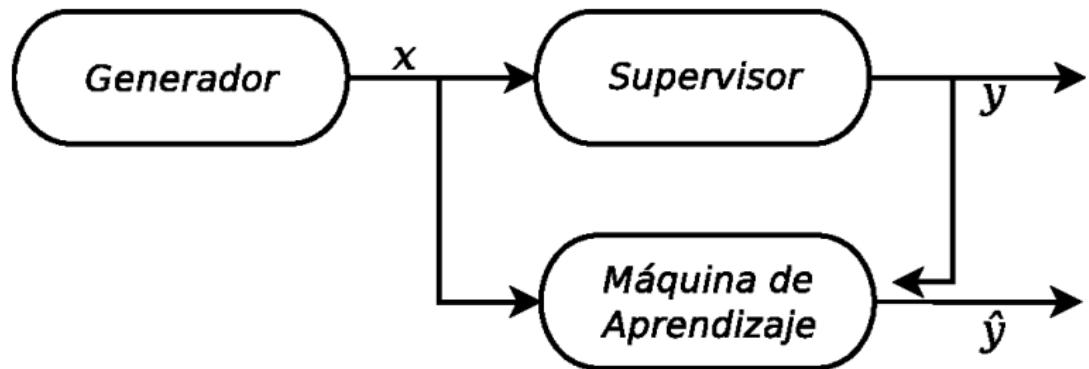
Fake?



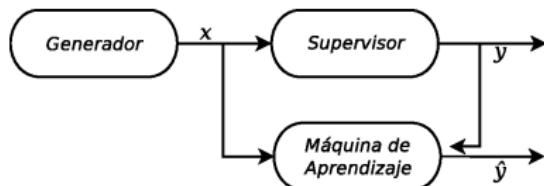
Modelo: Abstracción de la Realidad



Modelo de Machine Learning



Modelo de Machine Learning



El objetivo es retornar un valor \hat{y} lo más cercano a la respuesta del supervisor y para un x^{new} dado.

¿Cómo lo hacemos?

¿Identificamos al supervisor o Imitamos al supervisor?

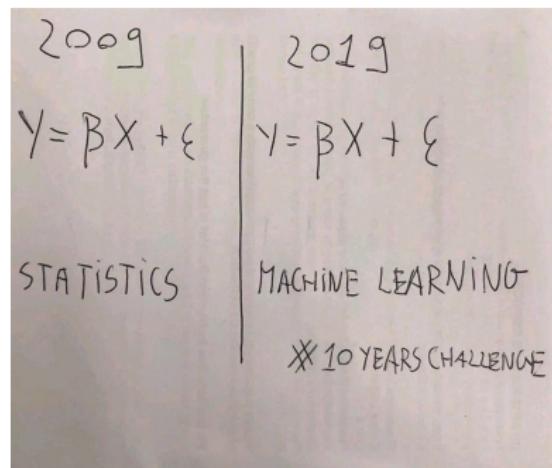
En este caso, imitación es mejor

La imitación es mejor, centra su elección en la calidad de la solución elegida desde un Espacio de Hipótesis, usando algún criterio de optimalidad

- Entonces un modelo de ML trata de imitar el comportamiento del supervisor en términos de la salida que genera al transformar un dato de entrada.
- Cuando la máquina observa un conjunto de entrenamiento $\{(x_i, y_i)\}$, esta muestra es independiente e idénticamente distribuida siguiendo la distribución conjunta $F(x, y) = F(x) \cdot F(y|x)$.
- Por ende, la hipótesis que busca la máquina será una función que mejor se aproxima al comportamiento de $F(y|x)$, basado en la muestra disponible y una función de costo (error).

Modelos Estadísticos vs Modelos de Machine Learning

¿Cuál es la diferencia?



Buscar relaciones subyacentes (patrones) y/u obtener una regla “general” que nos permita adquirir conocimiento sobre un evento de interés

- Identificando variables relevantes
- Descubriendo como interactuan esas variables (como se relacionan)
- Sin imponer supuestos distribucionales
- Generalizando a partir de datos históricos disponibles

datos → regla

Razonamiento Inductivo

¿Donde esta el truco?

¿Todos los modelos de ML siempre funcionan?

¿Donde esta el truco?

¿Todos los modelos de ML siempre funcionan?

!!! NO !!!

¿Donde esta el truco?

