



Definición

Aprendizaje Automático

Cómo crear programas de computadora que mejoren con la experiencia?
(Mitchell 1997)

Un **algoritmo de Machine Learning (ML)** es un algoritmo **capaz de aprender a partir de datos**

“Un programa se dice que **aprende** de una **experiencia E** con respecto a alguna clase de **tareas T** y una **medida de performance P**, si su **performance para las tareas en T**, medida por la **métrica P**, **mejora con la experiencia E**”.

Tipos de Aprendizaje

Aprendizaje Automático

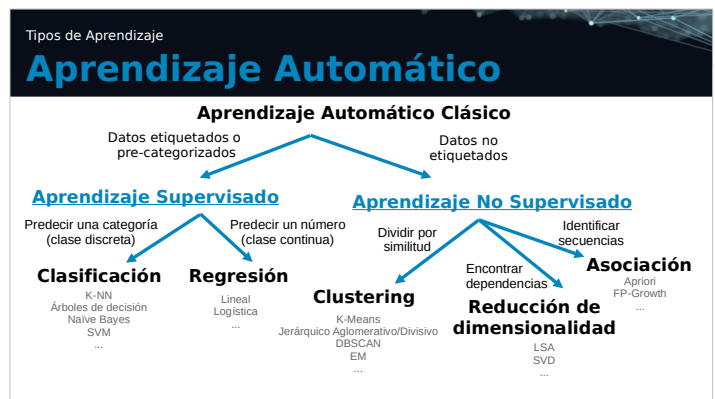
Dos formas principales de adquirir experiencia:

- A partir de ejemplos dados, es decir un conjunto de ejemplos clasificados o etiquetados

Aprendizaje Supervisado

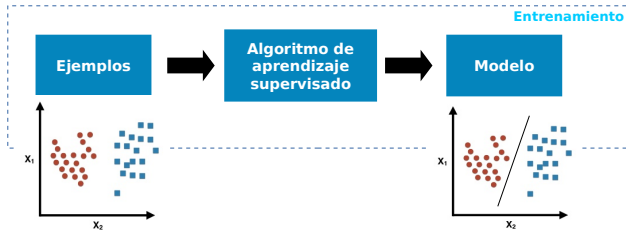
- Mediante exploración autónomas de los datos disponibles

Aprendizaje No Supervisado



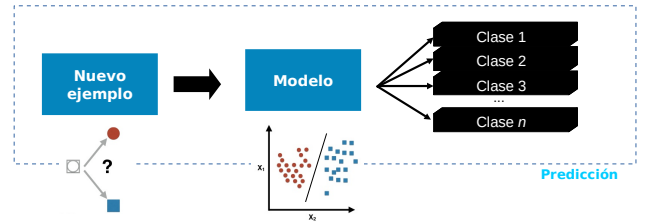
Aprendizaje Supervisado

Un paradigma de aprendizaje inductivo intenta aprender un **concepto**, **modelo o hipótesis** a través de **ejemplos** o **instancias** de tal concepto



Aprendizaje Supervisado

Para **predicción** se utilizan los conceptos aprendidos para inferir la clase de **nuevos** ejemplos



Clasificación

Funciones \mathcal{F}
 $f : \mathcal{X} \rightarrow \mathcal{Y}$

Datos de entrenamiento
 $\{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}\}$

Entrenamiento

Encontrar $\hat{f} \in \mathcal{F}$
 tal que $y_i \approx \hat{f}(x_i)$

Predicción

$x \Rightarrow y = \hat{f}(x)$

Clasificación

Problema: Es *spam* o no?

From: "" <akworld@hotmail.com>
 Subject: real estate is the only way...

Anyone can buy real estate with no money down
 Stop paying rent TODAY !
 There is no need to spend hundreds or even thousands for similar courses
 I am 22 years old and I have already purchased 6 properties using the methods
 outlined in this truly INCREDIBLE ebook.
 Change your life NOW !

