

Clase 1

Aspectos Generales

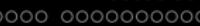
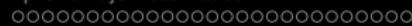
Marcelo Luis Errecalde^{1,2}

¹Universidad Nacional de San Luis, Argentina

²Universidad Nacional de la Patagonia Austral, Argentina
e-mails: merreca@unsl.edu.ar, merrecalde@gmail.com



Curso: Minería de Datos
Universidad Nacional de San Luis - Año 2018



Resumen

- 1 Aprendizaje Automático
- 2 Minería de Datos
- 3 El Proceso de KDD
- 4 Big Data (Analytics)
- 5 Razones para el uso de Python y Weka

Aprendizaje Automático (Machine Learning)

- Uno de los pilares fundamentales para la **Minería de Datos** es el concepto de **Aprendizaje Automático (Machine Learning)**, por lo que comenzaremos introduciendo sus principales ideas.

Aprendizaje Automático (Machine Learning)

- Uno de los pilares fundamentales para la **Minería de Datos** es el concepto de **Aprendizaje Automático (Machine Learning)**, por lo que comenzaremos introduciendo sus principales ideas.
- **Aprendizaje Automático**: área fundamental de la Inteligencia Artificial para obtener sistemas **flexibles** y **adaptativos** en problemas **complejos**.

Aprendizaje Automático (Machine Learning)

- Uno de los pilares fundamentales para la Minería de Datos es el concepto de Aprendizaje Automático (Machine Learning), por lo que comenzaremos introduciendo sus principales ideas.
- Aprendizaje Automático: área fundamental de la Inteligencia Artificial para obtener sistemas flexibles y adaptativos en problemas complejos.
- Entrenamiento en lugar de programación.

Aprendizaje Automático (Machine Learning)

- Uno de los pilares fundamentales para la Minería de Datos es el concepto de Aprendizaje Automático (Machine Learning), por lo que comenzaremos introduciendo sus principales ideas.
- Aprendizaje Automático: área fundamental de la Inteligencia Artificial para obtener sistemas flexibles y adaptativos en problemas complejos.
- Entrenamiento en lugar de programación.
- Conocimiento es descubierto en lugar de ser codificado.

Aprendizaje Automático (Machine Learning)

- Uno de los pilares fundamentales para la Minería de Datos es el concepto de Aprendizaje Automático (Machine Learning), por lo que comenzaremos introduciendo sus principales ideas.
- Aprendizaje Automático: área fundamental de la Inteligencia Artificial para obtener sistemas flexibles y adaptativos en problemas complejos.
- Entrenamiento en lugar de programación.
- Conocimiento es descubierto en lugar de ser codificado.
- Patrones y modelos “aproximados” en lugar de algoritmos exactos.



Aprendizaje automático (AA)

Aprendizaje automático (AA)

[Russell, 2009]

*“... cualquier sistema que se considere “inteligente” debería poseer la habilidad de **aprender**, es decir **mejorar** automáticamente con la **experiencia**. ”*

Aprendizaje automático (AA)

[Russell, 2009]

*“... cualquier sistema que se considere “inteligente” debería poseer la habilidad de **aprender**, es decir **mejorar** automáticamente con la **experiencia**. ”*

[Herbert Simon]

*“... cualquier cambio en un sistema que le permite **desempeñarse mejor** la próxima vez, sobre la misma tarea u **otra tomada de la misma población**”*

Aprendizaje automático (AA)

[Mitchell, 1997]

*“Un programa de computadora se dice que aprende desde la experiencia **E** con respecto a alguna clase de tareas **T** y medida de performance **P**, si mejora su performance con las tareas en **T**, con respecto a la medida **P**, basado en la experiencia **E**”*

Aprendizaje automático (AA)

[Mitchell, 1997]

*“Un programa de computadora se dice que aprende desde la **experiencia E** con respecto a alguna clase de **tareas T** y **medida de performance P**, si mejora su **perfomance** con las tareas en **T**, con respecto a la medida **P**, basado en la experiencia **E**”*

Ejemplo 1: atribución de autoría

- **Tarea T**: dado un documento arbitrario, identificar su autor.
 - **Experiencia E**: conjunto de documentos del grupo de escritores con que se trabajará.
 - **Medida de performance P**: porcentaje de aciertos de la autoría del documento (evaluada con otro conjunto **distinto** de documentos de los mismos autores).

Aprendizaje automático (AA)

[Mitchell, 1997]

*“Un programa de computadora se dice que aprende desde la **experiencia E** con respecto a alguna clase de **tareas T** y **medida de performance P**, si mejora su performance con las tareas en **T**, con respecto a la medida **P**, basado en la experiencia **E**”*

Ejemplo 2: jugar a las damas

- **Tarea T:** aprender a jugar a las damas.
 - **Experiencia E:** resultado (+1, 0 o -1) de partidas completas jugadas por el sistema contra sí mismo.
 - **Medida de performance P:** porcentaje de partidas ganadas al campeón mundial de ajedrez.

Aprendizaje automático

Algunos factores implícitos en las definiciones de AA:

- **cambios** en el comportamiento para lograr una mejor **performance** futura.

Aprendizaje automático

Algunos factores implícitos en las definiciones de AA:

- **cambios** en el comportamiento para lograr una mejor **performance** futura.
- existencia de algún tipo de **experiencia de entrenamiento**.

Aprendizaje automático

Algunos factores implícitos en las definiciones de AA:

- **cambios** en el comportamiento para lograr una mejor **performance** futura.
- existencia de algún tipo de **experiencia de entrenamiento**.
- capacidad de **generalizar** a objetos no observados previamente.

Aprendizaje automático

La componente más variable es el origen de la experiencia de entrenamiento:

- Interacción con el ambiente u otros agentes

Aprendizaje automático

La componente más variable es el origen de la experiencia de entrenamiento:

- Interacción con el **ambiente** u otros **agentes**
- Interacción **usuario-sistema** (agentes de interfaz)

Aprendizaje automático

La componente más variable es el origen de la experiencia de entrenamiento:

- Interacción con el **ambiente** u otros **agentes**
- Interacción **usuario-sistema** (agentes de interfaz)
- Aprendizaje por **observación** o asistido por otros agentes (**consejos**)

Aprendizaje automático

La componente más variable es el origen de la experiencia de entrenamiento:

- Interacción con el **ambiente** u otros **agentes**
- Interacción **usuario-sistema** (agentes de interfaz)
- Aprendizaje por **observación** o asistido por otros agentes (**consejos**)
- **Introspección** de los propios procesos internos

Aprendizaje automático

La componente más variable es el origen de la experiencia de entrenamiento:

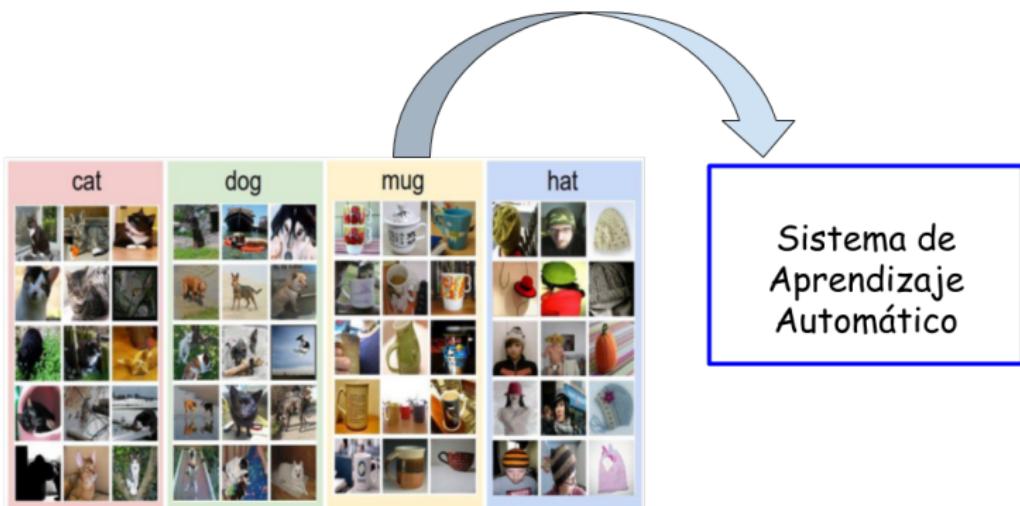
- Interacción con el **ambiente** u otros **agentes**
- Interacción **usuario-sistema** (agentes de interfaz)
- Aprendizaje por **observación** o asistido por otros agentes (**consejos**)
- **Introspección** de los propios procesos internos
- **Bases de datos**

Tipos de aprendizaje (de acuerdo a la retroalimentación)

- **Aprendizaje supervisado:** experiencia es un conjunto de ejemplos $\langle x, f(x) \rangle$, de la función f a ser aproximada.
- **Aprendizaje por refuerzos:** experiencia son secuencias de tri-uplas $\langle s, a, r \rangle$, donde a es la acción tomada por el agente en el estado s , y r es la evaluación numérica recibida desde el ambiente por la realización de esta acción.
- **Aprendizaje no supervisado:** **no existe** una retroalimentación explícita desde el ambiente.

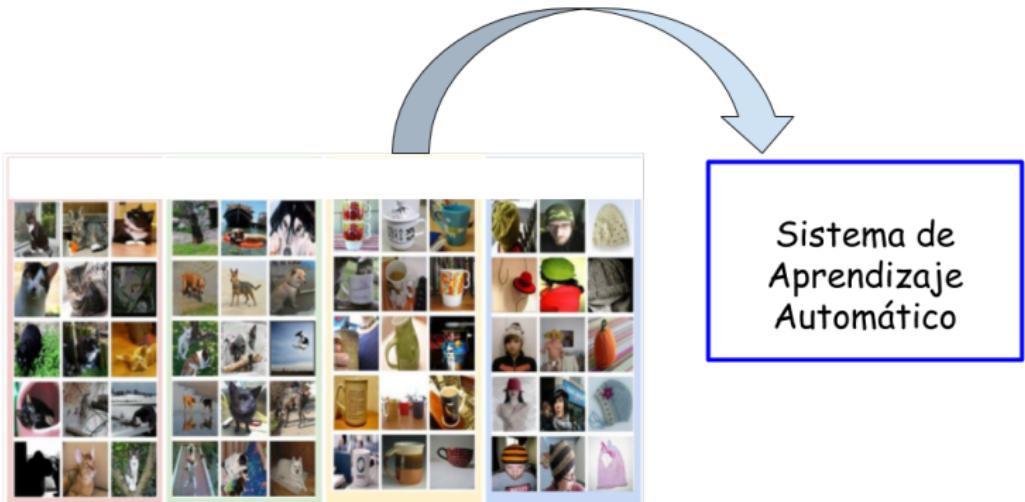
Retroalimentación en Aprendizaje Automático: supervisado

Retroalimentación = Ejemplos Etiquetados



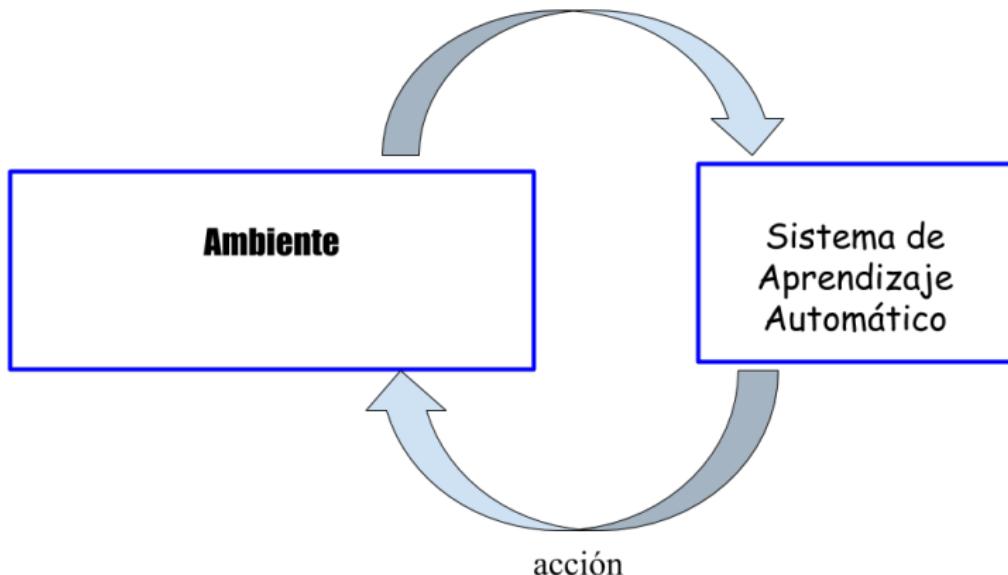
Retroalimentación en Aprendizaje Automático: no-supervisado

