



Introducción a la Ciencia de Datos

Maestría en Ciencias
de la Computación

Dr. Irvin Hussein López Nava



**THE DATASET
HAS NO LABELS SO...**

**USE UNSUPERVISED
LEARNING**





Classical Machine Learning

Task Driven

Supervised Learning

(Pre Categorized Data)

Classification

(Divide the socks by Color)

Regression

(Divide the Ties by Length)

Data Driven

Unsupervised Learning

(Unlabelled Data)

Clustering

(Divide by Similarity)

Association

(Identify Sequences)

Dimensionality Reduction

(Wider Dependencies)

Obj: Predictions & Predictive Models

Pattern/ Structure Recognition



06

Aprendizaje no supervisado

Recapitulando...

- Una tarea que involucra aprendizaje automático puede no ser lineal, pero tiene una serie de pasos bien conocidos:
 - Definición del problema.
 - Preparación de datos.
 - Aprendizaje de un modelo subyacente.
 - Mejorar el modelo subyacente mediante evaluaciones cuantitativas y cualitativas.
 - Despliegue del modelo.

Tipos de modelos de aprendizaje

- Para abordar un nuevo problema (o proyecto de curso) es importante identificar y definir el problema de la mejor manera posible y aprender un modelo que capture la **información significativa** de los datos.
- Si bien los problemas en el reconocimiento de patrones y el aprendizaje automático pueden ser de varios tipos, se pueden clasificar en términos generales en tres categorías:
 - Aprendizaje supervisado (abordado)
 - Aprendizaje no supervisado (ahora)
 - Aprendizaje por refuerzo (para otra ocasión)

Aprendizaje supervisado

El sistema se presenta con ejemplos de entradas y sus salidas deseadas, proporcionadas por un "maestro", y el objetivo es aprender una regla general que asigna entradas a salidas.

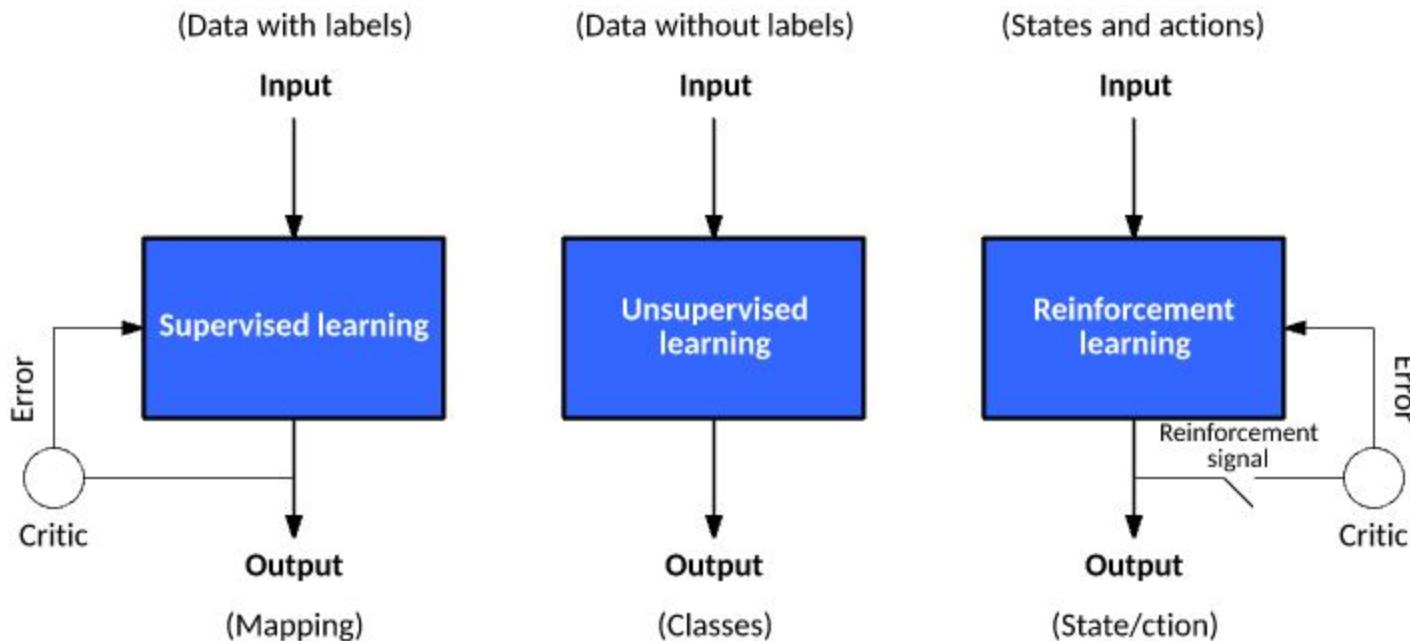
Aprendizaje no supervisado

No se asignan etiquetas al algoritmo de aprendizaje, dejándolo solo para que encuentre estructuras en las entradas. Puede ser un objetivo en sí mismo (descubrir patrones ocultos) o un medio para lograr un fin (aprendizaje de funciones).