Ejemplo: Clasificación de Spam

Clasificación

Cada mensaje se representa por un conjunto de características:

Errores de tipo	Remitente conocido	money	now	Mr.	...	Spam?
S	N	S	N	S		S
S	S	N	N	N		N
N	N	N	S	S		S
N	S	S	N	N		N
N	S	N	N	S		N
S	N	S	N	N		S
N	N	N	N	S		N

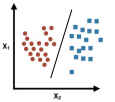
Ejemplo: Clasificación de Spam

Clasificación

Cada mensaje se representa por un conjunto de características:

Errores de tipo	Remitente conocido	money	now	Mr.	...	Spam?
S	N	S	N	S		S
S	S	N	N	N		N
N	N	N	S	S		S
N	S	S	N	N		N
N	S	N	N	S		N
S	N	S	N	N		S
N	N	N	N	S		N

Es **spam** si *remitente conocido=N* y (*money* o *now*)?



Ejemplo: Clasificación de Spam

Clasificación

Es *spam* o no?

From: "" <takworld@hotmail.com>
Subject: real estate is the only way...

Anyone can buy real estate with no money down
Stop paying rent TODAY !
There is no need to spend hundreds or even thousands for similar courses
I am 22 years old and I have already purchased 6 properties using the methods
outlined in this truly INCREDIBLE ebook.
Change your life NOW !

Es **spam** si *remitente conocido=N* y (*money* o *now*)?

Terminología

Clasificación

Objetivo: aprender un modelo a partir de una serie de ejemplos

Ejemplos o **instancias**, todos forman el **conjunto de datos** o **dataset**

Hipótesis o **modelo**: se infiere a partir de los ejemplos, se puede expresar de distintas formas (reglas, árbol de decisión, redes neurales, etc.)

Atributos o **características**: cada una de las propiedades que se observan de un ejemplo (valores reales o discretos)

Clases o **etiquetas**: es el atributo que se quiere predecir a partir de los demás

Clasificación

Errores de tipo	Remitente conocido	money	now	Mr.	...	Spam?
S	N	S	N	S		S
S	S	N	N	N		N
N	N	N	S	S		S
N	S	S	N	N		N
N	S	N	N	S		N
S	N	S	N	N		S
N	N	N	N	S		N

Dataset

Clasificación

Errores de tipo	Remitente conocido	money	now	Mr.	...	Spam?
S	N	S	N	S		S
S	S	N	N	N		N
N	N	N	S	S		S
N	S	S	N	N		N
N	S	N	N	S		N
S	N	S	N	N		S
N	N	N	N	S		N

Ejemplo o instancia

Clasificación

Errores de tipo	Remitente conocido	money	now	Mr.	...	Spam?
S	N	S	N	S		S
S	S	N	N	N		N
N	N	N	S	S		S
N	S	S	N	N		N
N	S	N	N	S		N
S	N	S	N	N		S
N	N	N	N	S		N

Atributo o característica

Clasificación

Errores de tipo	Remitente conocido	money	now	Mr.	...	Spam?
S	N	S	N	S		S
S	S	N	N	N		N
N	N	N	S	S		S
N	S	S	N	N		N
N	S	N	N	S		N
S	N	S	N	N		S
N	N	N	N	S		N

Clase, Categoría o etiqueta

Ejemplo: Clasificación de Spam

Clasificación

Objetivo: predecir si un email entrante es spam o no

Entrenamiento: colección de emails etiquetados como spam o no
→ alguien tiene que haberlos etiquetados manualmente

Atributos: las características usadas para tomar la decisión
→ palabras: *money*, *now*, etc.
→ patrones de texto: uso de mayúsculas
→ otros elementos: si el remitente está en los contactos, etc.

Clase: binaria *spam/no-spam*

Modelo o hipótesis: Es *spam* si remitente conocido=*N* y (*money* o *now*)

Ejemplo: Identificación de Flores

Clasificación

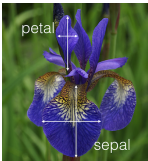
Problema: Podemos identificar el tipo de flor Iris?



Ejemplo: Identificación de Flores

Clasificación

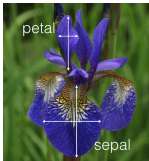
sepal length	sepal width	petal length	petal width	Tipo de Iris
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.9	3.1	4.9	1.3	versicolor
5.5	2.3	4.0	1.5	versicolor
6.3	3.3	6.0	2.5	virginica
5.8	2.7	5.1	1.9	virginica
7.1	3.0	5.9	2.1	virginica
6.3	2.9	5.6	1.8	virginica



Ejemplo: Identificación de Flores

Clasificación

sepal length	sepal width	petal length	petal width	Tipo de Iris
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.9	3.1	4.9	1.3	versicolor
5.5	2.3	4.0	1.5	versicolor
6.3	3.3	6.0	2.5	virginica
5.8	2.7	5.1	1.9	virginica
7.1	3.0	5.9	2.1	virginica
6.3	2.9	5.6	1.8	virginica
4.6	3.6	1.0	0.2	?