Retroalimentación = Ejemplos No Etiquetados

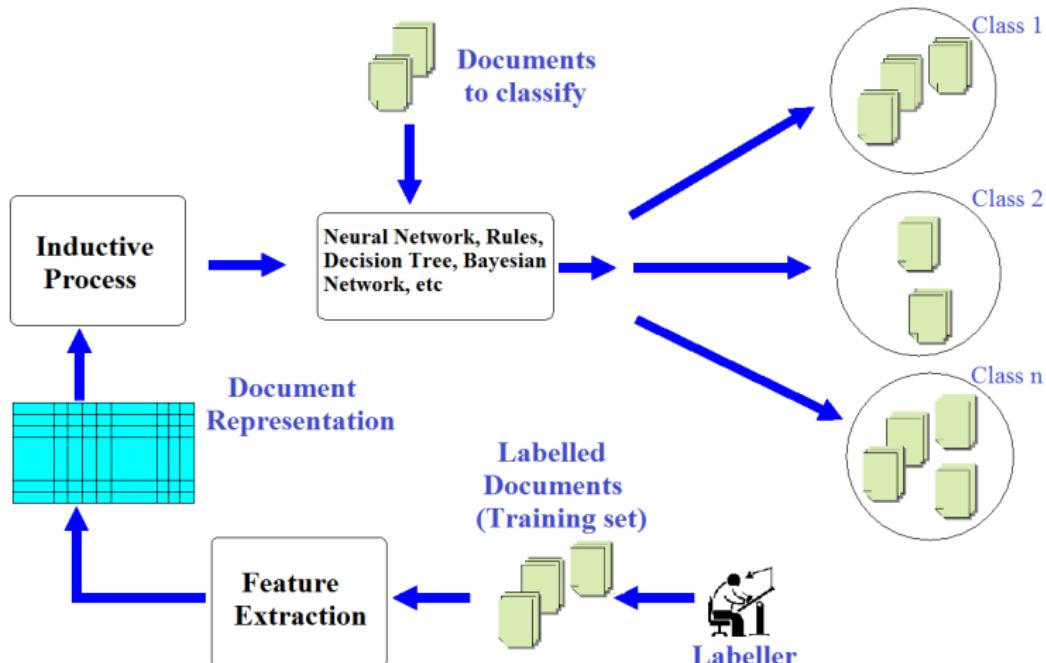


Retroalimentación en Aprendizaje Automático: refuerzo

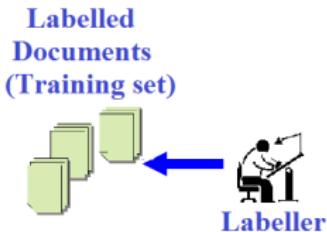
Retroalimentación = Recompensa/penalización



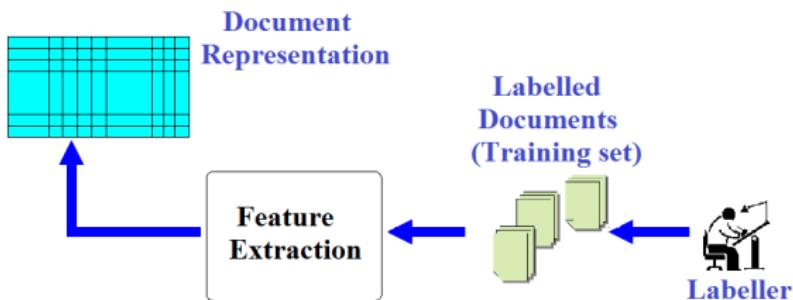
Ejemplo: Categorización Supervisada de Textos



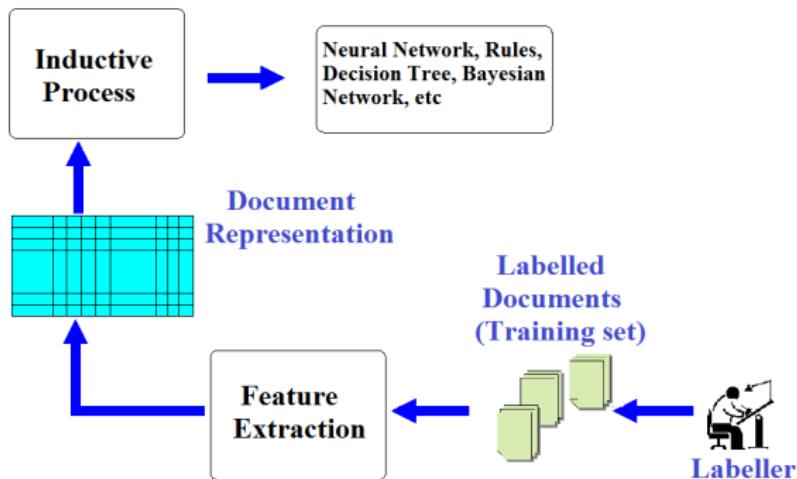
Ejemplo: Categorización Supervisada de Textos



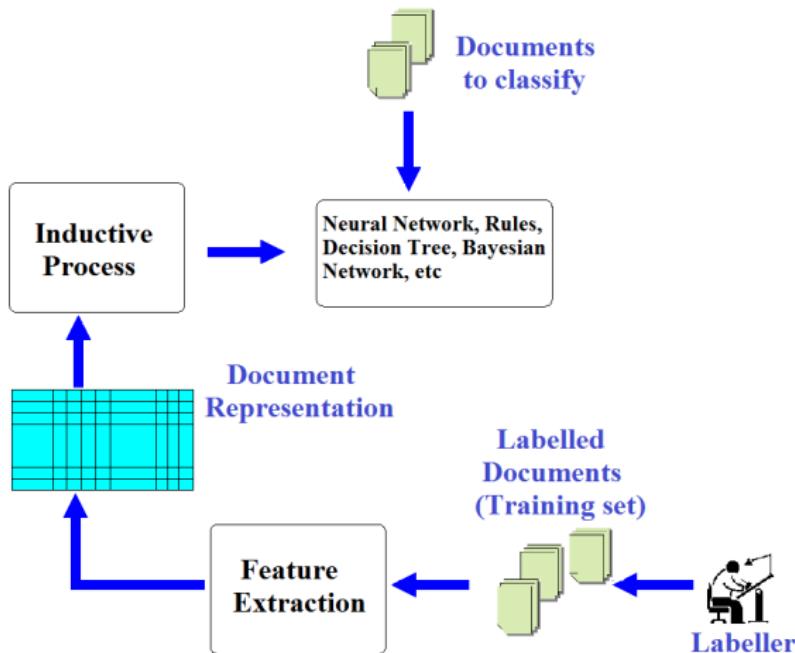
Ejemplo: Categorización Supervisada de Textos



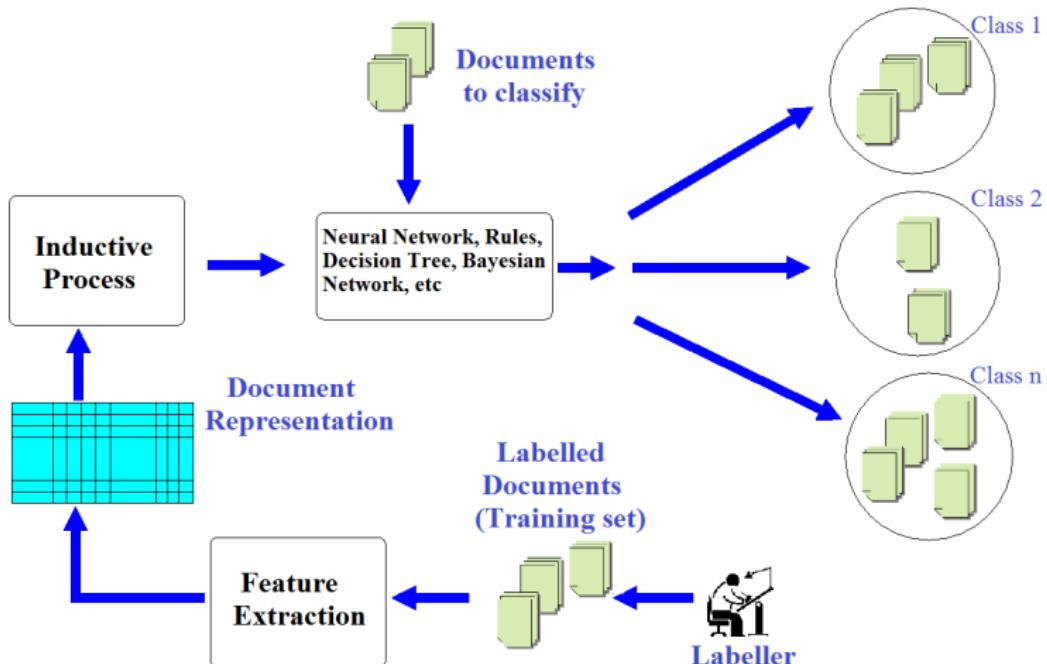
Ejemplo: Categorización Supervisada de Textos



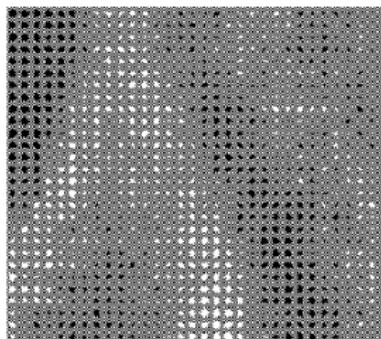
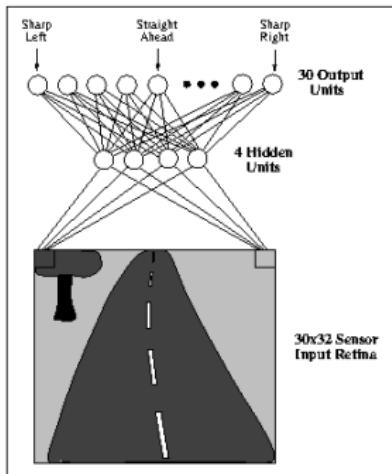
Ejemplo: Categorización Supervisada de Textos



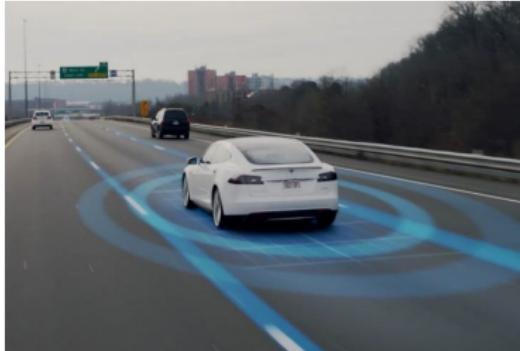
Ejemplo: Categorización Supervisada de Textos



Aprendizaje Automático: autos autónomos (de ALVINN ...)