Aprendizaje por refuerzo

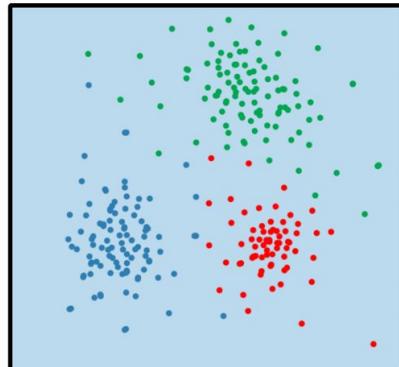
Un sistema interactúa con un entorno dinámico en el que debe cumplir un objetivo. El sistema recibe retroalimentación en términos de recompensas y castigos a medida que navega por su espacio de configuraciones.

En diagramas

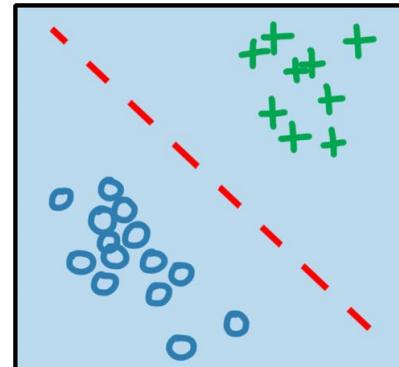


Con gráficos

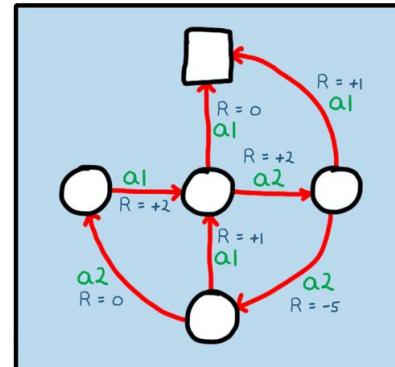
unsupervised
learning



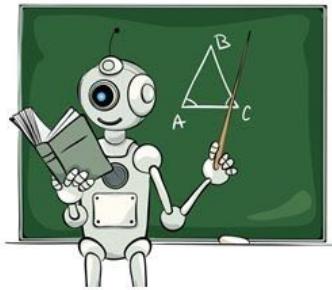
supervised
learning



reinforcement
learning

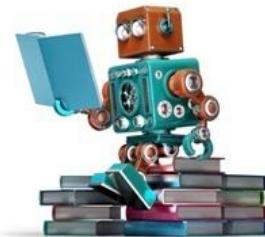


Con dibujos



**Supervised
Learning**

VS



**Unsupervised
Learning**

VS



**Reinforcement
Learning**

Aprendizaje no supervisado

- En algunos problemas, los datos de entrenamiento constan de un conjunto de vectores de entrada X sin ningún valor objetivo (y) correspondiente.
- El objetivo de estos problemas de **aprendizaje no supervisado** puede ser descubrir grupos de ejemplos similares dentro de los datos, lo que se denomina **agrupamiento**, o determinar cómo se distribuyen los datos en el espacio, lo que se conoce como **estimación de densidad**.
- Para expresarlo en términos más simples, para un espacio de n muestras x_1 a x_n , no se proporcionan etiquetas de clase para cada muestra, por lo que se conoce como aprendizaje sin maestro.

Algunos problemas

- En general, el aprendizaje no supervisado es más difícil que el supervisado...
- ¿Cómo saber si los resultados son significativos al no disponer de etiquetas de respuesta?
- Dejar que el experto observe los resultados (evaluación externa).
- Definir una función objetivo sobre la agrupación (evaluación interna).

¿Y por qué es necesario?

- Etiquetar grandes conjuntos de datos es muy costoso, por lo que muchas veces, sólo se pueden etiquetar manualmente unos pocos ejemplos.
- En otras ocasiones, no se sabe en cuántas o en qué clases están divididas los datos.
- Es posible querer utilizar el agrupamiento para conocer la estructura de los datos antes de diseñar un clasificador.

Parametric Unsupervised Learning

- En este caso, se asume una distribución paramétrica de los datos.
 - Supone que los datos de la muestra proceden de una población que sigue una distribución de probabilidad basada en un conjunto fijo de parámetros.
- Teóricamente, en una familia de distribuciones normales, todos los miembros tienen la misma forma y están parametrizados por la media y la desviación estándar.
 - Eso significa que si se conocen el valor de tendencia central y su dispersión, y que la distribución es normal, se conoce la probabilidad de cualquier observación futura.

Parametric Unsupervised Learning

- El **aprendizaje paramétrico no supervisado** implica la construcción de modelos de mezclas gaussianas y el uso del algoritmo *Expectation - Maximization* para predecir la clase de la muestra en cuestión.
- Este caso es mucho más difícil que el aprendizaje supervisado estándar porque no se dispone de etiquetas de respuesta y, por tanto, no se dispone de una medida correcta de la precisión para comprobar el resultado.