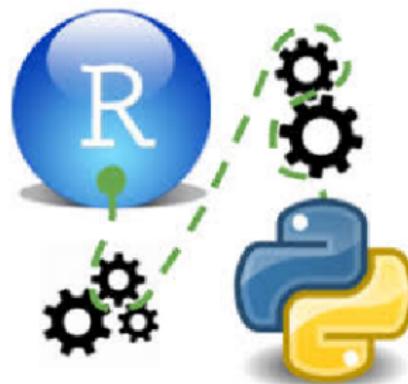
¿Todos los modelos de ML siempre funcionan?

¡¡¡ NO !!!

Los modelos tradicionales de ML también tienen algunos supuestos y/o requisitos que deben cumplirse:

- Preprocesamiento y Transformación de los datos
- Elección adecuada del modelo a utilizar (o modelos)
- Sintonización del modelo (o modelos)
- Diseñar una estrategia para mitigar el underfitting u overfitting
- Elegir una métrica de evaluación y/o estrategia de elección del modelo final (el cual será llevado a producción)

...y cómo lo llevamos a la práctica?



Opciones para trabajar en la nube (free)

- R

<https://rstudio.cloud>

- Python

<https://colab.research.google.com>

Construir el conjunto de datos disponible alimentar un modelo de ML

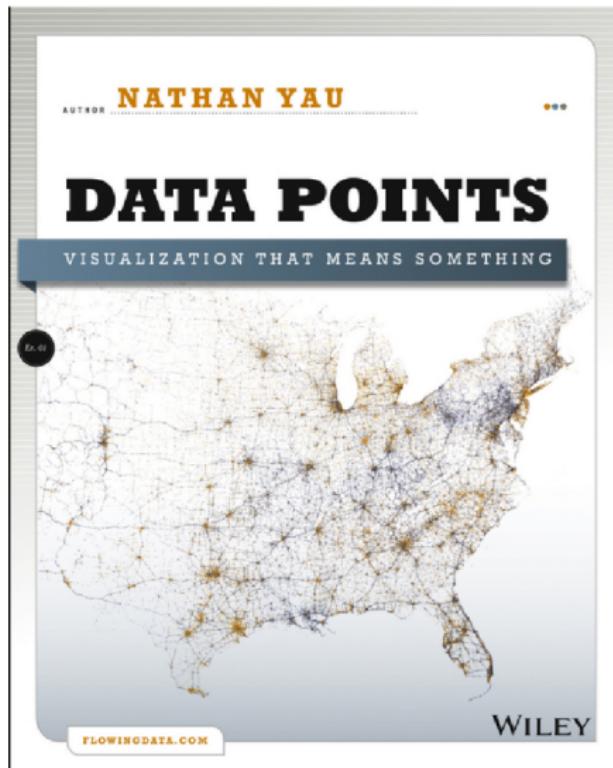
- Crear o adquirir un conjunto de datos de interés, seleccionando un subconjunto de variables o muestras de datos (o ambos).
 - Estructurado → SQL (Structured Query Language)
 - No Estructurado → NoSQL (Not Only SQL)
- Limpiar e integrar los datos seleccionados
 - Valores faltantes, datos con ruidos, datos atípicos, inconsistencias
 - Múltiples fuentes de datos con formatos diferentes, redundancia
- Transformar el dataset para un correcto análisis y modelamiento
 - Estandarizar o Normalizar (escalamiento)
 - Reducción vertical u horizontal
 - Discretizar (categorización de datos numéricos, variables dummy, etc.)
 - **Crear nuevas variables**

Primeros pasos (0)

Iniciamos con la manipulación de la estructura de datos **DataFrame** proporcionado por el modulo **Pandas**

```
import pandas as pd
```

Visualización de Datos

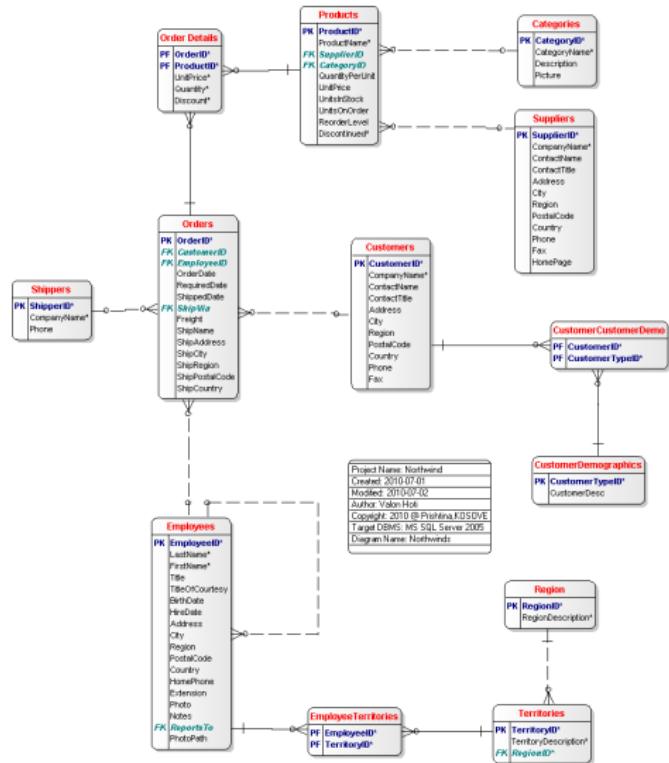


Primeros pasos (1)