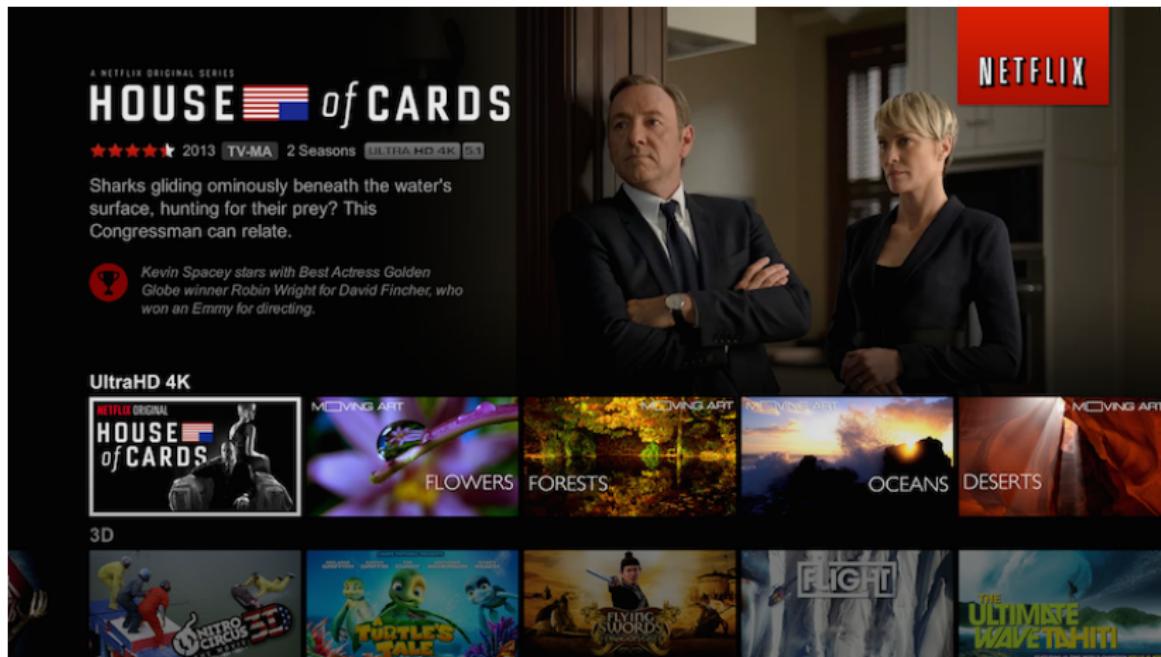
Aprendizaje Automático: autos autónomos (... al presente)



Aprendizaje Automático: vuelo invertido



Aprendizaje Automático: software que se adapta al usuario



Aprendizaje Automático: edificios que se adaptan al usuario



[home](#) / [about](#) / [research projects](#) / [tools](#) / [datasets](#) / [publications](#)

TOOLS

WSU CASAS Tools

- Real-time activity profiling
- CASAS-It
- AL activity learning (recognition, discovery, and prediction)
- AR activity recognition
- AR activity discovery
- Rule-based activity prediction
- AD pattern visualizer
- AV activity visualization
- Real-time annotation tools
- Data sampling tools (SMOTEBoost, RUSBoost, RACOG, wRACOG)
- ALZ sequential prediction
- Multiview transfer learning techniques

GIVE TO CASAS



CONNECT WITH US



LEADERSHIP

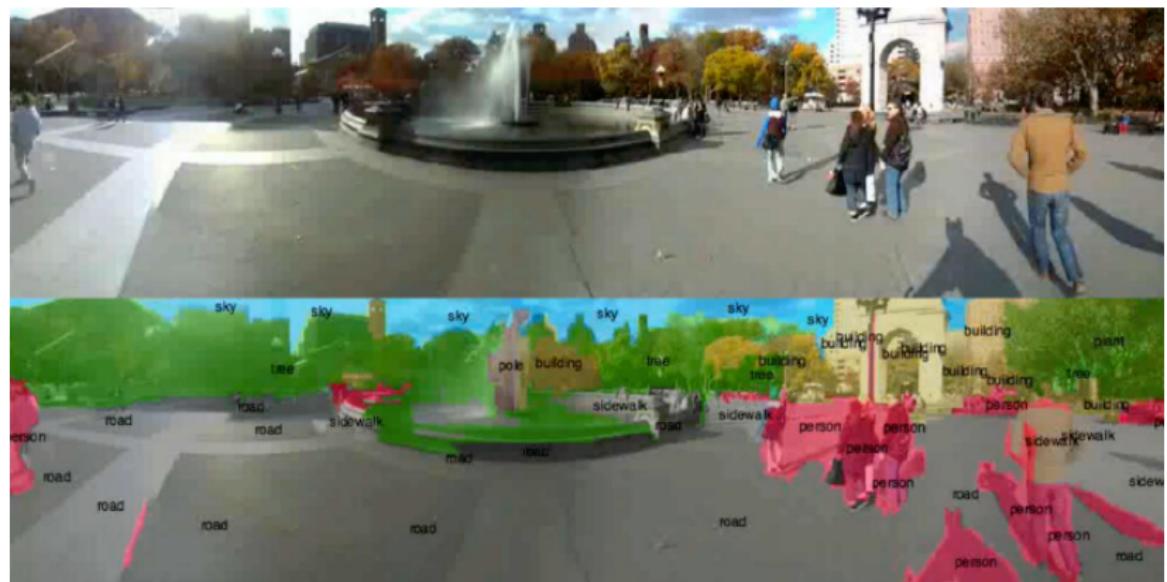
- Aaron Crandall
- Behrooz Shirdzi
- Diane Cook

Center for Advanced Studies in Adaptive

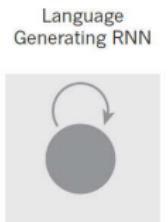
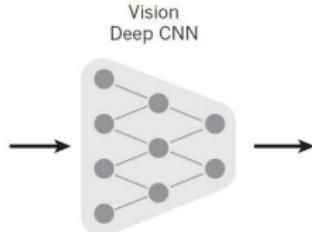
Systems (CASAS)
School of Electrical Engineering and
Computer Science
EME 121 Spokane Steel

<http://ailab.eecs.wsu.edu/casas/>

Aprendizaje Automático: rotulado de escenas



Aprendizaje Automático: descripción de escenas



A group of people
shopping at an outdoor
market.

There are many
vegetables at the
fruit stand.

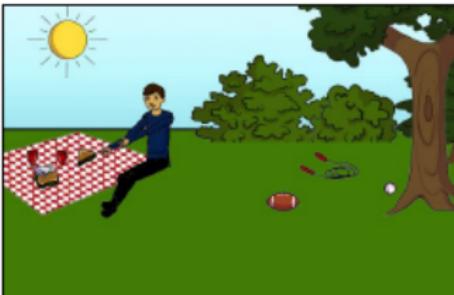
Aprendizaje Automático: responder preguntas sobre imágenes



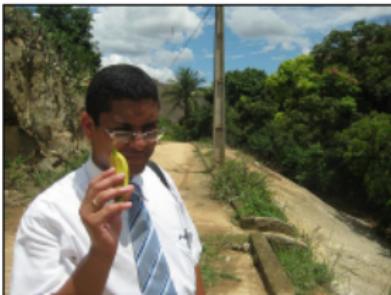
What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?



Does it appear to be rainy?
Does this person have 20/20 vision?

Aprendizaje Automático: transferencia de estilos

A



B



C



D



Aprendizaje Automático: reconocimiento de emociones

Face Emotion Recognition

Happy: 100%

Sad: 0%

Surprise: 0%

Fear: 0%

Disgust: 0%

Angry: 0%

Neutral: 0%

Happy Sad Surprise Fear Disgust Angry Neutral



Please read the following sentence once you click on OK

Cancel OK

Voice Emotion Recognition

```
1 1:i? 2:Neutral 8.923
*** Predictions on test data ***
inst# actual predicted error prediction ()
1 1:i? 2:Neutral 8.677

1 1:i? 2:Neutral 8.804
*** Predictions on test data ***
inst# actual predicted error prediction ()
1 1:i? 2:Neutral 8.577

1 1:i? 3:Happy 8.813
*** Predictions on test data ***
inst# actual predicted error prediction ()
1 1:i? 3:Happy 8.813
```

Please read the following sentence once you click on OK

Cancel OK

PowerPoint Slides

Sender Emotions: Happy

Actually, I have some really good news for you.

Aprendizaje Automático en este curso

- Nos centraremos en su utilización en la Ciencia de Datos y la Minería de Datos.

Aprendizaje Automático en este curso

- Nos centraremos en su utilización en la Ciencia de Datos y la Minería de Datos.
 - Usaremos como herramienta el lenguaje Python y la biblioteca Scikit-Learn.

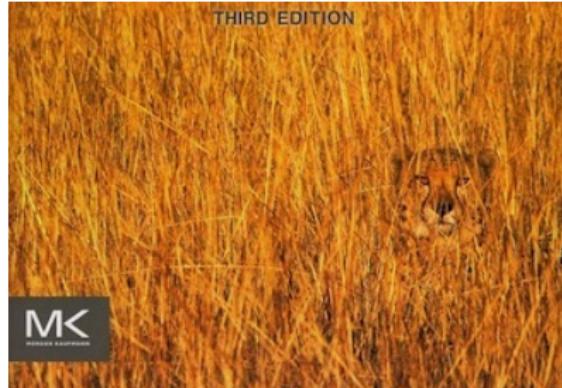
Aprendizaje Automático en este curso

- Nos centraremos en su utilización en la **Ciencia de Datos** y la **Minería de Datos**.
 - Usaremos como herramienta el lenguaje **Python** y la biblioteca **Scikit-Learn**.
 - También ejemplificaremos utilizando el sistema **Weka**.

Aprendizaje Automático en este curso

- Nos centraremos en su utilización en la **Ciencia de Datos** y la **Minería de Datos**.
 - Usaremos como herramienta el lenguaje **Python** y la biblioteca **Scikit-Learn**.
 - También ejemplificaremos utilizando el sistema **Weka**.

¿Qué es la Minería de Datos (MD)?



¿Qué es la Minería de Datos (MD)?

[Witten y Frank, 2011]

*“...es el proceso of descubrir **patrones** en los **datos**. El proceso debe ser automático o (más usualmente) semi-automático. Los **patrones** descubiertos deben ser **significativos**, en el sentido que conducen a alguna **ventaja** (usualmente económica). Los **datos** se presentan, invariablemente, en **grandes cantidades**”*

[Mitchell, 1999]

*“... uso de **datos históricos** para descubrir **regularidades generales** y mejorar las decisiones futuras”*

¿Qué es la Minería de Datos (MD)?

[H. Orallo, 2004]

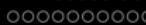
“... proceso que tiene como objetivo convertir *datos* en *conocimiento*”

[Fayyad, 1996]

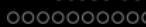
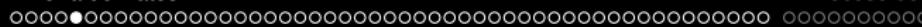
“... es un *paso* particular en el proceso de KDD que consiste en la aplicación de algoritmos específicos para extraer *patrones* (*o modelos*) desde los *datos*”

[Tan-Steinbach-Kumar, 2006]

“... es el proceso de *descubrir* información *útil*, de forma automática, en *grandes repositorios de datos*”

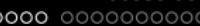
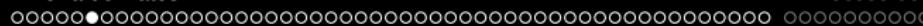


¿Por qué surge la Minería de Datos?