Non-parametric Unsupervised Learning

- En la versión no parametrizada del aprendizaje no supervisado, los datos se agrupan en *clústeres*, donde cada clúster ('con suerte') dice algo sobre las categorías y clases presentes en los datos.
- Este método suele utilizarse para modelar y analizar datos con muestras de tamaño pequeño.
- A diferencia de los modelos paramétricos, los modelos no paramétricos no requieren que el modelador haga ninguna suposición sobre la distribución de la población, por lo que a veces se denominan **métodos sin distribución**.

6.1

Agrupamiento

¿Qué es el Agrupamiento?



Definiciones

ChatGPT

Es una técnica en el campo del aprendizaje automático que se utiliza para dividir un conjunto de datos en grupos o clústeres, de tal manera que los elementos dentro de un mismo grupo sean más similares entre sí que con los elementos de otros grupos.

<https://chat.openai.com/>

Wikipedia

El análisis de *clústeres* o agrupamiento es la tarea de agrupar un conjunto de objetos de tal manera que los objetos del mismo grupo (*clúster*) sean más similares entre sí que con los de otros grupos (*clústeres*).

[https://en.wikipedia.org/wiki/
Cluster_analysis](https://en.wikipedia.org/wiki/Cluster_analysis)

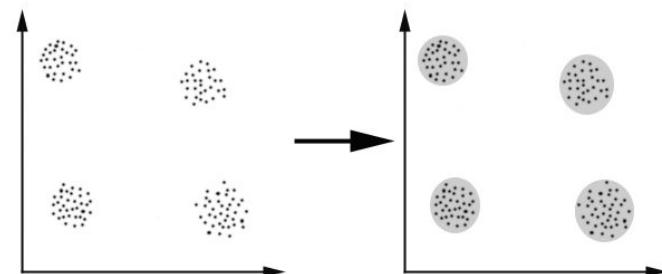
Gemini

Es un tipo de aprendizaje no supervisado que se utiliza para encontrar grupos de datos que son similares entre sí. Los grupos se denominan clústeres. Su objetivo es identificar los grupos de datos que comparten características comunes.

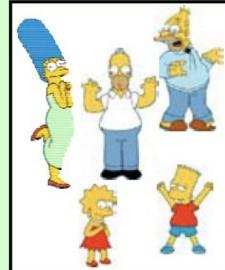
<https://gemini.google.com/>

¿Qué es *clustering*?

- El agrupamiento puede considerarse el problema de aprendizaje no supervisado **más importante**; así, como cualquier otro problema de este tipo, trata de encontrar **una estructura** en una colección de datos sin etiquetar.
- Un cluster es, por tanto, una colección de objetos que son "similares" entre sí y son "disimilares" a los objetos pertenecientes a otros clusters.
- i.e.
 - *high intra-class similarity,*
 - *low inter-class similarity.*



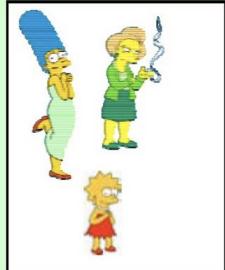
Clustering is subjective



Simpson's Family



School Employees



Females



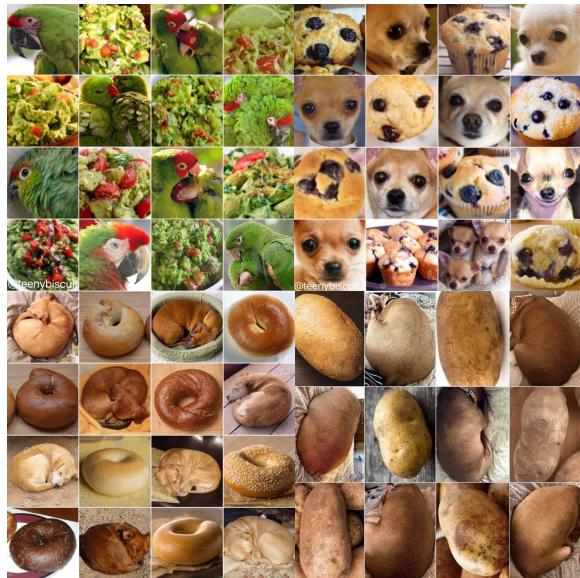
Males

¿Qué es similaridad?

- El verdadero significado de la **similitud** es una cuestión filosófica.
- En un enfoque más pragmático: pensar en términos de **distancia** (en lugar de similitud) entre vectores o correlaciones entre variables aleatorias.



Difícil de definir, pero lo sabemos al verlo



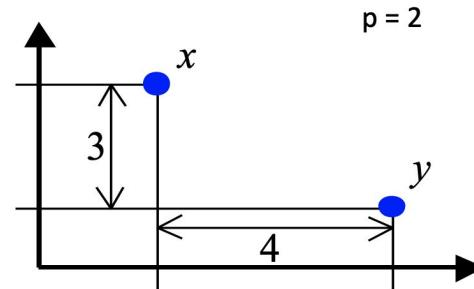
2 grupos?

4 grupos?

8 grupos?