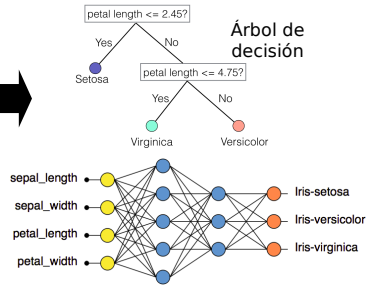
Ejemplo: Identificación de Flores

Clasificación

sepal_length	sepal_width	petal_length	petal_width	Tipo de Iris
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.9	3.1	4.9	1.3	versicolor
5.5	2.3	4.0	1.5	versicolor
6.3	3.3	6.0	2.5	virginica
5.8	2.7	5.1	1.9	virginica
7.1	3.0	5.9	2.1	virginica
6.3	2.9	5.6	1.8	virginica

Entrenamiento

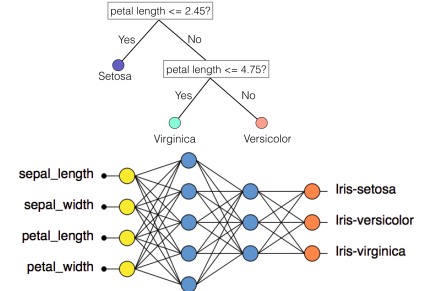
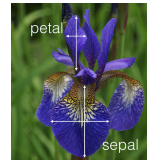
Red neuronal



Ejemplo: Identificación de Flores

Clasificación

Predicción



Ejemplo: Identificación de Flores

Clasificación

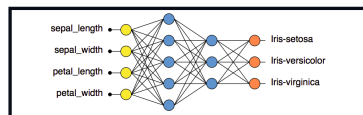
Objetivo: identificar flores de Iris

Entrenamiento: observaciones de flores previamente clasificadas

Atributos: medidas de la flor observada
→ alto y ancho de pétalo y sépalo

Clase: multi-clase *setosa/versicolor/virginica*

Modelo o hipótesis:



Ejemplo: Identificación de Flores

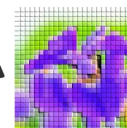
Clasificación

Atributos: se puede usar la imagen?

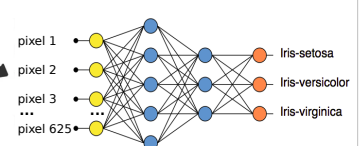
Entrenamiento: conjunto de imágenes previamente clasificadas



Iris flower sample



25 x 25 pixel image
(625 pixels)



Ejemplo: Juego de Tenis

Clasificación

Problema: Es un día para jugar al tenis?

Día	Panorama	Temperatura	Humedad	Viento	Juega tenis?
1	Soleado	Caluroso	Alta	Leve	NO
2	Soleado	Caluroso	Alta	Fuerte	NO
3	Nublado	Caluroso	Alta	Leve	SI
4	Lluvioso	Templado	Alta	Leve	SI
5	Lluvioso	Frio	Normal	Leve	SI
6	Lluvioso	Frio	Normal	Fuerte	NO
7	Nublado	Frio	Normal	Fuerte	SI
8	Soleado	Templado	Alta	Leve	NO
9	Soleado	Frio	Normal	Leve	SI
10	Lluvioso	Templado	Normal	Leve	SI
11	Soleado	Templado	Normal	Fuerte	SI
12	Nublado	Templado	Alta	Fuerte	SI
13	Nublado	Caluroso	Normal	Leve	SI
14	Lluvioso	Templado	Alta	Fuerte	NO

Ejemplo: Juego de Tenis

Clasificación

Objetivo: predecir si una persona jugará al tenis o no

Entrenamiento: observaciones de varios días

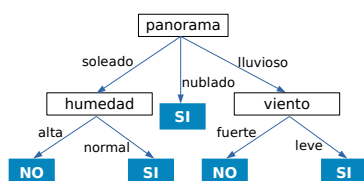
Atributos: condiciones climáticas
→ panorama, temperatura, humedad, viento

Clase: binaria *si/no*

Ejemplo: Juego de Tenis

Clasificación

Modelo o hipótesis:



Ejemplo: Minería de Opinión

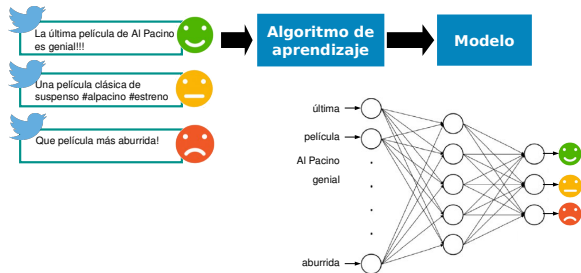
Clasificación

Problema: Una opinión es positiva, negativa o neutral?