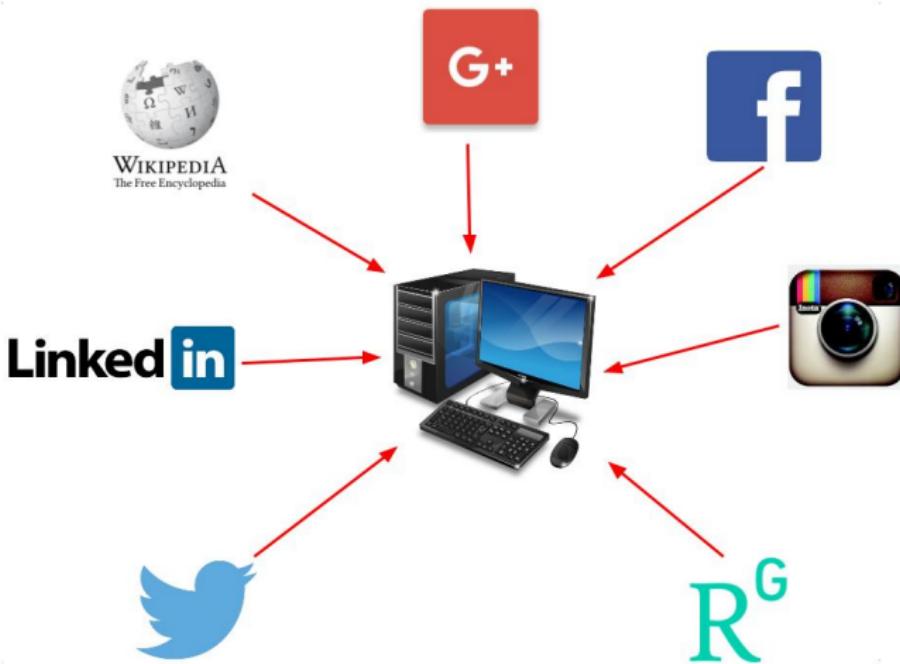


El acceso a la información hoy en día, ...

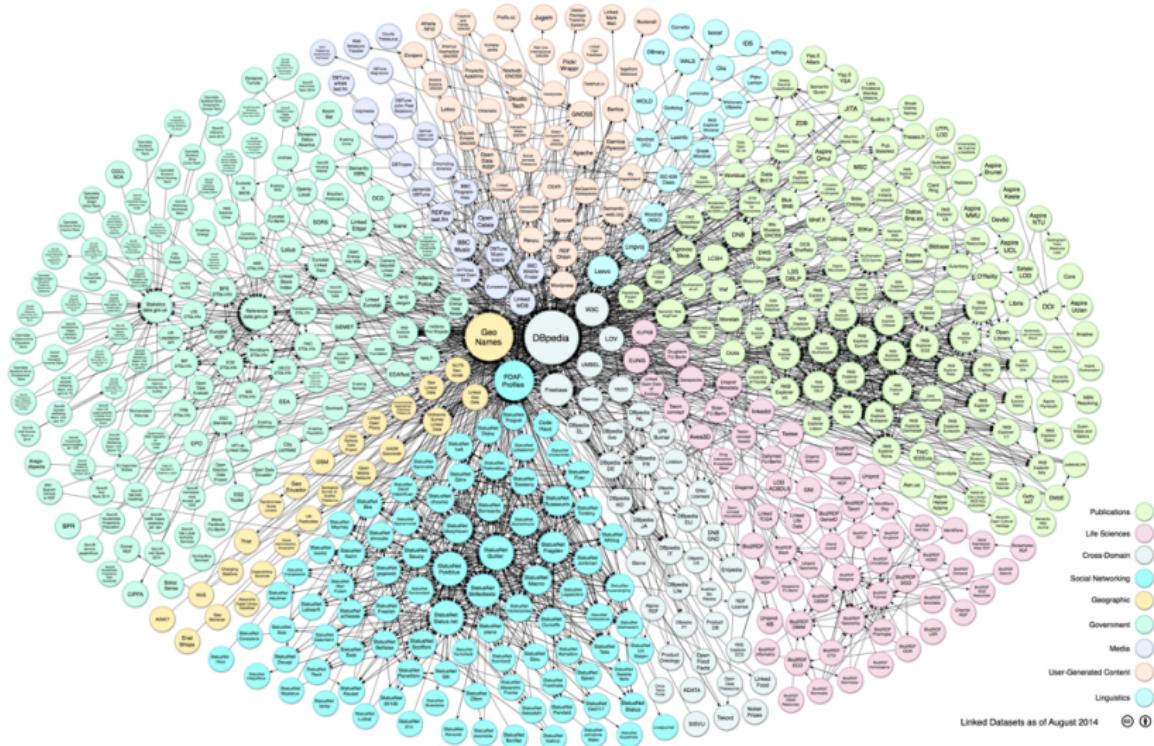




se extiende a una multitud de medios sociales ...



y la tendencia sigue ... Linking Open Data (LOD)

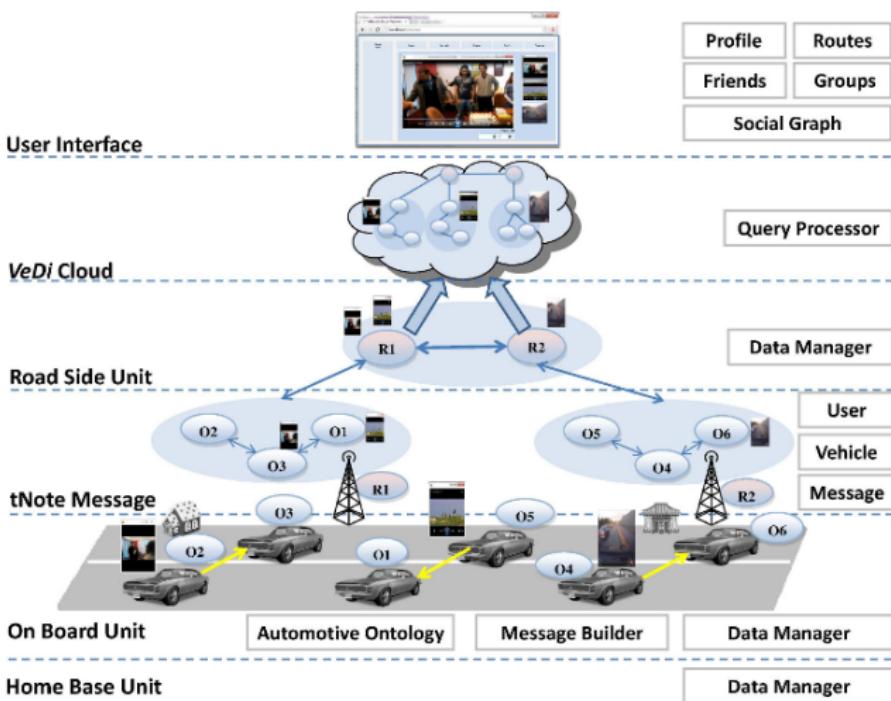


y en un futuro ... Web of Things (WoT)

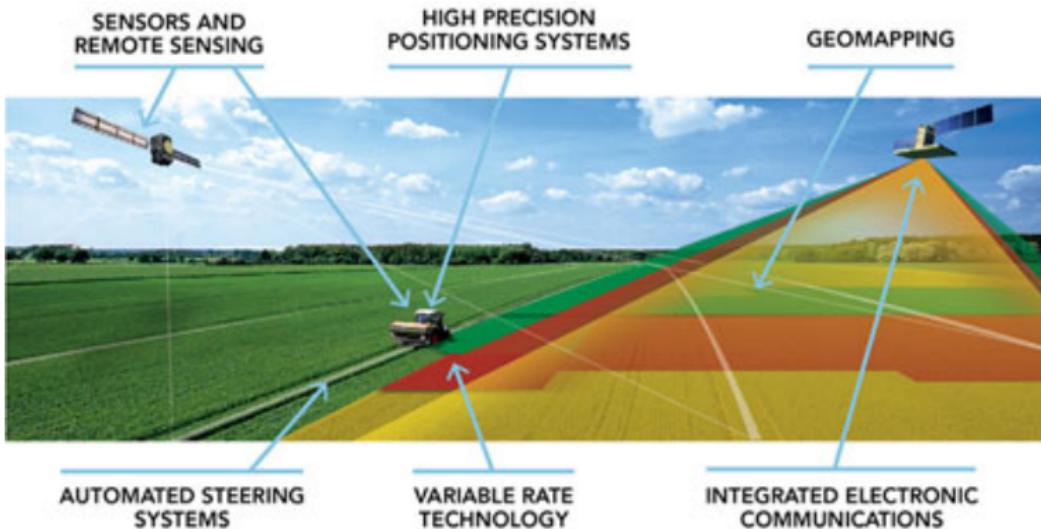


con tendencias similares en la organización vehicular ...

2



y también en el agro



Resumen: ¿Por qué surge la Minería de Datos?

- El análisis e interpretación **manual** de los datos se torna **impráctico**.
- Gran **acumulación de datos**. Causas:
 - El **registro electrónico** de transacciones comerciales.
 - El **registro electrónico** de actividades en Internet.
 - **Flujos de datos masivos** originados en fuentes diversas.
 - El **almacenamiento** es más **barato**.
- Algoritmos **eficientes** y **robustos** para el procesamiento de estos datos.
- Poder computacional más barato.
- Ventajas **comerciales** y **científicas**.
- Alta dimensionalidad.
- Datos **heterogéneos** y **complejos**.

Ejemplo 1: Análisis de datos del carro de compras

Idcar	Huevos	Aceite	Pañales	Vino	Leche	Manteca	Salmón	Lechugas	...
1	si	no	no	si	no	si	si	si	...
2	no	si	no	no	si	no	no	si	...
3	no	no	si	no	si	no	no	no	...
4	no	si	si	no	si	no	no	no	...
5	si	si	no	no	no	si	no	si	...
6	si	no	no	si	si	si	si	no	...
7	no	no	no	no	no	no	no	no	...
8	si	si	si	si	si	si	si	no	...
...

Descripción de los datos:

- **Objetos/instancias:** contenido de un “**carro de compras**”.
- **Atributos:** corresponden a los distintos **productos del supermercado**.
- **Valores de los atributos:** todos de tipo **booleano**.

Ejemplo 1: Análisis de datos del carro de compras

Idcar	Huevos	Aceite	Pañales	Vino	Leche	Manteca	Salmón	Lechugas	...
1	si	no	no	si	no	si	si	si	...
2	no	si	no	no	si	no	no	si	...
3	no	no	si	no	si	no	no	no	...
4	no	si	si	no	si	no	no	no	...
5	si	si	no	no	no	si	no	si	...
6	si	no	no	si	si	si	si	no	...
7	no	no	no	no	no	no	no	no	...
8	si	si	si	si	si	si	si	no	...
...

De estos datos, un algoritmo de MD podría determinar que:

- 100 % de las veces que se compra **pañales** también se compra **leche**
- 50 % de las compras de **huevos** también incluyen **aceite**
- 33 % de las veces que se compra **vino** y **salmón** también se compra **lechuga**

Ejemplo 2: Ventas mensuales durante el último año

Producto	mes-12	...	mes-4	mes-3	mes-2	mes-1
televisor plano 30' Phlipis	20	...	52	14	139	74
video-dvd-recorder Miesens	11	...	43	32	26	59
discman mp3 LJ	50	...	61	14	5	28
frigorífico no frost Jazzussi	3	...	21	27	1	49
microondas con grill Sanson	14	...	27	2	25	12
...

Descripción de los datos:

- **Objetos/instancias:** ventas mensuales por electrodoméstico.
- **Atributos:** corresponden a los distintos meses del año.
- **Valores de los atributos:** todos de tipo entero (numéricos).

Ejemplo 2: Ventas mensuales durante el último año

Producto	mes-12	...	mes-4	mes-3	mes-2	mes-1
televisor plano 30' Phlipis	20	...	52	14	139	74
video-dvd-recorder Miesens	11	...	43	32	26	59
discman mp3 LJ	50	...	61	14	5	28
frigorífico no frost Jazzussi	3	...	21	27	1	49
microondas con grill Sanson	14	...	27	2	25	12
...