Métricas de distancia

$$\begin{aligned}x &= (x_1, x_2, \dots, x_p) \\y &= (y_1, y_2, \dots, y_p)\end{aligned}$$



Euclidean distance

$$d(x, y) = \sqrt{ \sum_{i=1}^p |x_i - y_i|^2 } \quad 5$$

Manhattan distance

$$d(x, y) = \sum_{i=1}^p |x_i - y_i| \quad 7$$

Sup-distance

$$d(x, y) = \max_{1 \leq i \leq p} |x_i - y_i| \quad 4$$

Coeficientes de correlación

$$\mathbf{x} = (x_1, x_2, \dots, x_p)$$

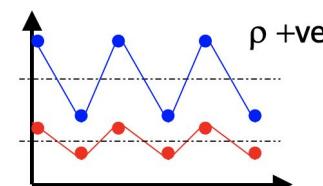
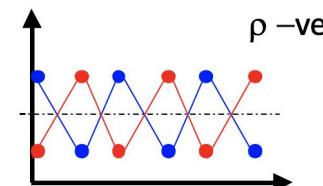
$$\mathbf{y} = (y_1, y_2, \dots, y_p)$$

Random vectors (e.g. expression levels
of two genes under various drugs)

Pearson correlation coefficient

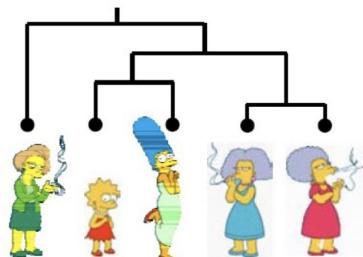
$$\rho(x, y) = \frac{\sum_{i=1}^p (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^p (x_i - \bar{x})^2 \times \sum_{i=1}^p (y_i - \bar{y})^2}}$$

$$\text{where } \bar{x} = \frac{1}{p} \sum_{i=1}^p x_i \text{ and } \bar{y} = \frac{1}{p} \sum_{i=1}^p y_i.$$



Algoritmos de agrupamiento

- Partition algorithms
 - K means clustering
 - Mixture-Model based clustering
- Hierarchical algorithms
 - Single-linkage
 - Average-linkage
 - Complete-linkage
 - Centroid-based



Algoritmos de partición

- Método de partición: Construir una partición de n objetos en un conjunto de K clusters.
- Dado: un conjunto de objetos y el número K
- Buscar: una partición de K clusters que optimice el criterio de partición elegido
 - Óptimo global: enumerar exhaustivamente todas las particiones.
 - Método heurístico eficaz: Algoritmo **K-means**.

K-means

Algorithm

Entrada – Número deseado de *clústers*, k

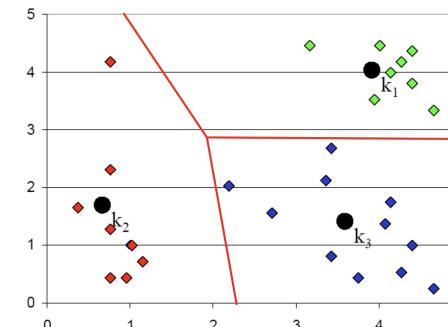
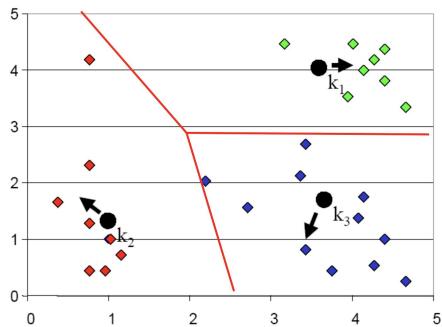
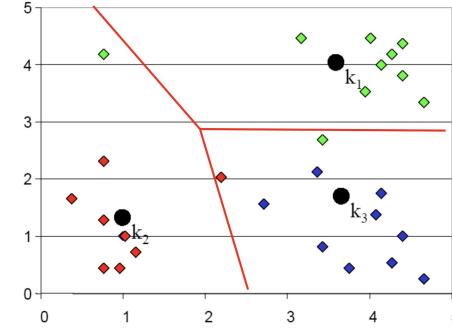
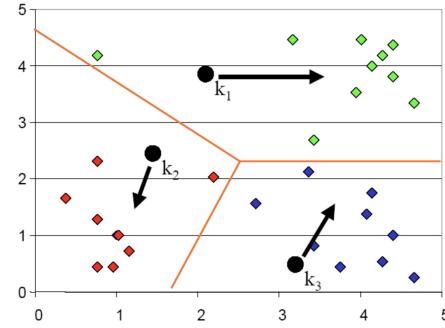
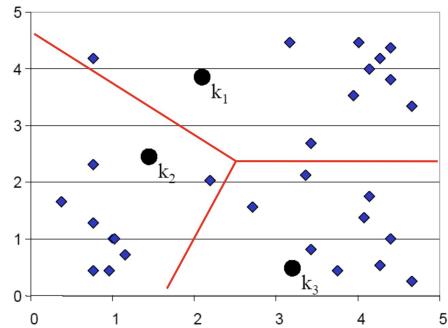
Iniciar – los k centros de los *clusters* (aleatoriamente si es necesario)

Iterar –

1. Asignar puntos a los centros de clúster más cercanos.
2. Volver a estimar los k centros de los *clusters* (a.k.a. centroide), suponiendo que los miembros encontrados anteriormente son correctos.