- Cuál es la reacción ante el estreno de una película?
- Qué piensa la gente de un producto nuevo?
- Qué piensa la gente de un candidato o situación?

Ejemplo: Minería de Opinión

Clasificación



Ejemplo: Minería de Opinión

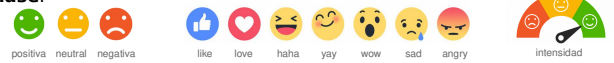
Clasificación

Objetivo: descubrir las opiniones, emociones y sentimientos hacia un producto o un servicio → **Minería de opinión**

Entrenamiento: colección de tweets/reviews con polaridad conocida

Atributos: características del tweet/review
 → palabras en el texto
 → elementos de la red social (por ej. *hashtags*)
 → características del autor
 → características del contexto

Clase:



Ejemplo: Minería de Opinión

Clasificación

Minería de Opinión: descubrir las opiniones, emociones y sentimientos hacia un producto o un servicio



Ejemplo: Emisión de CO2

Regresión

Problema: Cuál será la emisión de dióxido de carbono de un auto que se está por fabricar?

Tamaño del motor	Cilindros	Consumo de gasolina	Emisión CO2
2.0	4	8.5	196
2.4	4	9.6	221
1.5	4	5.9	136
3.5	6	11.1	255
3.5	6	10.6	244
3.5	6	10.0	230
3.5	6	10.1	232
3.7	6	11.1	255
3.7	6	11.6	267

Ejemplo: Emisión de CO2

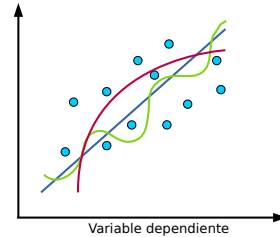
Regresión

Variables independientes			Variable dependiente
Tamaño del motor	Cilindros	Consumo de gasolina	
2.0	4	8.5	196
2.4	4	9.6	221
1.5	4	5.9	136
3.5	6	11.1	255
3.5	6	10.6	244
3.5	6	10.0	230
3.5	6	10.1	232
3.7	6	11.1	255
3.7	6	11.6	267

Ejemplo: Emisión de CO2

Regresión

Problema:Cuál será la emisión de dióxido de carbono de un auto que se está por fabricar?



Resumen

Clasificación

Clasificación y regresión son tareas **predictivas**

- Los ejemplos pertenecen a un conjunto de **clases** finitas y conocidas de antemano
- Las **características** que describen los ejemplos deben seleccionarse cuidadosamente
- Normalmente los datos requieren pre-procesamiento antes del entrenamiento
- Los algoritmos de ML aplican análisis estadístico sobre los datos a fin de generar un **modelo** que los describa
- Cuando mayor cantidad de ejemplos se utilicen para entrenamiento, mejor será el modelo resultante (mejora con la experiencia)
- El modelo se utiliza finalmente para la toma de decisión
- La confiabilidad del modelo depende de varios factores y debe ser evaluada

Definición

Clustering

El agrupamiento o clustering es un método **no supervisado** para organizar un conjunto de ejemplos **no etiquetados** en una serie de grupos o **clusters**



Definición

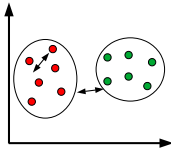
Clustering

Informalmente, se podría decir que el clustering busca encontrar un agrupamiento natural de los datos

Un *cluster* es un subconjunto de los ejemplos que son **similares** entre sí, y **disímiles** a los ejemplos en otros *clusters*

Un buen agrupamiento debería general *clusters* con:

- alta similitud intra-cluster
- baja similitud inter-cluster



Definición

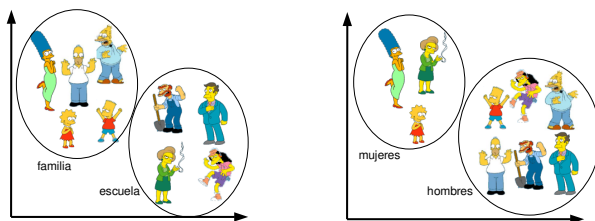
Clustering

Cuál es la forma natural de agrupar estos datos?