Un algoritmo de MD podría generar un modelo para **predecir** las **ventas** del mes siguiente

Ejemplo 3: Análisis de riesgo en créditos bancarios

Datos (Entrada)

IDC	D-crédito	C-crédito	Salario	Casa	Cuentas-Mor	...	Devuelve-C
101	15	60.000	2.200	si	2	...	no
102	2	30.000	3.500	si	0	...	si
103	9	9.000	1.700	si	1	...	no
104	15	18.000	1.900	no	0	...	si
105	10	24.000	2.100	no	0	...	no
...

Descripción de los datos:

- **Objetos/instancias:** información por cada **cliente bancario**.
- **Atributos:** crédito otorgado, salario, casa, devolución crédito, etc.
- **Valores de los atributos:** de tipo **numéricos, booleanos**, etc.

Ejemplo 3: Análisis de riesgo en créditos bancarios

Datos (Entrada)

IDC	D-crédito	C-crédito	Salario	Casa	Cuentas-Mor	...	Devuelve-C
101	15	60.000	2.200	si	2	...	no
102	2	30.000	3.500	si	0	...	si
103	9	9.000	1.700	si	1	...	no
104	15	18.000	1.900	no	0	...	si
105	10	24.000	2.100	no	0	...	no
...

Patrones (Posible Salida)

SI Cuentas-Mor > 0 ENTONCES Devuelve-C = no

SI Cuentas-Mor = 0 Y [(Salario > 2500) O (D-crédito > 10)] entonces Devuelve-C = si

Ejemplo 4: Determinar grupos de empleados (Entrada)

Id	Sueldo	Casado	Coche	Hijos	Alq/prop	Sindicado	Bajas/a no	Antig.	Sexo
1	1000	Si	No	0	Alquiler	No	7	15	H
2	2000	No	Si	1	Alquiler	Si	3	3	M
3	1500	Si	Si	2	Prop.	Si	5	10	H
4	3000	Si	Si	1	Alquiler	No	15	7	M
5	1000	Si	Si	0	Prop.	Si	1	6	H
6	4000	No	Si	0	Alquiler	Si	3	16	M
7	2500	No	No	0	Alquiler	Si	0	8	H
8	2000	No	Si	0	Prop	Si	2	6	M
9	2000	Si	Si	3	Prop	No	7	5	H
10	3000	Si	Si	2	Prop	No	1	20	H
11	5000	No	No	0	Alquiler	No	2	12	M
12	800	Si	Si	2	Prop	No	3	1	H
13	2000	No	No	0	Alquiler	No	27	5	M
14	1000	No	Si	0	Alquiler	Si	0	7	H
15	8 00	No	Si	0	Alquiler	No	3	2	H
...

- **Objetos/instancias:** información de cada **empleado**.
- **Atributos:** distintos datos de un empleado.
- **Valores de los atributos:** de tipo **numéricos, booleanos, nominales**, etc.

Ejemplo 4: Determinar grupos de empleados (Salida)

Grupo 1

Sueldo: 1535.2

Casado: No → 0.777, Si → 0.223

Coche: No → 0.82, Si → 0.18

Hijos: 0.05

Alq/Pro: Al → 0.99, Pr → 0.01

Sindic: No → 0.8, Si → 0.2

Bajas/Año: 8.3

Antiguedad: 8.7

Sexo: H → 0.61, M → 0.39

Grupo 2

Sueldo: 1428.7

Casado: No → 0.98, Si → 0.02

Coche: No → 0.01, Si → 0.99

Hijos: 0.3

Alq/Pro: Al → 0.75, Pr → 0.25

Sindic: Si → 1.0

Bajas/Año: 2.3

Antiguedad: 8

Sexo: H → 0.25, M → 0.75

Grupo 3

Sueldo: 1233.8

Casado: Si → 1.0

Coche: No → 0.05, Si → 0.95

Hijos: 2.3

Alq/Pro: Al → 0.17, Pr → 0.83

Sindic: No → 0.67, Si → 0.33

Bajas/Año: 5.1

Antiguedad: 8.1

Sexo: H → 0.83, M → 0.17

Ejemplo 5: Problema del tiempo

Datos (Entrada)

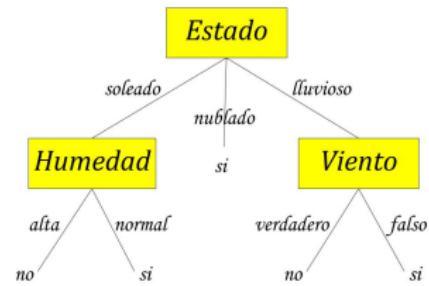
Ej.	Atributos				
	Estado	Temp	Humedad	Viento	JT
e ₁	soleado	caluroso	alta	falso	no
e ₂	soleado	caluroso	alta	verdadero	no
e ₃	nublado	caluroso	alta	falso	si
e ₄	lluvioso	templado	alta	falso	si
e ₅	lluvioso	fresco	normal	falso	si
e ₆	lluvioso	fresco	normal	verdadero	no
e ₇	nublado	fresco	normal	verdadero	si
e ₈	soleado	templado	alta	falso	no
e ₉	soleado	fresco	normal	falso	si
e ₁₀	lluvioso	templado	normal	falso	si
e ₁₁	soleado	templado	normal	verdadero	si
e ₁₂	nublado	templado	alta	verdadero	si
e ₁₃	nublado	caluroso	normal	falso	si
e ₁₄	lluvioso	templado	alta	verdadero	no

Ejemplo 5: Problema del tiempo

Datos (Entrada)

Ej.	Atributos				
	Estado	Temp	Humedad	Viento	JT
e ₁	soleado	caluroso	alta	falso	no
e ₂	soleado	caluroso	alta	verdadero	no
e ₃	nublado	caluroso	alta	falso	si
e ₄	lluvioso	templado	alta	falso	si
e ₅	lluvioso	fresco	normal	falso	si
e ₆	lluvioso	fresco	normal	verdadero	no
e ₇	nublado	fresco	normal	verdadero	si
e ₈	soleado	templado	alta	falso	no
e ₉	soleado	fresco	normal	falso	si
e ₁₀	lluvioso	templado	normal	falso	si
e ₁₁	soleado	templado	normal	verdadero	si
e ₁₂	nublado	templado	alta	verdadero	si
e ₁₃	nublado	caluroso	normal	falso	si
e ₁₄	lluvioso	templado	alta	verdadero	no

Modelo (Posible Salida)



Ejemplo 6 (Web): Usuarios de Twitter

- **Objetos/instancias**: información de cada usuario.
 - **Atributos**: distintos datos de su **estructura de red**, comportamiento de **comunicación**, características **socio-lingüisticos** y n -gramas (de palabras y caracteres) .
 - **Valores de los atributos**: de tipo **numéricos**, **booleanos**, **nominales**, etc.

Algunas características de la estructura de red y comportamiento de comunicación.

Ejemplo 6 (Web): Usuarios de Twitter

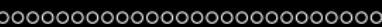
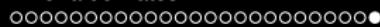
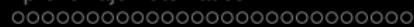
- P-FER-FING: proporción **followers/followings**.
 - F-FER: frecuencia (número) de **followers**.
 - F-FING: frecuencia (número) de los que está **following**.
 - P-RET: porcentaje de tweets del usuario que son **retweets**.
 - P-RES: porcentaje de tweets con **respuestas** del usuario.
 - P-TWE: porcentaje de **tweets** generados por el usuario.

Algunas características de la estructura de red y comportamiento de comunicación.

Ejemplo 6 (Web): Usuarios de Twitter (II)

Características socio-lingüisticas

<i>FEATURE</i>	<i>Description/Example</i>
SIMLEYS	A list of emoticons compiled from the Wikipedia.
OMG	Abbreviation for 'Oh My God'
ELLIPSSES	'....'
POSSESSIVE BIGRAMS	E.g. my_XXX, our_XXX
REPATED ALPHABETS	E.g. niceeeeeee, nooooo waaaaay
SELF	E.g., Lxxx, Im_XXX
LAUGH	E.g. LOL, ROTFL, LMFAO, haha, hehe
SHOUT	Text in ALLCAPS
EXASPERATION	E.g. Ugh, mmmm, hmmmm, ahh, grrr
AGREEMENT	E.g. yea, yeah, ohya
HONORIFICS	E.g. dude, man, bro, sir
AFFECTION	E.g. xoxo
EXCITEMENT	A string of exclamation symbols (!!!!!)
SINGLE EXCLAIM	A single exclamation at the end of the tweet
PUZZLED PUNCT	A combination of any number of ? and ! (!?!!??!)



Ejemplo 6 (Web): Usuarios de Twitter (III)

O derivadas del **contenido** (textos) de los **tweets** (por ejemplo, **palabras**, ver Ejemplo siguiente).

En base a cualquiera de estas características (o **features**), un algoritmo de MD podría generar un modelo para **predecir** los **grupos etarios** (rangos de edad) de cada usuario.

Ejemplo 7: Textos arbitrarios (Representación BOW)

Documentos

- ① "pintaron el banco de la plaza"
- ② "si paso la prueba, iremos paso a paso"
- ③ "no me banco ir al banco a cobrar cheques"

Pesos binarios

ID	a	al	banco	cheques	cobrar	de	el	ir	iremos	la	me	no	paso	pintaron	plaza	prueba	si
t1	0	0	1	0	0	1	1	0	0	1	0	0	0	1	1	0	0
t2	1	0	0	0	0	0	0	0	1	1	0	0	1	0	0	1	1
t3	1	1	1	1	1	0	0	1	0	0	1	1	0	0	0	0	0

Ejemplo 7: Textos arbitrarios (Representación BOW)

Documentos

- ① "pintaron el banco de la plaza"
- ② "si paso la prueba, iremos paso a paso"
- ③ "no me banco ir al banco a cobrar cheques"

Pesos *TF* (Frecuencia del término)

ID	a	al	banco	cheques	cobrar	de	el	ir	iremos	la	me	no	paso	pintaron	plaza	prueba	si
t1	0	0	1	0	0	1	1	0	0	1	0	0	0	1	1	0	0
t2	1	0	0	0	0	0	0	0	1	1	0	0	3	0	0	1	1
t3	1	1	2	1	1	0	0	1	0	0	1	1	0	0	0	0	0

Ejemplo 7: Textos arbitrarios - características estáticas

Documentos

- ① "pintaron el banco de la plaza"
- ② "si paso la prueba, iremos paso a paso"
- ③ "no me banco ir al banco a cobrar cheques"

Número de palabras (NP), longitud de palabra más larga (LPL), longitud promedio de palabras (LPP), verbos en pasado (VP)

ID	NP	LPL	LPP	VP
t1	6	8	4	1
t2	8	6	3,625	0
t3	9	7	3.55	0

Ejemplo 8: clasificando lirios de Iris



- Iris dataset: tiene información de **150 flores** de lirio, con **50 ejemplos** de cada una de las siguientes **especies** de lirio: **Setosa, Versicolour y Virginica**.
- Cada flor de lirio es caracterizada por **5 atributos**
 - **4 numéricos** con las **características (features)** de la flor
 - **1 nominal** con la **clase (especie)** de la flor

Ejemplo 8: clasificando lirios de Iris



- **Atributos** de un lirio:
 - 1 sepal_length: **longitud del sépalo** (en centímetros).
 - 2 sepal_width: **ancho del sépalo** (en centímetros).
 - 3 petal_length: **longitud del pétalo** (en centímetros).
 - 4 petal_width: **longitud del pétalo** (en centímetros).
 - 5 species: **tipo de lirio** (setosa, versicolour o virginica).