Finalización – Si ninguno de los objetos ha cambiado de pertenencia en la última iteración, salir. En caso contrario, ir al paso 1.

Ejemplo de K-means



K-means recap

- Iniciar aleatoriamente k centroides.

$$\mu^{(0)} = \mu_1^{(0)}, \dots, \mu_k^{(0)}$$

- Iterar $t = 0, 1, 2, \dots$

- Clasificación: Asignar cada punto $j \in \{1, \dots, m\}$ al centroide más cercano.

$$C^{(t)}(j) \leftarrow \arg \min_{i=1, \dots, k} \|\mu_i^{(t)} - x_j\|^2$$

- Re-centrar: μ_i se convierte en el centroide de sus puntos:

$$\mu_i^{(t+1)} \leftarrow \arg \min_{\mu} \sum_{j:C^{(t)}(j)=i} \|\mu - x_j\|^2 \quad i \in \{1, \dots, k\}$$

- Equivalente a $\mu_i \leftarrow$ promedio de sus puntos.

¿Cómo se optimiza K-means?

- Funciones potenciales $F(\mu, C)$ de centros μ y C asignaciones de puntos:

$$\begin{aligned} F(\mu, C) &= \sum_{j=1}^m \|\mu_{C(j)} - x_j\|^2 \\ &= \sum_{i=1}^k \sum_{j:C(j)=i} \|\mu_i - x_j\|^2 \end{aligned}$$

- Optimal K-means:
 - $\min_{\mu} \min_C F(\mu, C)$

K-means óptimo

- Optimizar la función potencial.

$$\min_{\mu} \min_C F(\mu, C) = \min_{\mu} \min_C \sum_{i=1}^k \sum_{j:C(j)=i} \|\mu_i - x_j\|^2$$

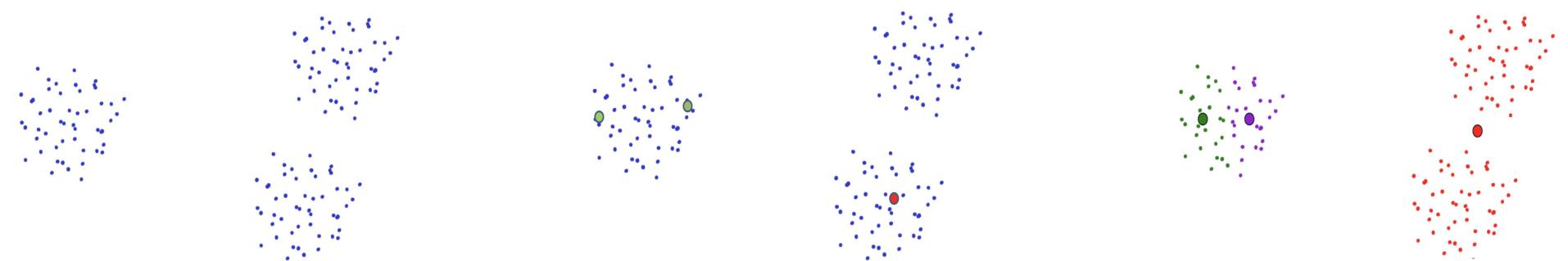
- K-means algorithm:

1. Fijar μ , optimizar C Esperado: asignación de *clusters*
2. Fijar C , optimizar μ Máxima verosimilitud para el centro

- Similar al algoritmo *EM / Baum Welch* para aprender Parámetros HMM.

Elección de la semilla

- Los resultados son muy sensibles a la elección de la semilla.

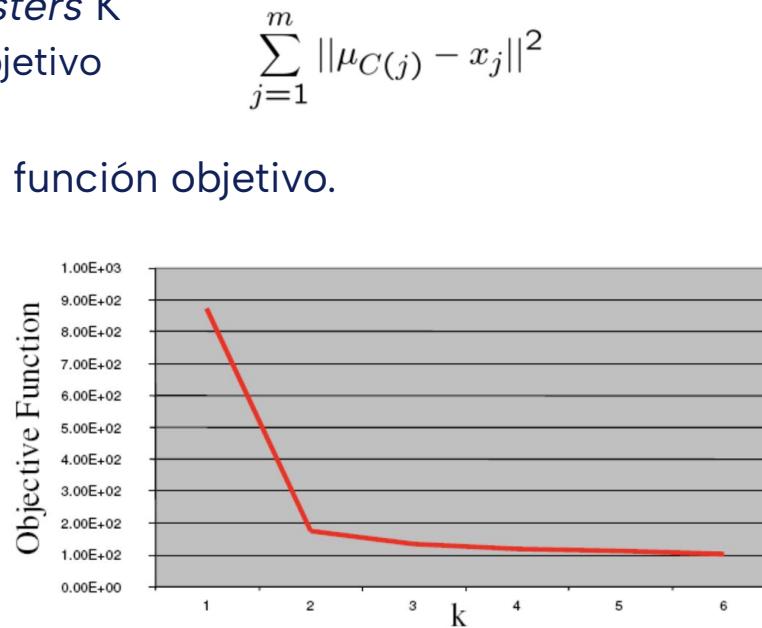


Selección de la semilla

- Los resultados pueden **variар** en función de la selección aleatoria de semillas.
- Algunas semillas pueden dar lugar a una tasa de convergencia pobre, o convergencia a una agrupación subóptima.
 - Probar varios puntos de inicio (¡muy importante!)
 - Probar diferentes implementaciones del algoritmo K-means, e.g., Arthur & Vassilvitskii.
- **Idea clave:** elegir centros alejados entre sí (probabilidad de elegir un punto como centro del cluster \square distancia del centro más cercano elegido hasta el momento).

