ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN PROFESIONAL

Diplomado en Programación y Aplicaciones de Python

Aplicaciones en Ciencia de Datos e Inteligencia Artificial

Profesor:
Francisco Pérez Galarce

Evaluaciones

Evaluación escrita de conceptos	**20%**		
• 2 controles (contenido teórico e implementación)	10%	**1 de 2**	**Hoy el segundo**
• Prueba final de finalización del curso	10%		
Desarrollo de tareas de programación	**80%**		
• 2 actividades de implementación en clases	20%	**1 de 2**	**Hoy la segunda**
• 2 Mini proyectos	40%	**1 de 2**	
• Repositorio en Github	20%		

Fechas de evaluaciones

Fecha	Actividad/Evaluación
29-10-24	Introducción al aprendizaje de máquina: exploración y procesamiento de datos con Python **Actividad 1 (No evaluada)**
05-11-24	Aprendizaje supervisado con Python : regresiones **Actividad 2 (Evaluada)**
12-11-24	**Actividad 2 (Evaluada)** **Control 1**
19-11-24	Aprendizaje supervisado con Python naive Bayes y métricas de evaluación **Mini Proyecto 1**
26-11-24	Aprendizaje supervisado con Python : decision tree, random forest **Mini Proyecto 1**
03-12-24	Aprendizaje no supervisado con Python: k-means **Actividad 4 (Evaluada) – Control 2**
10-12-24	Redes Neuronales I **Mini Proyecto 2**
17-12-24	Redes Neuronales II **Prueba Final / Portafolio en Github**

- Introducción al aprendizaje de máquinas.
- Procesamiento de Datos
- Aprendizaje supervisado.
- Aprendizaje no supervisado
- Redes Neuronales

Hasta hoy!!

APPRENDIZAJE NO SUPERVISADO

Profesor: **Francisco Pérez Galarce**

Aprendizaje no supervisado

Algoritmos que…

Aprenden desde los datos sin considerer una variable output.

Explorar la base de datos para encontrar patrones ocultos:

- Grupos similares
- Objetos anómalos

Input

$$\mathcal{D} = \{x_1, x_2, \ldots, x_{N-1}, x_N, \}$$

Ejemplos de aprendizaje no supervisado

- Segmentación de clientes
- Identificación de reglas de compras
- Generación de música e imágenes artificiales