Ejemplo 8: clasificando lirios de Iris

- Este conjunto de datos puede ser accedido como un DataFrame de Pandas usando la biblioteca **Seaborn**

Ejemplo 8: clasificando lirios de Iris

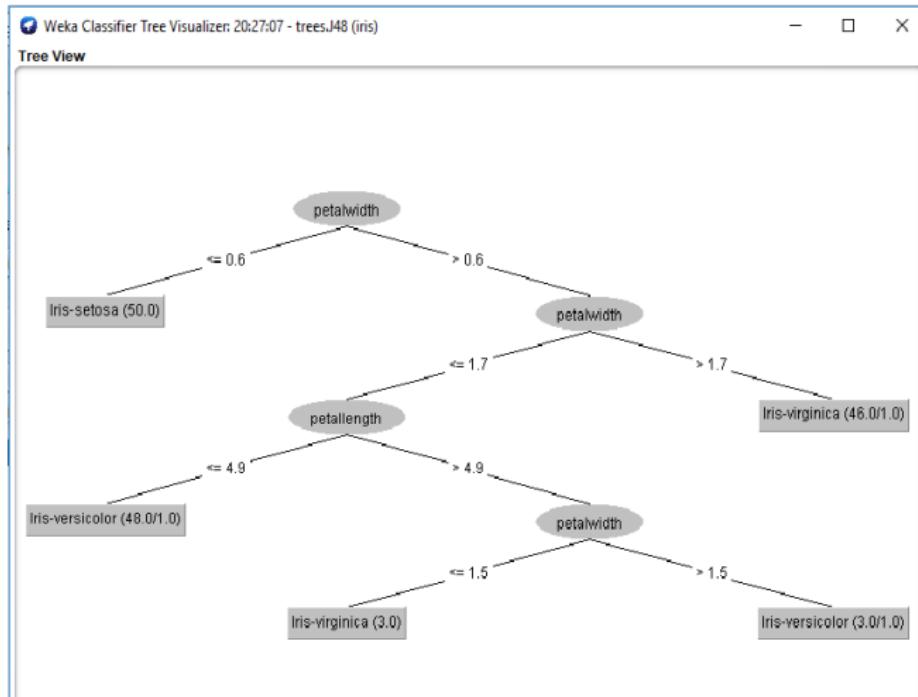
- Este conjunto de datos puede ser accedido como un DataFrame de Pandas usando la biblioteca **Seaborn**
- Cada fila refiere a cada flor observada (muestra)

```
In [1]: import seaborn as sns  
iris = sns.load_dataset('iris')  
iris.head()
```

```
Out[1]:  sepal_length  sepal_width  petal_length  petal_width  species  
0          5.1          3.5          1.4          0.2  setosa  
1          4.9          3.0          1.4          0.2  setosa  
2          4.7          3.2          1.3          0.2  setosa  
3          4.6          3.1          1.5          0.2  setosa  
4          5.0          3.6          1.4          0.2  setosa
```

Ejemplo 8: clasificando lirios de Iris

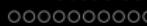
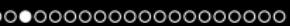
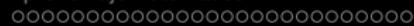
Con Iris, Weka podría aprender este árbol de decisión



Tareas de la MD

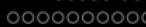
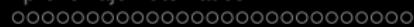
Las tareas de la MD suelen ser divididas en dos grandes categorías:

- Tareas *predictivas*: predicen el valor de un atributo particular (variable **dependiente** u **objetivo**) en base a los valores de los otros atributos (variables **independientes** o **explicatorias**). Ejemplos: **clasificación** y **regresión**.
- Tareas *descriptivas*: obtienen **patrones** (correlaciones, tendencias, grupos, trayectorias y anomalías) que resumen las relaciones subyacentes en los datos. Usualmente son tareas **exploratorias** que requieren pasos posteriores de validación y explicación. Ejemplos: **análisis de grupos** y **análisis de asociación**.



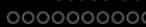
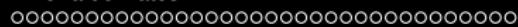
Algunas tareas de la MD

- Categorización / Clasificación
- Regresión
- Análisis de grupos (*cluster analysis*)
- Análisis de asociaciones (*association analysis*)
- Análisis de correlaciones



Tareas Predictivas

- Categorización / Clasificación <——
- Regresión <——
- Análisis de grupos (*cluster analysis*)
- Análisis de asociaciones (*association analysis*)
- Análisis de correlaciones



Tareas Descriptivas

- Categorización / Clasificación
- Regresión
- Análisis de grupos (*cluster analysis*) <——
- Análisis de asociaciones (*association analysis*) <——
- Análisis de correlaciones <——

Clasificación



08	02	22	97	38	15	00	40	00	75	04	05	07	78	52	12	50	77	81
49	49	99	40	17	81	18	57	60	87	17	40	98	43	69	48	01	56	62
81	49	31	73	55	79	14	29	93	71	40	67	53	30	03	49	13	36	65
52	70	95	23	04	60	11	42	03	68	56	01	32	56	71	37	02	36	91
22	31	16	71	51	60	03	59	41	92	36	54	22	40	40	28	66	33	80
24	47	55	60	39	03	45	02	44	75	33	53	78	36	84	20	35	17	12
12	98	81	28	64	23	67	10	26	38	40	67	59	54	70	66	18	38	64
67	26	20	68	02	62	12	20	95	63	94	39	63	08	40	91	66	49	94
24	55	58	05	66	73	99	26	97	17	78	78	96	83	14	88	34	89	63
21	36	23	09	75	00	76	44	20	45	35	14	00	61	33	97	34	31	95
78	17	53	28	22	75	31	67	15	94	03	80	04	62	16	14	09	53	56
92	16	39	05	42	96	35	31	47	55	58	88	24	01	17	54	24	36	29
57	86	56	00	48	35	71	89	07	05	44	44	37	44	60	21	58	51	54
58	19	80	81	68	05	94	47	69	28	73	92	13	86	52	17	77	04	89
40	52	08	83	97	35	99	16	07	97	57	32	16	26	26	79	33	27	98
66	44	48	87	57	62	20	72	03	46	33	67	46	55	12	32	63	93	53
69	04	42	16	73	38	56	39	11	24	94	72	18	08	46	29	32	40	62
36	20	69	36	41	72	30	23	88	31	53	52	69	82	67	59	85	74	04
16	20	73	35	29	78	31	90	01	74	31	49	71	88	94	41	16	23	57
54	01	70	54	71	83	51	54	69	16	92	33	48	61	43	52	01	89	14
48	40																	

What the computer sees

image classification

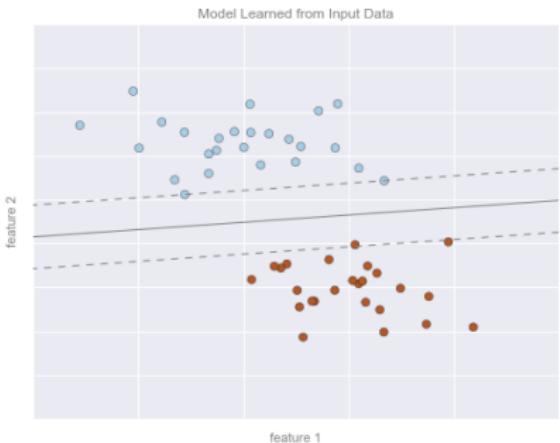
82% cat
15% dog
2% hat
1% mug

Clasificación

Datos Etiquetados

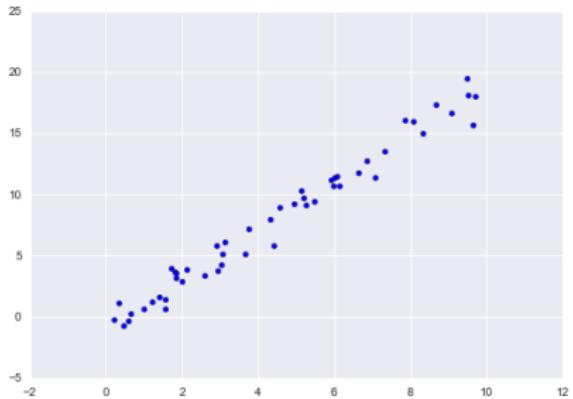


Modelo Aprendido

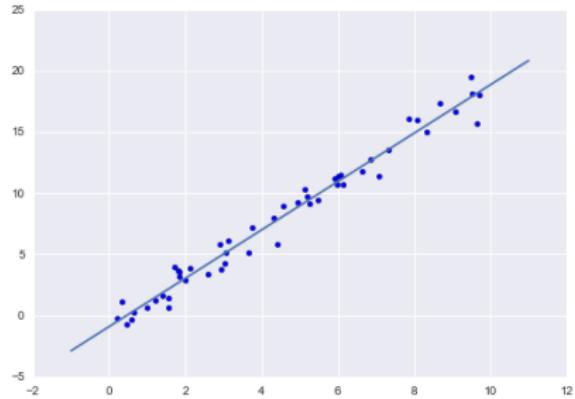


Regresión

Datos Etiquetados



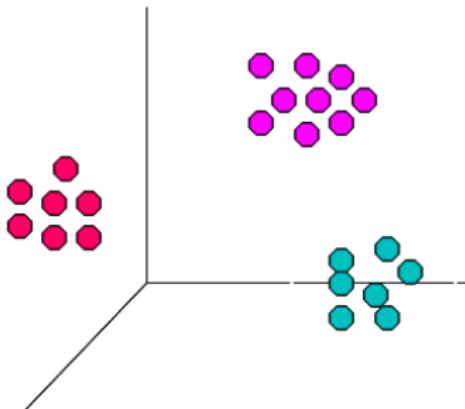
Modelo Aprendido



Análisis de Clusters

Definición

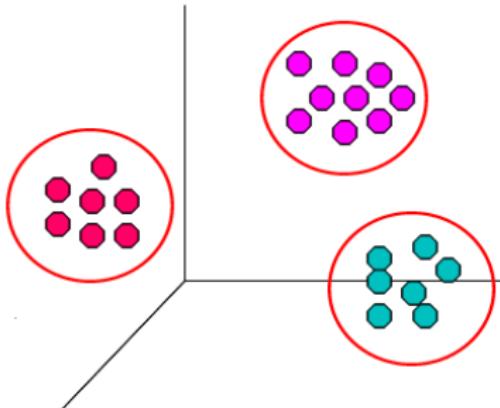
Encontrar grupos de objetos tal que los de un mismo grupo sean similares (o están relacionados) y sean diferentes (o están poco relacionados) con los objetos de los otros grupos.



Análisis de Clusters

Definición

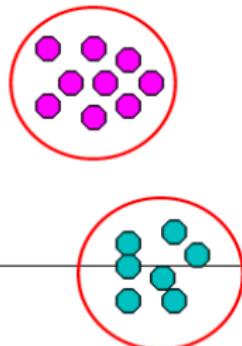
Encontrar **grupos** de objetos tal que los de un mismo grupo sean similares (o están relacionados) y sean diferentes (o están poco relacionados) con los objetos de los otros grupos.



Análisis de Clusters

Definición

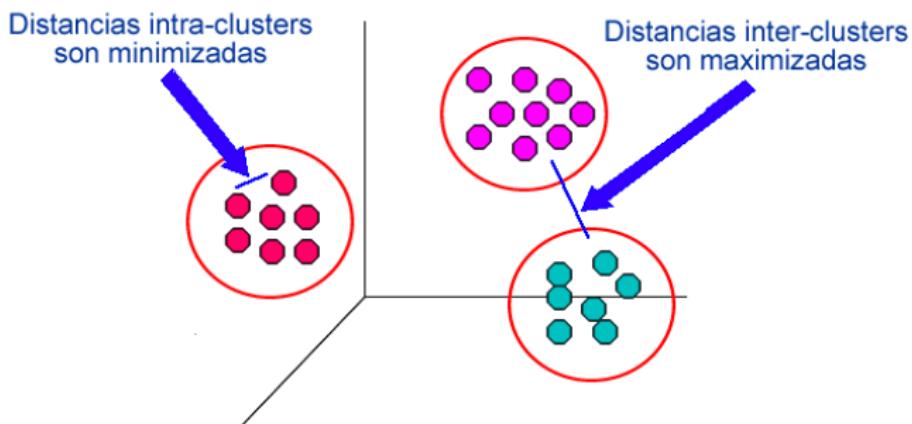
Encontrar grupos de objetos tal que los de un **mismo grupo** sean **similares** (o están **relacionados**) y sean diferentes (o están poco relacionados) con los objetos de los otros grupos.