Vamos a importar un dataset desde una base de datos MySQL

DB Northwind

<https://theaccessbuddy.wordpress.com/2011/07/03/northwind-database-explained/>



Primeros pasos (2)

Revisamos algunos ejemplos de preprocesamiento y transformación tradicionales

Datos Faltantes (missing values)

Técnicas de imputación tradicionales:

- Medida de tendencia central: promedio, mediana, moda, etc...
- Regresión Lineal
- Regresión Logística?
- Modelo no lineal: redes neuronales, regresión polinomial, etc...
- *Modelo de recomendación basado en similaridades*

¿Qué porcentaje es aceptable imputar?

Hint: Segmentar (si es posible) las observaciones antes de imputar

Escalamiento de los datos

- A veces la data presenta una distribución exponencial en sus valores, entonces es recomendable escalarlo aplicando una transformación logarítmica.
- De esta forma se evita ciertas distorsión/ruido en la visualización de los datos.
- Es posible que sea recomendable aplicar otra transformación.

Variables dummy

Son aquellas variables “indicador”, los cuales indican (valga la redundancia) la presencia o ausencia de un valor específico.

Por ejemplo, si la muestra de una variable fuera

A

A

B

A

su equivalente a dummy sería

1 0

1 0

0 1

1 0

Variables Dicotómicas

Son aquellas variables cualitativas que solo pueden tomar dos valores, entonces:

- Preservar sus valores originales
- Transformarlo a dos números enteros a elección (recomendado por el experto del negocio)
- Transformarlo a escala binaria: 0 , 1
- Transformarlo a una variable dummy

Variables categóricas

- Ordinales: es posible transformarla a una escala numérica de valores enteros (\mathbb{Z}).
- Nominales: Solo existiría una transformación a escala numérica por convención de los expertos del negocio, pero no es recomendable.

En ambos casos, es posible transformarlo a una variable dummy.

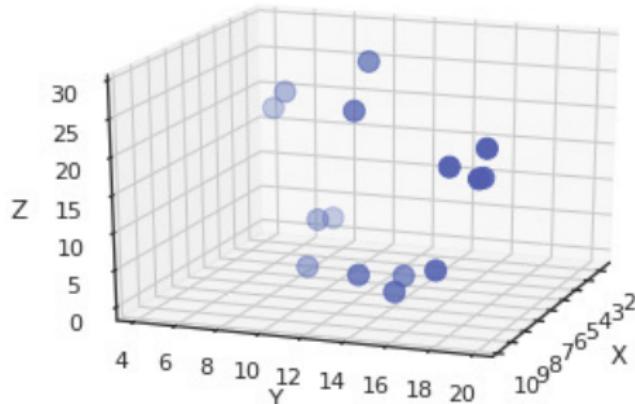
Crear nuevas variables

También es recomendable crear nuevas variables a partir de las características disponibles, permitiendo encontrar nuevas relaciones en la etapa de modelamiento