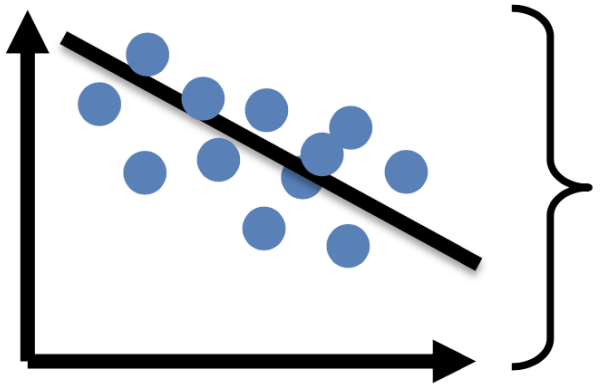
Clustering

Profesor: **Francisco Pérez Galarce**

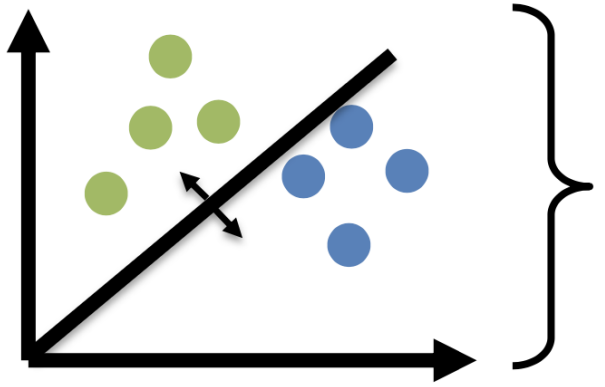
Clustering VS Clasificación Supervisada

Profesor: **Francisco Pérez Galarce**

Clustering VS Clasificación Supervisada

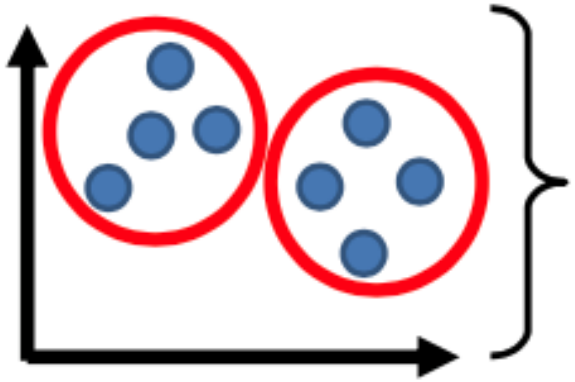


Regresión

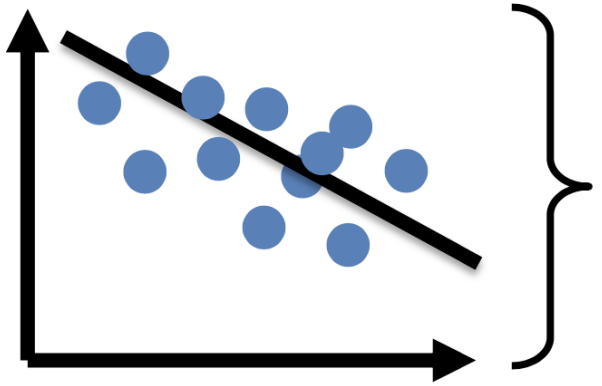


Clasificación

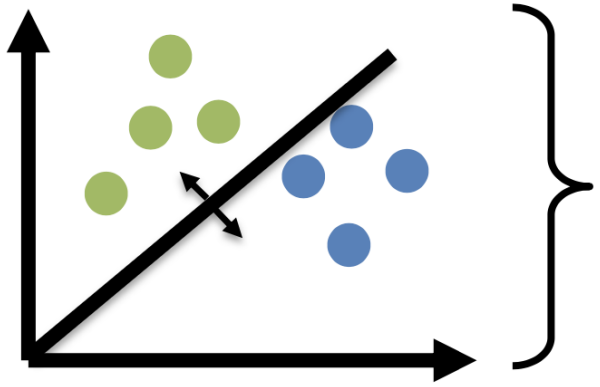
Clustering VS Clasificación Supervisada



Clustering



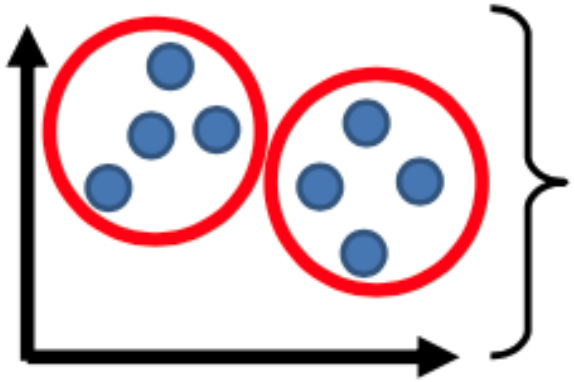
Regresión



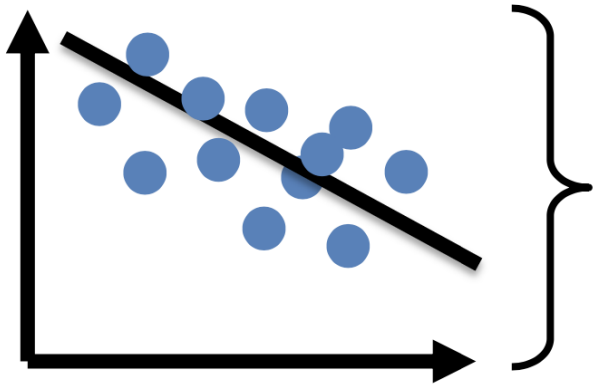
Clasificación

Profesor: **Francisco Pérez Galarce**

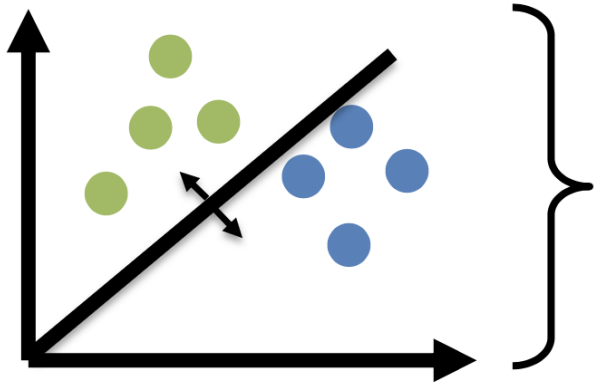
Clustering VS Clasificación Supervisada



Clustering



Regresión



Clasificación

Clustering:
El objetivo es en agrupar objetos con las mismas características.