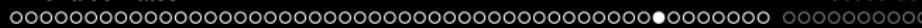


Análisis de Clusters

Definición

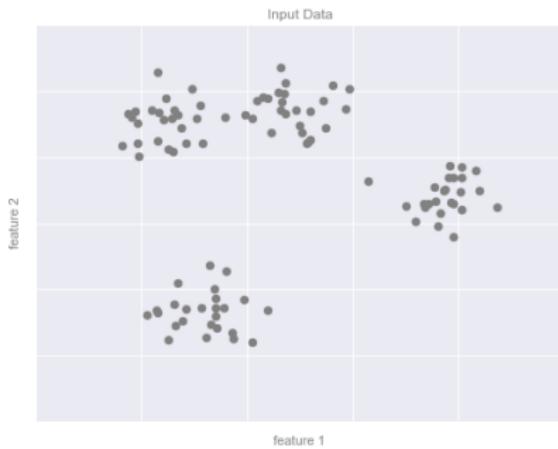
Encontrar grupos de objetos tal que los de un mismo grupo sean similares (o están relacionados) y sean **diferentes** (o están **poco relacionados**) con los objetos de los **otros grupos**.



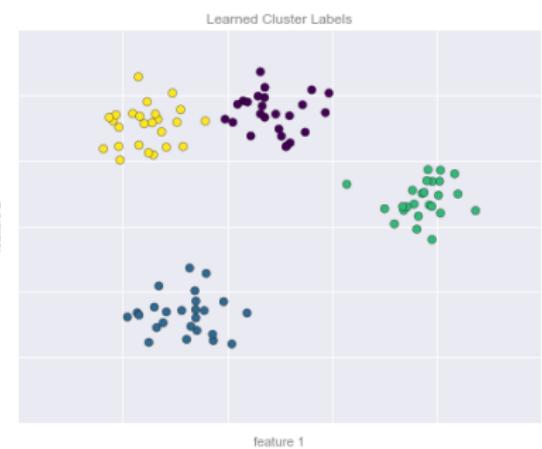


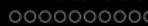
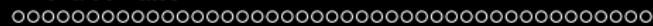
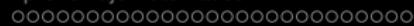
Clustering

Datos (sin etiquetar)



Grupos obtenidos





Agrupamiento

Ejemplos:

- Identificar **grupos de clientes** en una librería en base a sus preferencias de compras.
- Identificar **grupos de personas** con los mismos gustos para una agencia de viaje.
- Determinar **grupos de empleados** diferenciados.
- Agrupar **documentos** que tratan temas/tópicos similares.
- Agrupar documentos escritos por el mismo autor.

Algunas áreas de aplicación de la MD

- Aplicaciones financieras y bancarias, análisis de mercado y comercio en general.
- Seguros y salud privada
- Educación
- Psicología
- Procesos industriales
- Medicina
- Biología, bioingeniería y otras ciencias
- Telecomunicaciones
- Internet
- Turismo, policiales, deportes, política, agro ... y muchas más

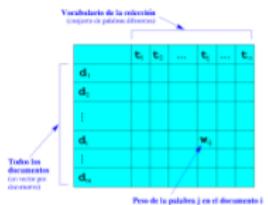
Algunos tipos particulares de MD

- Minería de **textos** (**text mining**)
- Minería de la **Web** (**web mining**)
- Minería de **imágenes** (**image mining**)
- Minería de **audios** y **videos** (**audio and video (data) mining**)
- Minería de datos en **flujos continuos** (**stream (data) mining**)
- Minería de datos de la **educación** (**educational data mining**)

Ejemplo: minería de textos



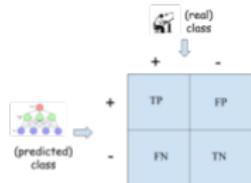
Ejemplo: minería de textos



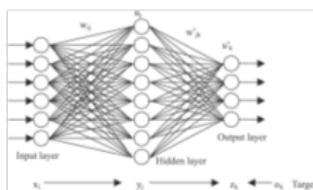
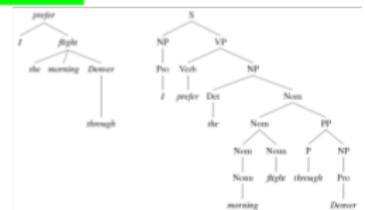
$$\begin{bmatrix} X \\ |V| \times c \end{bmatrix} = \begin{bmatrix} W \\ |V| \times m \end{bmatrix} \begin{bmatrix} C \\ m \times m \end{bmatrix} \begin{bmatrix} m \times c \end{bmatrix}$$



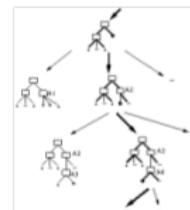
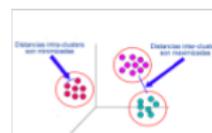
Recuperación de la Información



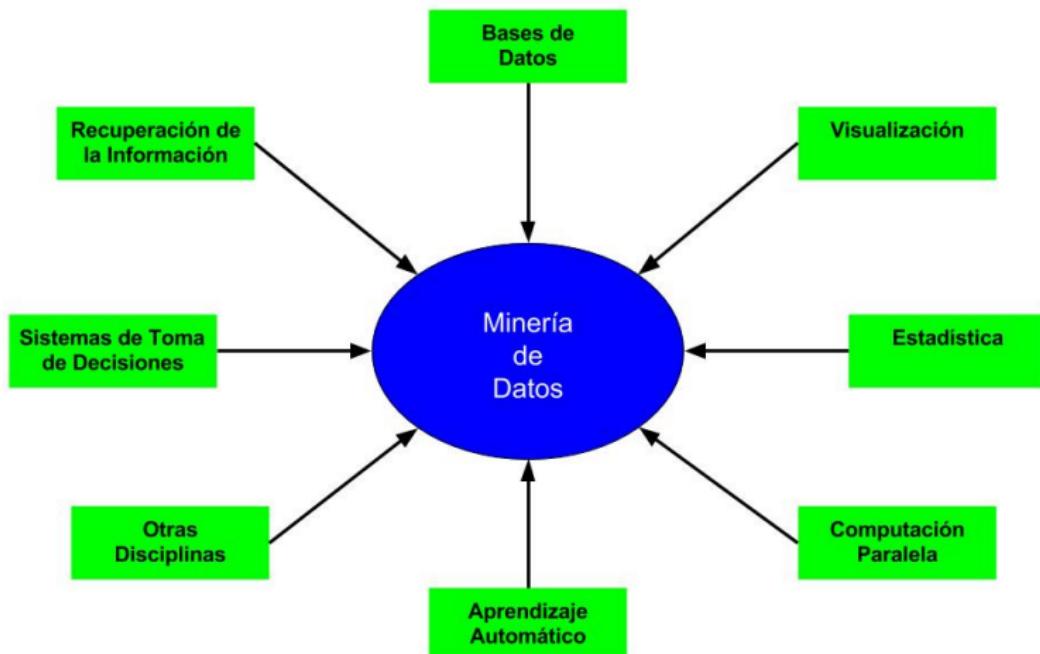
Procesamiento del Lenguaje Natural



Aprendizaje Automático



Relación de la MD con otras disciplinas



Aprendizaje automático versus Minería de Datos

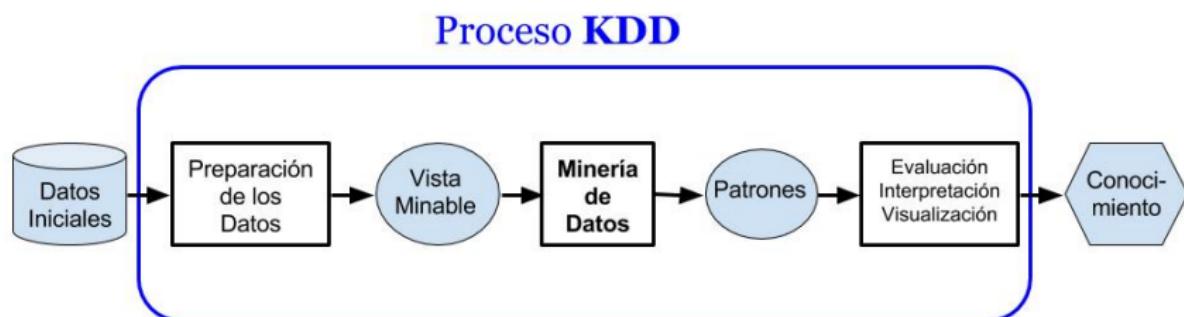
Algunos autores consideran que $AA \equiv MD$ pero ésto no es así.

- MD toma la **experiencia** desde grandes **Bases de datos**. AA incluye **otras formas** de entrenamiento.
- En MD no sólo importa la performance sino también una **representación explícita** del conocimiento adquirido.
- MD suele requerir una **participación humana** considerable.
- MD incluye técnicas **estadísticas** que no son propias del AA.

Minería de Datos y KDD

KDD (Knowledge Discovery in Databases): Proceso no trivial de identificar **patrones** válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles, a partir de los datos.

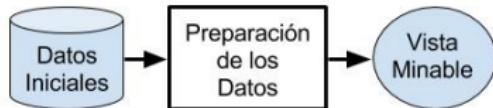
Fases del Proceso KDD



Fase de Preparación de los datos

KDD (*Knowledge Discovery in Databases*): Proceso no trivial de identificar **patrones** válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles, a partir de los datos.

Fases del Proceso KDD



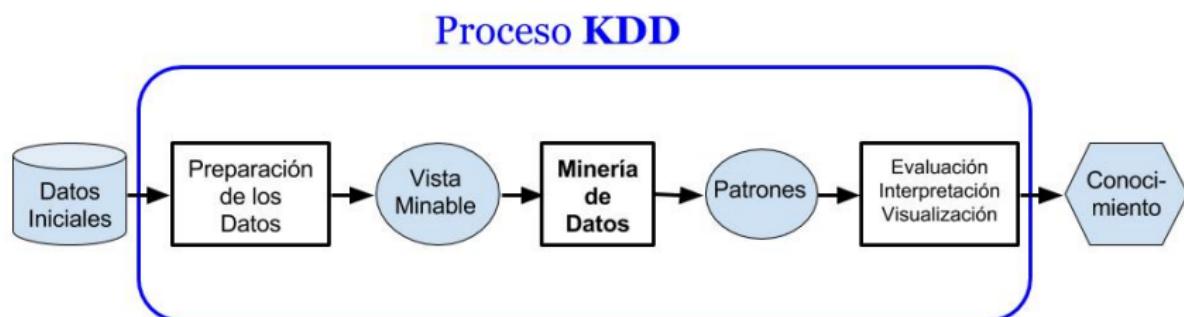
Fase de Preparación de los datos

- Sub-fase de **recopilación e integración de los datos**
 - Determinar fuentes de información útiles y dónde conseguirlas.
 - Coleccionar múltiples bases de datos **heterogéneas** en un único **almacén de datos**.
- Sub-fase de **selección, limpieza y transformación**
 - Detección de valores **anómalos** (no siempre eliminados).
 - Tratamiento de datos **faltantes** (o perdidos).
 - Selección de atributos **relevantes** (columnas).
 - Selección de una **muestra** de datos (filas).
 - Construcción de **nuevos atributos** (agrupamiento, numerización, discretización, normalización).