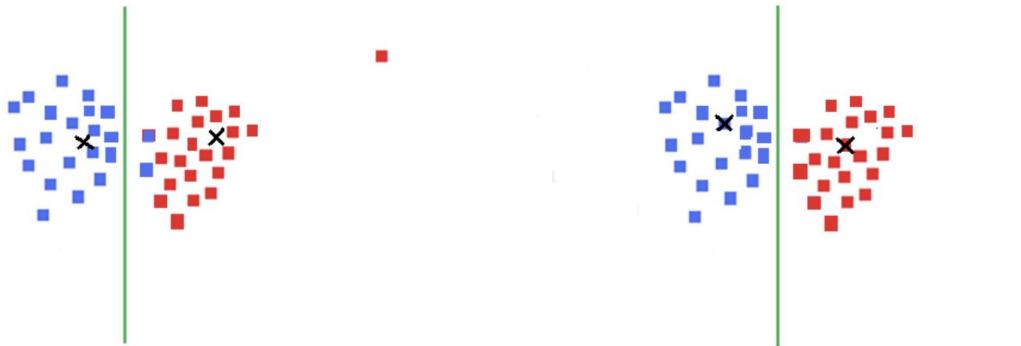
Otros problemas

- Número de *clusters* K
 - Función objetivo
- Buscar sobre la función objetivo.

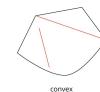
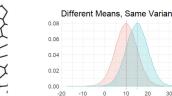
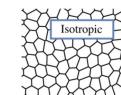


Otros problemas

- Sensibilidad a valores atípicos: usar k-medoids.



- Forma de los clusters: supone *clusters* isótropos, de igual varianza y convexos.



Algoritmos de partición

- K-means.
 - Asignación dura: cada objeto pertenece a un solo clúster.
- Mixture modeling.
 - Asignación suave: probabilidad de que un objeto pertenezca a un clúster.
 - Enfoque generativo.

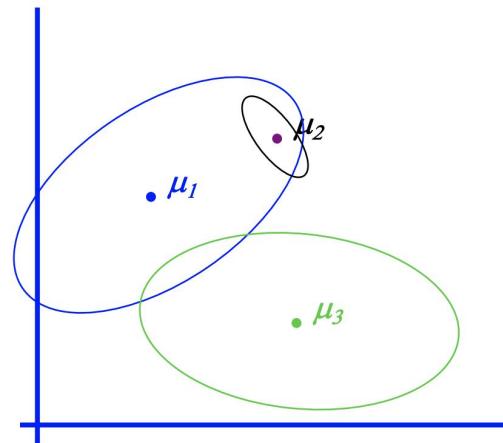
Mixture models

- GMM – Gaussian Mixture Model (distribución multimodal)

$$p(x|y=i) \sim N(\mu_i, \Sigma_i)$$

$$p(x) = \sum_i p(x|y=i) P(y=i)$$

↓ ↓
Mixture component Mixture proportion

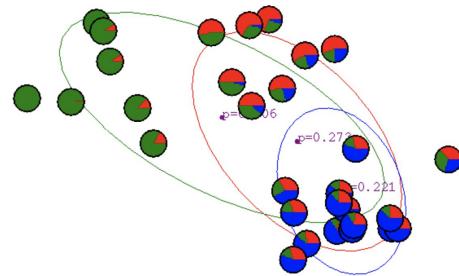


Gaussian Mixture Model

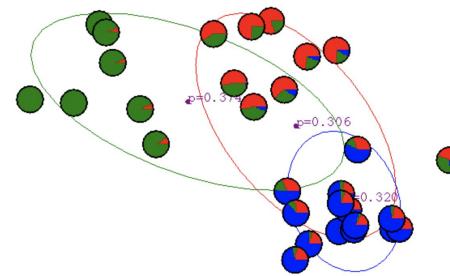
- Hay k componentes.
- El componente i tiene asociado un vector de media μ_i ,
- Cada componente genera datos a partir de una gaussiana con media μ_i y matriz de covarianza S_i ,
- Cada punto de datos se genera de acuerdo a los siguientes pasos:
 1. Elija un componente al azar:
Elija el componente i con probabilidad $P(y=i)$.
 2. Punto de datos $x \sim N(\mu_i, \Sigma_i)$.