No tenemos una etiqueta para los datos

Clustering

- Segmentación de clientes
- Optimización de despachos
- Clasificación de documentos
- Detección de fraudes

Profesor: **Francisco Pérez Galarce**

Dada una base de datos DB con N registros y M descriptores.

x_{11}	...	x_{1M}
...	...	...
x_{N1}	...	x_{NM}

Dada una base de datos DB con N registros y M descriptores.

x_{11}	...	x_{1M}
...	...	...
x_{N1}	...	x_{NM}

¿Cómo agrupamos los registros similares?

K-MEANS

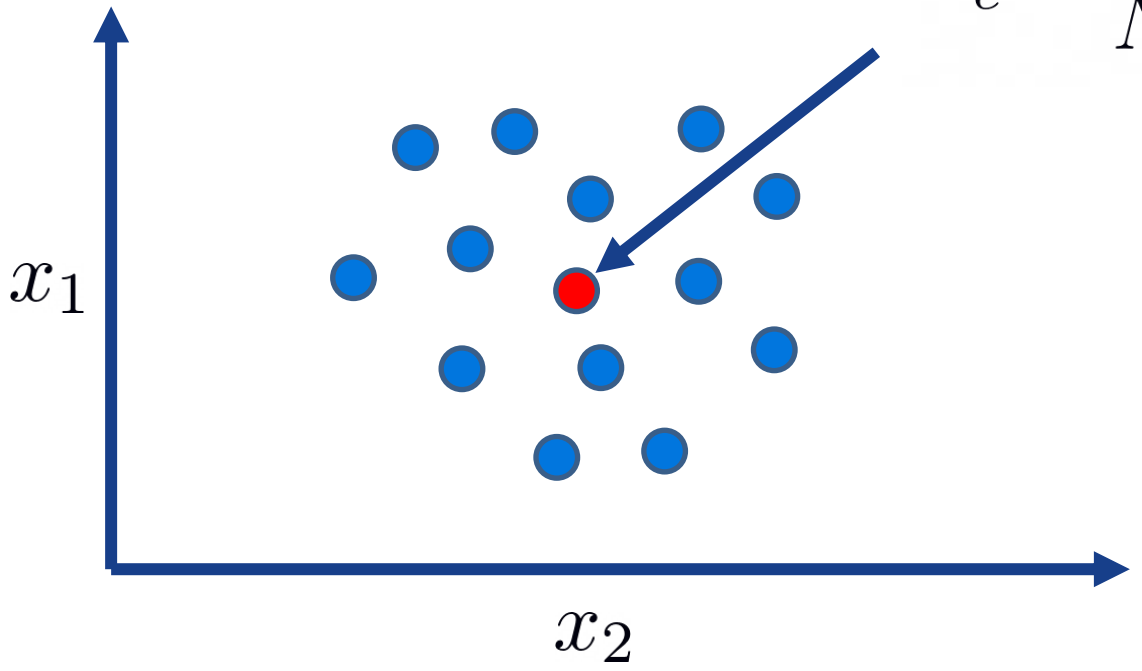
Profesor: **Francisco Pérez Galarce**

K-MEANS

1. Se debe aplicar en dominios numéricos.
2. Genera como resultado conjuntos disjuntos de objetos.
3. El número de segmentos debe ser entregado como un *input*.
4. Algoritmo simple que ha sido utilizado por décadas.

K-MEANS : Centroides

$$\mathbf{m}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} \mathbf{x}_i$$



K-MEANS

Input

Base de datos, número de segmentos/clases (K)

Output

Objetos pertenecientes a cada segmento.

Paso 1

- Elegir aleatoriamente K centroides.
- Ejecutar paso 1 y paso 2 mientras no se cumpla el criterio de detención.