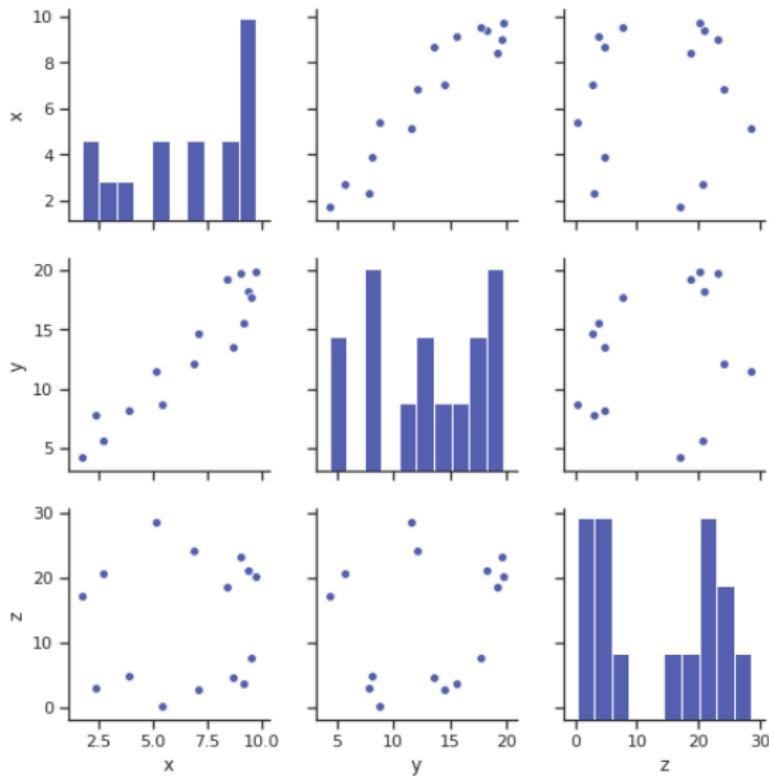
- Explorando interacciones sugeridas por el experto del negocio
- Las limitaciones teóricas y de tiempo o hardware, dificultan descubrir relaciones interesantes que entreguen mayor información

Reducción de dimensionalidad

	x	y	z
0	1.718904	4.298863	17.119145
1	2.343304	7.760359	2.926768
2	2.697904	5.645916	20.745013
3	3.907229	8.126480	4.777099
4	5.108628	11.523384	28.599180
5	5.400889	8.696050	0.220503
6	6.856521	12.066713	24.068945
7	7.058479	14.572488	2.726902
8	8.414407	19.144303	18.684308
9	8.650349	13.528268	4.757091
10	8.985672	19.639019	23.263027
11	9.132386	15.556656	3.675323
12	9.390741	18.211383	21.055203
13	9.531774	17.731447	7.645226
14	9.729324	19.767204	20.285475



Analizar la correlación entre las variables



Análisis de Componentes Principales

Los componentes principales corresponden a ejes de un nuevo sistema de coordenadas. Por ejemplo, basado en los datos anteriores, se pasará de un espacio de 3 dimensiones a otro espacio de 2 dimensiones:

$$PCA : (x, y, z) \longrightarrow (x^*, y^*)$$

donde

$$x^* = \alpha_1 x + \alpha_2 y + \alpha_3 z$$

$$y^* = \beta_1 x + \beta_2 y + \beta_3 z$$

es decir, la transformación que se aplica es una combinación lineal, provocando una proyección de los datos originales al nuevo espacio de menor dimensión.

Al aplicar PCA, los nuevos ejes son ortogonales, por lo cual, son vectores bases de un nuevo espacio vectorial.

Análisis de Componentes Principales

```
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler

# Sea la variable X el dataframe con los datos, entonces

scaler = StandardScaler()
X = scaler.fit_transform(X)

pca = PCA(2)
pca.fit(X)

print(pca.explained_variance_)
print(pca.explained_variance_ratio_.cumsum())
```

Output

```
[2.12361594 1.03875472]
[0.66068052 0.98384865]
```

Análisis de Componentes Principales

```
pca.components_
```

```
[ [-0.68657205 -0.69984144 -0.19708062]  
[-0.2086277 -0.07003179 0.97548451] ]
```

$$PCA : (x, y, z) \longrightarrow (x^*, y^*)$$

$$x^* = -0,6866 \cdot x - 0,6998 \cdot y - 0,1971 \cdot z$$

$$y^* = -0,2086 \cdot x - 0,0700 \cdot y + 0,9755 \cdot z$$

Análisis de Componentes Principales

Los datos transformados al nuevo espacio se obtiene con

```
pca.transform(X)
```

Output

```
array([[ 2.34234986,  0.8798128 ],
       [ 2.0095969 , -0.69203503],
       [ 1.83730341,  1.16459047],
       [ 1.53111594, -0.62278733],
       [ 0.26434082,  1.71881914],
       [ 1.17687269, -1.21779916],
       [-0.15012818,  1.10768334],
       [-0.09605425, -1.16314327],
       [-1.3974096 ,  0.3321317 ],
       [-0.391435 , -1.05789417],
       [-1.70386128,  0.75863391],
       [-0.76744944, -1.23482955],
       [-1.56204575,  0.51786187],
       [-1.24922997, -0.88179473],
       [-1.84396616,  0.39074999]])
```

Para regresar los datos al espacio vectorial original

```
pca.inverse_transform(X2)
```

Para regresar los datos a la escala original

```
scaler.inverse_transform(X2)
```

¡¡¡Alerta!!!