Ejemplo de GMM

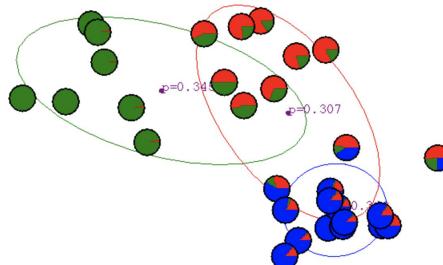
1st iteration



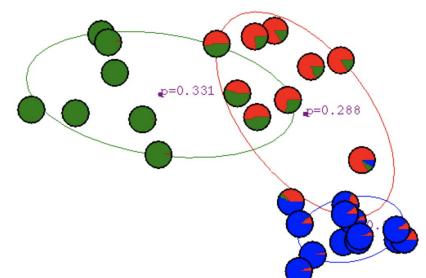
2nd iteration



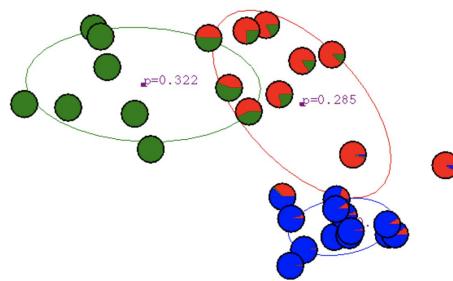
3rd iteration



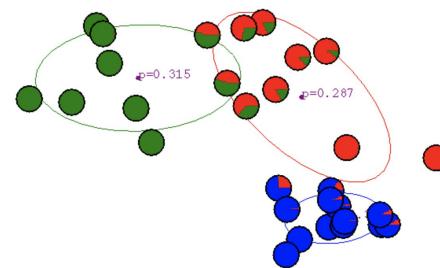
4th iteration



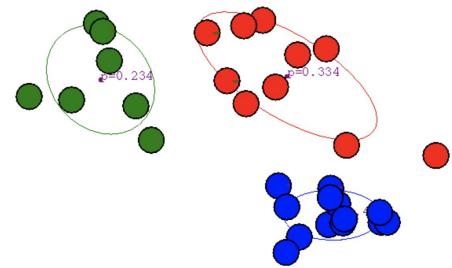
Ejemplo de GMM



5th iteration

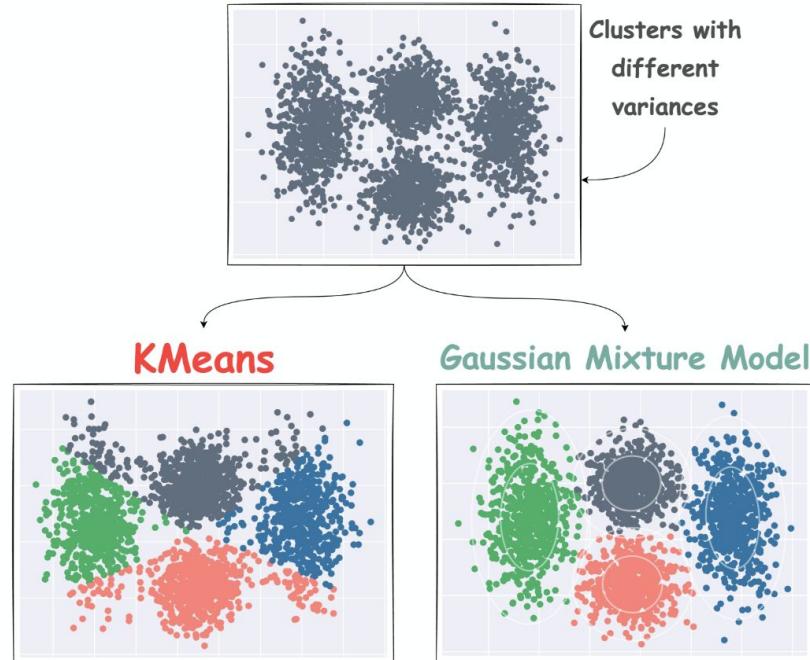


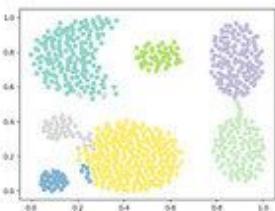
6th iteration



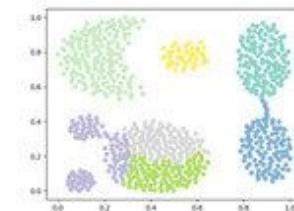
20th iteration

K-means vs GMM

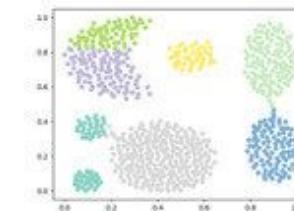




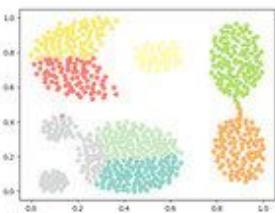
(a) PFA-GMM



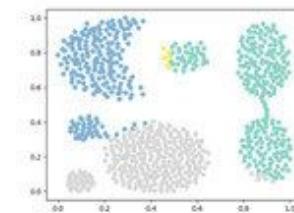
(b) PFA-KM



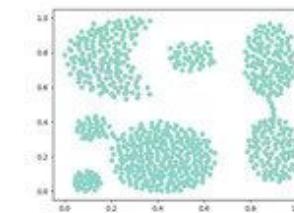
(c) GMM



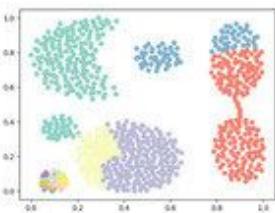
(d) K-means



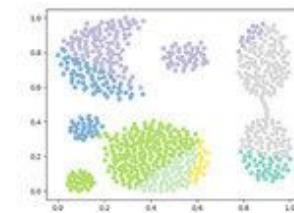
(e) FCM



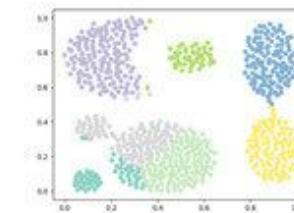
(f) DBSCAN



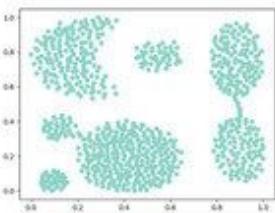
(g) AP



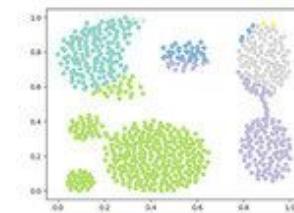
(h) PSO



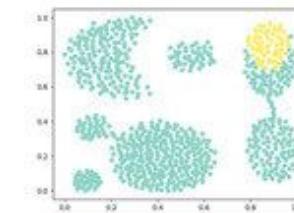
(i) PSO-FCM



(j) GA

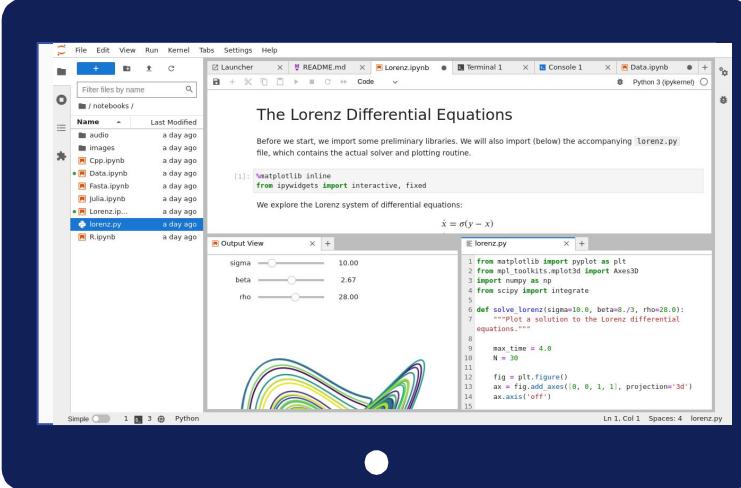


(k) ABC



(l) DE

(Go to live notebook)



The screenshot shows a Jupyter Notebook interface with several tabs open:

- Launcher
- README.md
- Lorenz.ipynb (active tab)
- Terminal 1
- Console 1
- Data.ipynb