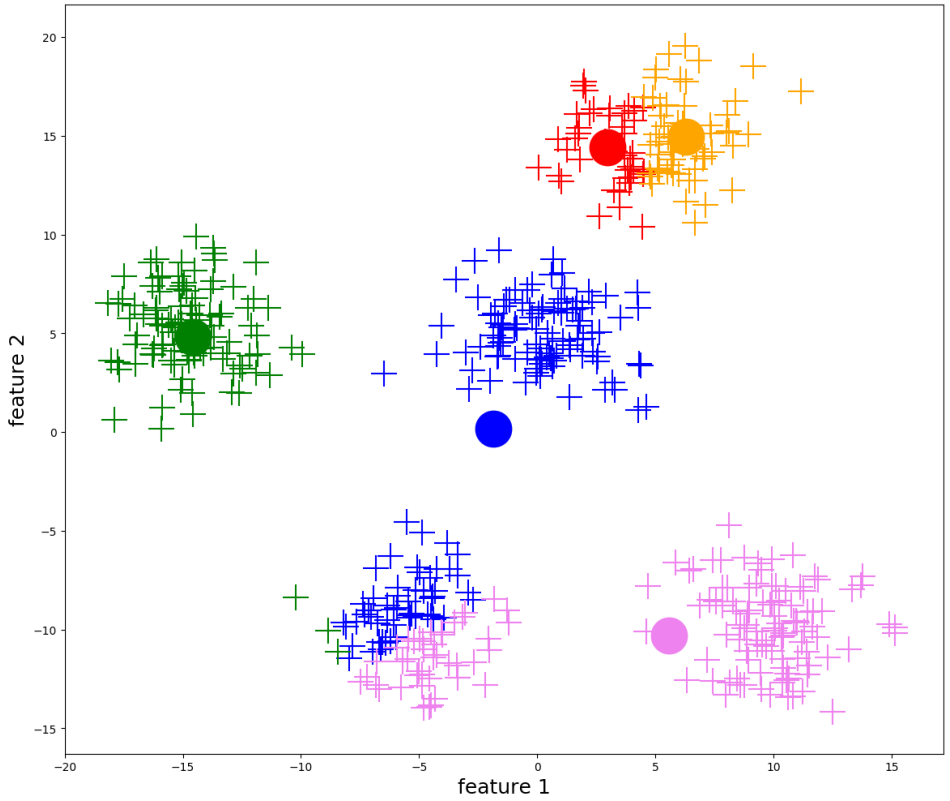
Paso 2

- Asignar cada objeto de la base de datos a su centroide más cercano.

Paso 3

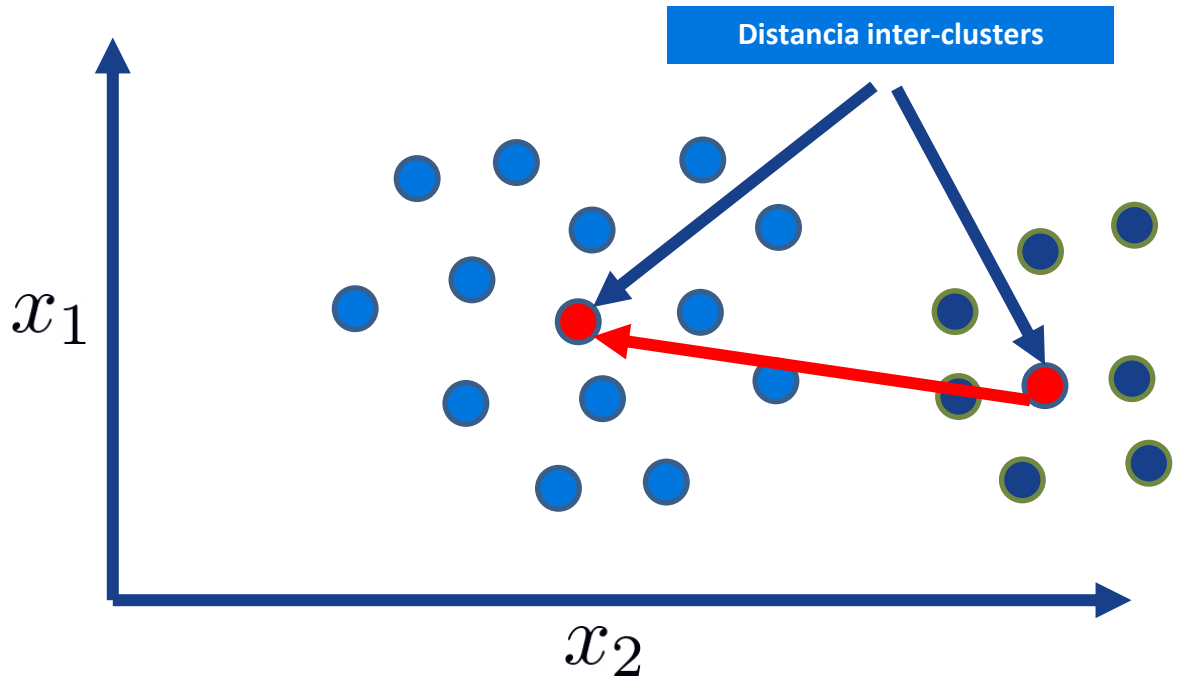
- Con las nuevas asignaciones recalcular centroides.