Definición

Clustering



Es una tarea descriptiva, el agrupamientos depende de un criterio de relación expresado a través de una función de similitud/distancia

Definición

Clustering

- Medida de **similitud** o proximidad



Euclidiana

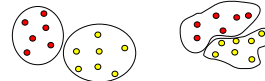


Manhattan



Coseno

- Un criterio o función de **evaluación** de un agrupamiento



- Un algoritmo para calcular el agrupamiento (por ej. optimizando la medida de evaluación)

Ejemplo: Segmentación de Clientes

Clustering

Datos: conjunto de usuarios de cable con su historial

	We	BBC	oh!	Animal	Travel	TLC	Health	Fox News	CNN
Usuario 1	X		X				X		X
Usuario 2				X	X	X	X	X	X
Usuario 3			X					X	
Usuario 4	X	X							
Usuario 5	X		X		X				X
Usuario 6		X		X	X		X	X	

Ejemplo: Segmentación de Clientes

Clustering

	We	BBC	oh!	Animal	Travel	TLC	Health	Fox News	CNN
Usuario 1	X		X				X		X
Usuario 2				X	X	X	X	X	X
Usuario 3			X					X	
Usuario 4	X	X	X						
Usuario 5	X		X		X				X
Usuario 6		X		X	X		X	X	

Los usuarios 1, 4 y 5 se agrupan en un mismo cluster y sus preferencias típicas serían:

We oh! Muy alta CNN Alta BBC Travel Health Baja

Ejemplo: Segmentación de Clientes

Clustering

	We	BBC	oh!	Animal	Travel	TLC	Health	Fox News	CNN
Usuario 1	X		X				X		X
Usuario 2				X	X	X	X	X	X
Usuario 3			X					X	
Usuario 4	X	X	X						
Usuario 5	X		X		X				X
Usuario 6		X		X	X		X		

Los usuarios 2 y 6 se agrupan en un mismo cluster y sus preferencias típicas serían:

Animal Travel Health Muy alta BBC TLC CNN Baja

Ejemplo: Segmentación de Clientes

Clustering



Ejemplo: Segmentación de Clientes

Clustering



Paquete I



Paquete II



Ejemplo: Segmentación de Clientes

Clustering

Datos: conjunto de usuarios de cable con su historial

- Identificar segmentos de consumidores a fin de crear ofertas diferenciadas en base a preferencias comunes
- Por ejemplo: armando de paquetes de cable, recomendación de canales, direccionamiento de publicidad

Ejemplo: Agrupamiento de Documentos

Clustering

Datos: conjunto de páginas Web

- resultado de un consulta
- almacenadas en el servidor de una compañía
- colección de documentos médicos, legales, etc.

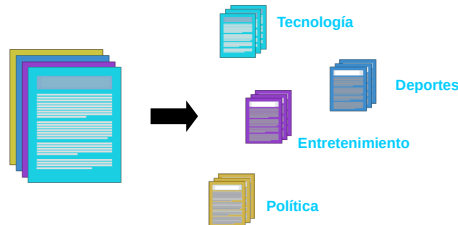
Ejemplo: Agrupamiento de Documentos

Clustering

	nova	galaxy	heat	hollywood	film	role	diet	fur
doc 1	1.0	0.5	0.3					
doc 2	0.5	1.0						
doc 3	1.0	0.8	0.7					
doc 4		0.9	1.0	0.5				
doc 5				1.0	1.0			
doc 6					0.9		1.0	
doc 7	0.5		0.7				0.9	
doc 8		0.6		1.0	0.3	0.2		0.8
doc 9			0.7	0.5	0.1	0.3		

Ejemplo: Agrupamiento de Documentos

Clustering



Ejemplo: Agrupamiento de Documentos

Clustering

yippy

Web News Images Video

clustering

Search

Results: 1,398 of about 23,398,376 Details

Sources Sites Time Topics

Top 323 Clusters

Search (2)

Machine Learning (2)

Cluster Analysis (2)