- Python 3 (ipykernel)

The main content area displays the "The Lorenz Differential Equations" example. It includes a code cell with the following imports:

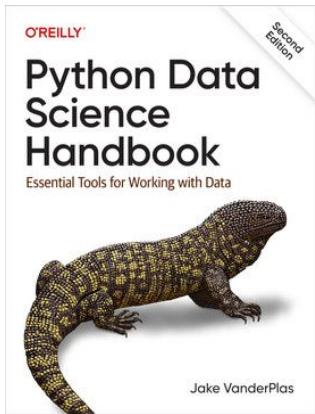
```
[1]: %matplotlib inline
from ipywidgets import interactive, fixed
```

Below the imports, the text states: "We explore the Lorenz system of differential equations:" followed by the differential equation $\dot{x} = \sigma(y - x)$.

The "Output View" shows three sliders for parameters: sigma (10.00), beta (2.67), and rho (28.00). To the right, the "lorenz.py" file is shown with its code:

```
1 from matplotlib import pyplot as plt
2 from mpl_toolkits.mplot3d import Axes3D
3 import numpy as np
4 from scipy import integrate
5
6 def solve_lorenz(sigma=10.0, beta=8./3., rho=28.0):
7     """Plot a solution in the Lorenz differential
    equations."""
8
9     max_time = 4.0
10    N = 30
11
12    fig = plt.figure()
13    ax = fig.add_axes([0, 0, 1, 1], projection='3d')
14    ax.axis('off')
```

Extra Libro



- 05.11-K-Means.ipynb
- 05.12-Gaussian-Mixtures.ipynb

Gracias!

¿Alguna pregunta?

hussein@cicese.mx

<https://sites.google.com/view/husseinlopeznava>



CREDITS: This presentation was based on a template by [Slidesgo](#), and includes icons by [Flaticon](#).