K-MEANS: ejemplo

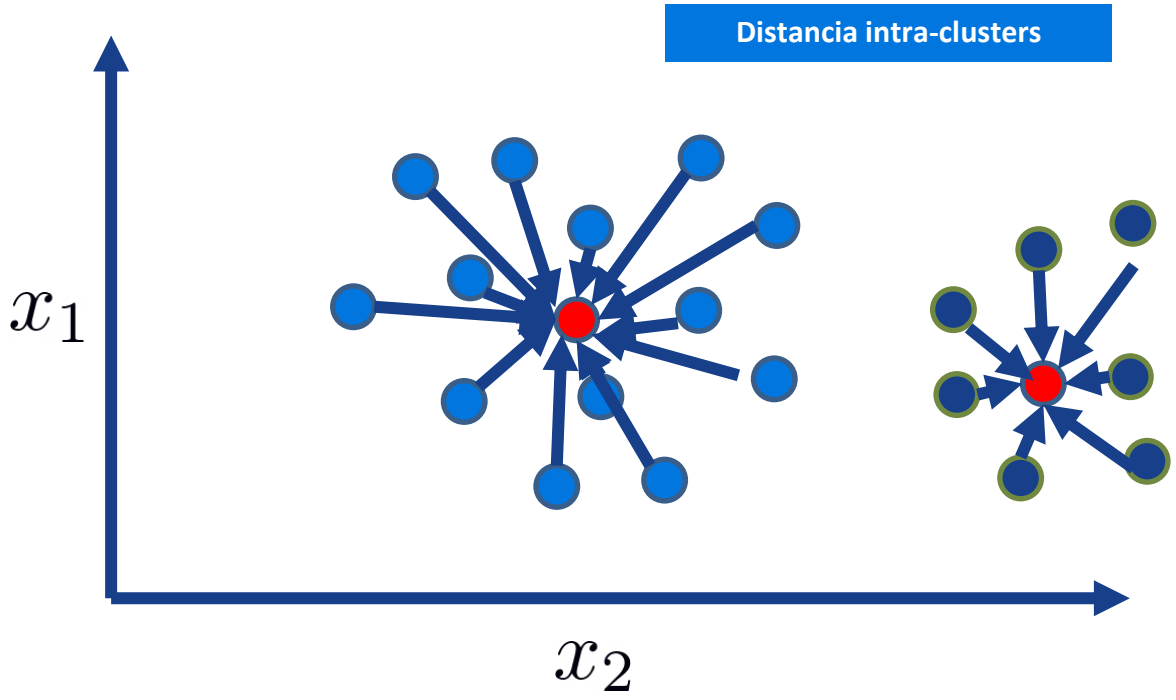


https://ai.plainenglish.io/understanding-k-means-clustering-hands-on-visual-approach-c2dc46f0ed18

K-MEANS: Análisis de clusters



K-MEANS: Análisis de clusters



EVALUACIÓN DE CLUSTERS

Profesor: **Francisco Pérez Galarce**

Davies-Bouldin index

$$\text{DB index} = \frac{1}{n}\sum_{i=1}^{n}\max_{j\neq i}\frac{(\sigma_i + \sigma_j)}{d(c_i, c_j)}$$

$$n = \text{Número de } \textit{clusters}$$
$$\sigma_i = \text{Distancia intra-}\textit{cluster}\text{ del } \textit{cluster } i$$
$$d(c_i, c_j) = \text{Distancia entre los centroides de los } \textit{clusters } i \text{ y } j$$

Silhouette score *(coherencia)*

Distancia promedio más pequeña del **objeto *i*** al resto de los puntos en **otros *clusters***.

Distancia promedio del **objeto *i*** al resto de los **puntos del *cluster***.

Número que indica **coherencia del objeto *i*** dentro del ***cluster*** que fue **asignado.**

$$S(i) = \frac{\overbrace{b(i)} - \overbrace{a(i)}}{\max\{a(i), b(i)\}}$$

$$-1 \leq S(i) \leq 1$$

Profesor: **Francisco Pérez Galarce**

El valor del Silhouette Score puede variar de -1 a 1:

- Un valor de -1 indica que los puntos están asignados incorrectamente a los clusters y que están más cerca de los puntos de otros clusters.

- Un valor de 0 indica que los puntos están cerca del límite entre dos clusters. Un valor de 1 indica que los puntos están bien agrupados y lejos de los puntos de otros clusters.

El mejor valor del Silhouette Score cercano a 1 ☺

ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN
PROFESIONAL

www.educacionprofesional.ing.uc.cl