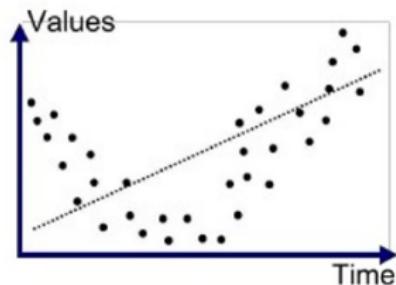
Antes de empezar a modelar, recordar

Los **modelos tradicionales** de aprendizaje de máquina requieren que el conjunto de observaciones (los datos) sean

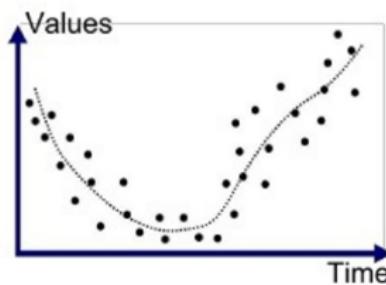
INDEPENDIENTES

(y suponen que son idénticamente distribuidos)

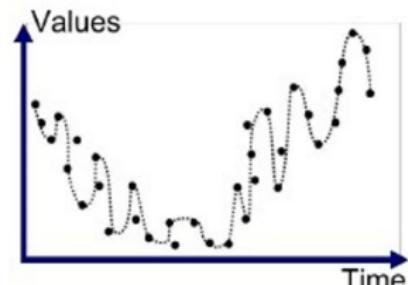
Underfitting and Overfitting



Underfitted



Good Fit/Robust



Overfitted

Overfitting



Training and Testing Dataset

Available dataset

Training dataset

Testing dataset

Training, Validation and Testing Dataset

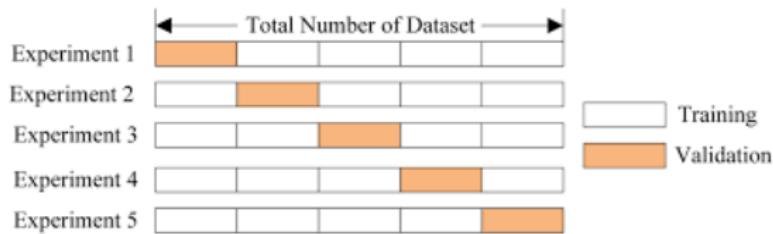
Available dataset

Training dataset

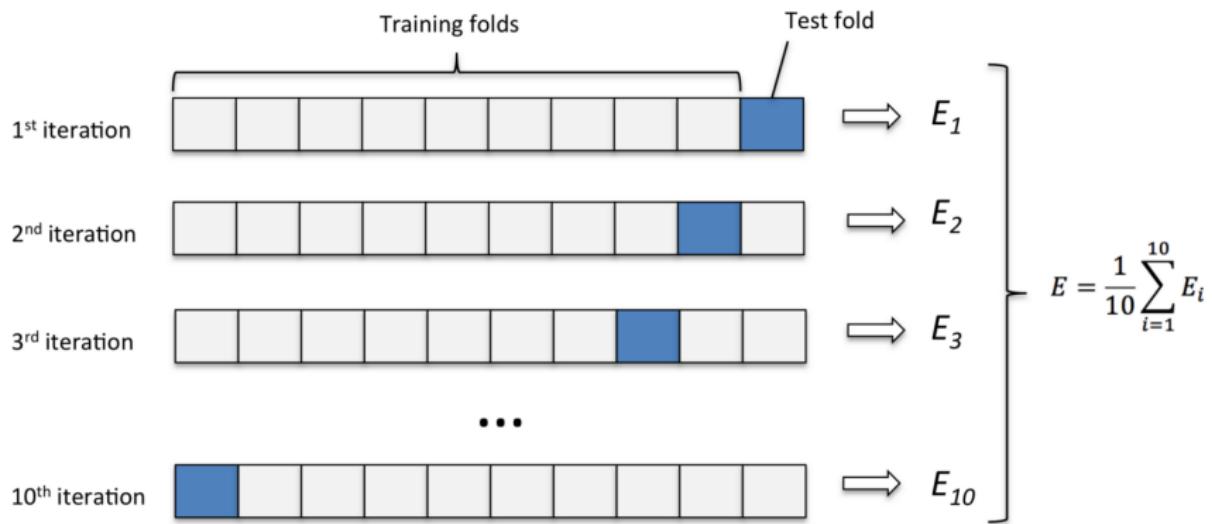
Validation dataset

Testing dataset

Cross-Validation



Cross-Validation



Preguntas?

eelopezo[at]pucp.edu.pe
LATEX