Definition (2)

Concept, Document and field level (2)

Free (2)

Data clustering (2)

Science, Methods (2)

Hierarchical clustering (2)

Clustering Methods (2)

University (2)

Code (2)

Clustering software (2)

Apache, Software Foundation Blog (2)

Image (2)

Market (2)

Protein (2)

Servers, Computing (2)

Neural networks (2)

Risk, Solutions (2)

Cluster analysis - Wikipedia

Date: 2020-04-13T13:05:00

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).

https://en.wikipedia.org/wiki/Cluster_analysis - Yippy Index V

2. Clustering: Introduction & different methods of clustering

Date: 2020-04-13T14:29:00

5. Hierarchical Clustering. Hierarchical clustering, so the name suggests is an algorithm that builds hierarchy of clusters. This algorithm starts with all the data points assigned to a cluster of their own.

https://www.analyticsvidhya.com/blog/2016/05/5-different-methods-of-clustering/ - Yippy Index V

3. Clustering: Definition of Clustering at Dictionary.com

Date: 2020-04-13T13:05:00

Clustering definition, a number of things of the same kind, growing or held together; a bunch; a cluster of grapes. See more.

https://www.dictionary.com/browse/clustering - Yippy Index V

4. The 5 Clustering Algorithms Data Scientists Need to Know

Date: 2020-04-13T13:05:00

Clustering is a method of unsupervised learning and is a common technique for statistical data analysis used in many fields. In Data Science, we can use clustering analysis to gain some valuable insights from our data by seeing what groups the data points fall into when we apply a clustering algorithm.

https://towardsdatascience.com/...-algorithms-weed-to-know-216d113d4861 - Yippy Index V

5. What is Clustering? - Definition from Techopedia

Date: 2020-04-13T13:05:00

Clustering is a method of unsupervised learning and is a common technique for statistical data analysis used in many fields. In Data Science, we can use clustering analysis to gain some valuable insights from our data by seeing what groups the data points fall into when we apply a clustering algorithm.

https://towardsdatascience.com/...-algorithms-weed-to-know-216d113d4861 - Yippy Index V

6. What is Clustering? - Definition from Techopedia

Date: 2020-04-13T13:05:00

Clustering is a method of unsupervised learning and is a common technique for statistical data analysis used in many fields. In Data Science, we can use clustering analysis to gain some valuable insights from our data by seeing what groups the data points fall into when we apply a clustering algorithm.

https://towardsdatascience.com/...-algorithms-weed-to-know-216d113d4861 - Yippy Index V

Ejemplo: Agrupamiento de Documentos

Clustering

Clustering de los documentos recuperados para una consulta

- se presentan los resultados de una búsqueda en forma más organizada que facilita la navegación (por ej. *Yippy*)

Clustering de documentos en una colección

- hipótesis de clustering, documentos similares tienden a ser relevantes a la misma consulta

Clustering para generación automática de taxonomías

- facilita la exploración de documentos (por ej. *Yahoo!*)

Definición

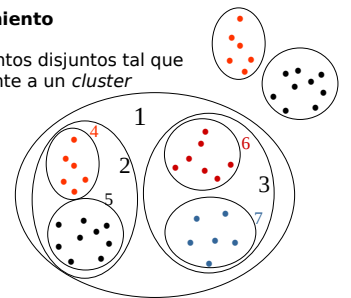
Clustering

Algoritmos basados en **particionamiento**

- dividen los ejemplos en subconjuntos disjuntos tal que cada objeto pertenece exactamente a un *cluster*

Algoritmos **jerárquicos**

- generan un conjunto de *clusters* anidados en un árbol jerárquico



Clustering

El clustering es una tarea **descriptiva**

- Los ejemplos no tienen una clase definida
- Un algoritmo de clustering genera un posible agrupamiento de los ejemplos, generalmente optimizando un criterio de evaluación
- En algunas aplicaciones el clustering no es la tarea primaria, sino que asiste en el desarrollo de otra tarea
- El clustering es difícil de evaluar, es dependiente de la aplicación y, se podría decir que es subjetivo

Próximas clases

Aprendizaje Supervisado

- Algoritmos de clasificación/regresión (k-NN, árboles, etc.)
- Evaluación de la clasificación

Aprendizaje No Supervisado

- Algoritmos de clustering (k-Means, aglomerativo, etc.)
- Evaluación del clustering