Minería de Datos y KDD

KDD (Knowledge Discovery in Databases): Proceso no trivial de identificar **patrones** válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles, a partir de los datos.

Fases del Proceso KDD



Fase de Minería de datos

KDD (*Knowledge Discovery in Databases*): Proceso no trivial de identificar **patrones** válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles, a partir de los datos.

Fases del Proceso KDD



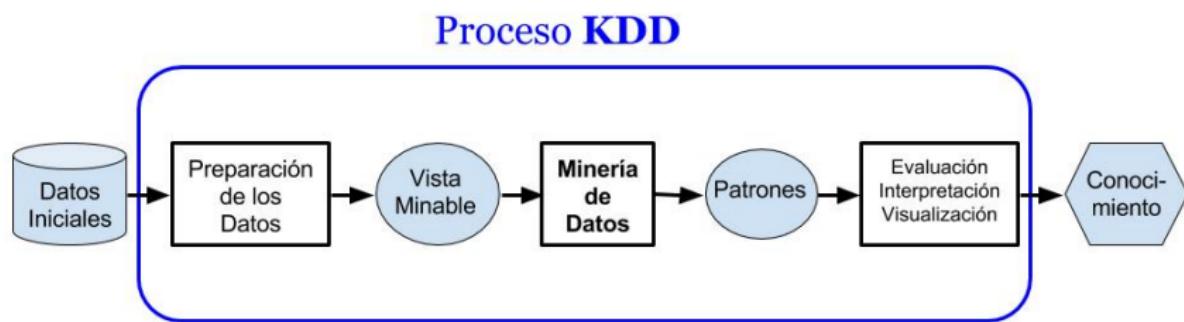
Fase de Minería de datos

- Determinar qué tipo de **tarea** de MD es el más apropiado (clasificación, agrupamiento, etc).
- Elegir tipo de **modelo** (árboles de decisión, reglas de clasificación, Redes Neuronales).
- Elegir el **algoritmo** de minería (o aprendizaje) (CART, C5.0, Backpropagation)

Minería de Datos y KDD

KDD (Knowledge Discovery in Databases): Proceso no trivial de identificar **patrones** válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles, a partir de los datos.

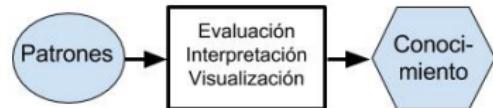
Fases del Proceso KDD

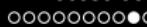
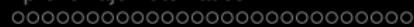


Fase de evaluación, interpretación y visualización

KDD (*Knowledge Discovery in Databases*): Proceso no trivial de identificar **patrones** válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles, a partir de los datos.

Fases del Proceso KDD

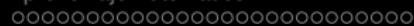




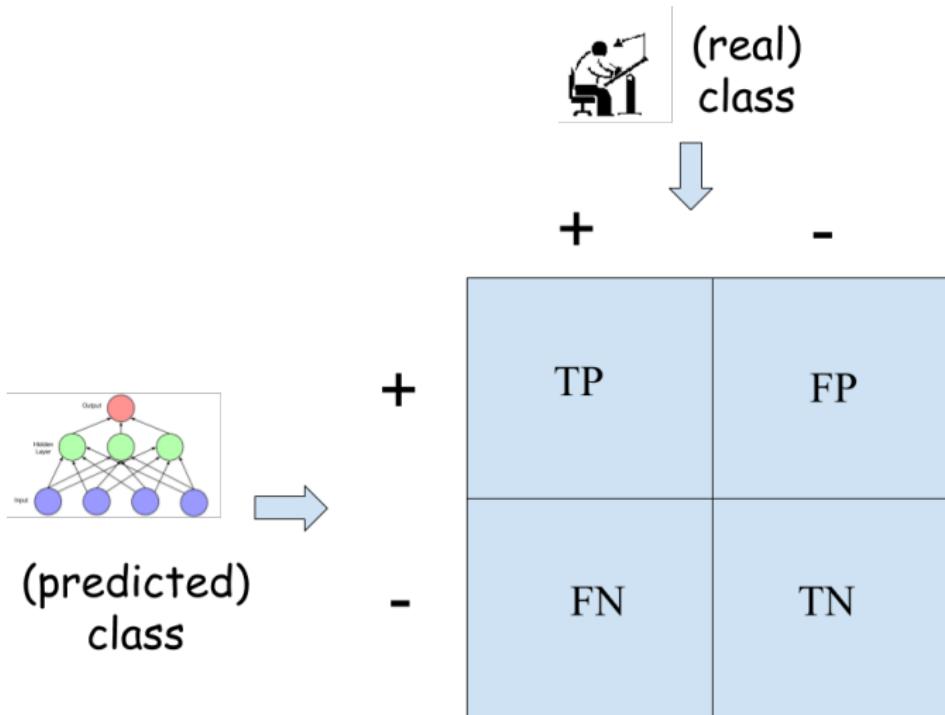
Fase de evaluación, interpretación y visualización

Criterios para la **evaluación** de los modelos (patrones) descubiertos:

- **Precisos**
- **Comprensibles** (inteligibles)
- **Interesantes** (útiles y novedosos)



Evaluación (supervisada) clásica

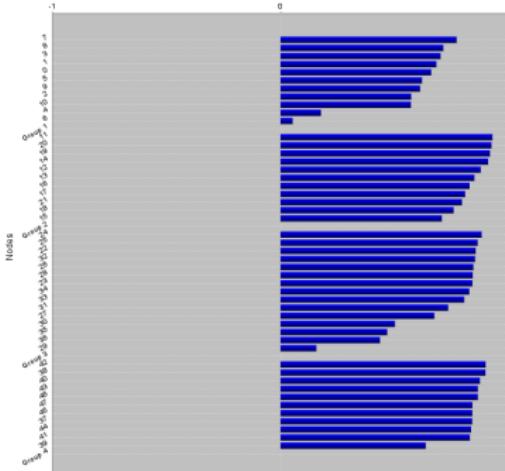


Evaluando agrupamientos

Agrupamiento bueno

Silhouette Graphic

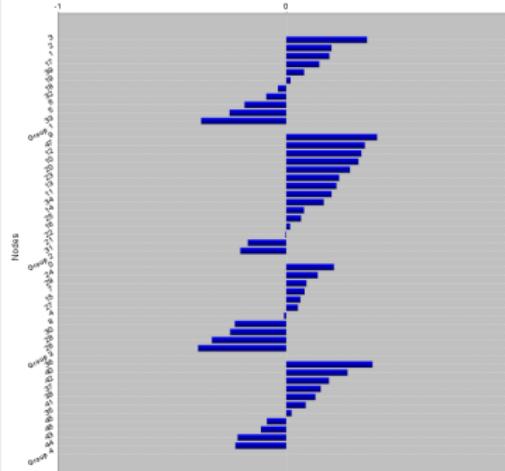
Silhouette Value



Agrupamiento malo

Silhouette Graphic

Silhouette Value



Visualizando información de textos



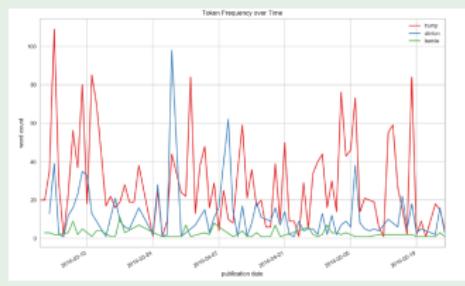
a a a
correlation strength

relative frequency

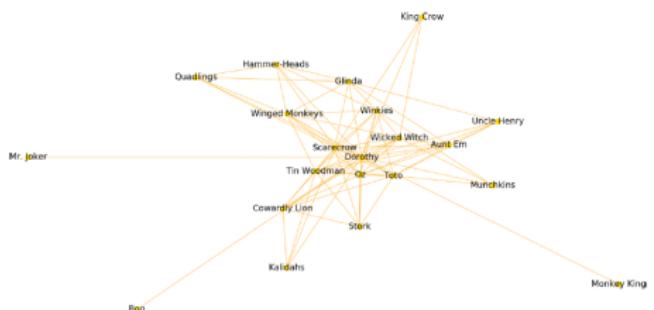
b b b
prevalence in topic

Visualizando información de textos

Frecuencia de palabras en el tiempo

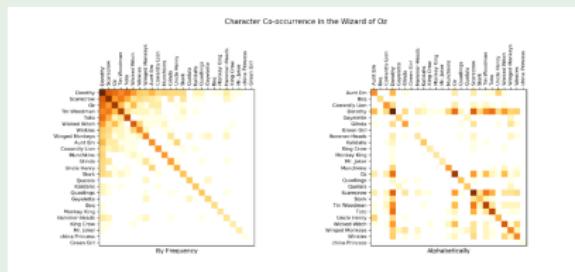


Grafo social de palabras

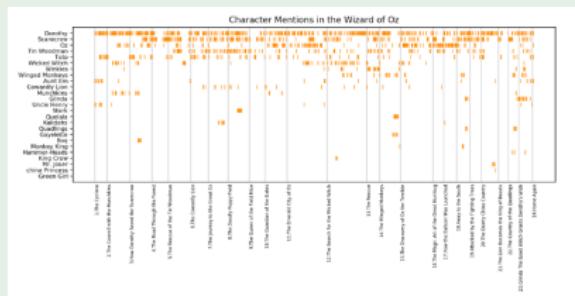


Visualizando información de textos

Matrices de co-ocurrencia

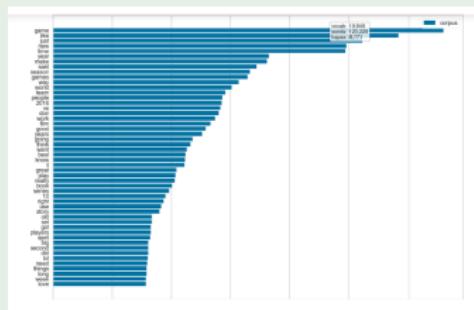


Gráficas de dispersión

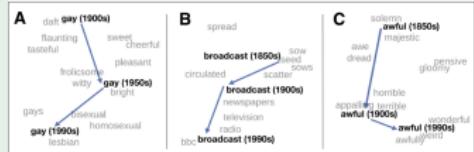


Visualizando información de textos

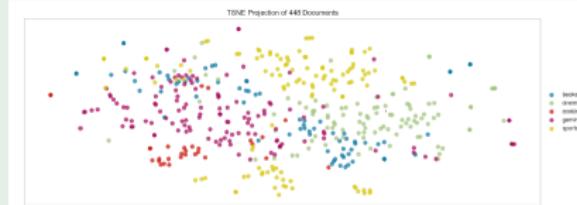
Frecuencias



Evolución en el tiempo

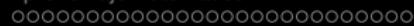


Visualización con t-SNE



Cercanía de embeddings





La *entrada* del proceso de MD

Vista minable: tabla única con todos los atributos relevantes para el proceso de MD.

Atributos: Nos concentraremos en dos tipos de atributos

- Atributos **numéricos**: enteros, reales.
- Atributos **nominales**: también referenciados como atributos **categóricos, enumerados o discretos**.

Ejemplo: tipos de datos en Weka

Sistemas como **Weka**, por ejemplo, trabajan con los tipos de atributos

- ① **numeric**
- ② **integer**: tratado como **numeric**
- ③ **real**: tratado como **numeric**
- ④ **string**
- ⑤ **nominal**: se especifica enumerando los elementos
- ⑥ **date**

Ejemplo: archivo *ARFF* en Weka

```
% Este es un ejemplo sencillo
```

```
@relation ejemplo
```

```
@attribute estadoCivil {sol, cas, sep, vi}
```

```
@attribute alquila {SI, NO}
```

```
@attribute edad integer
```

```
@attribute peso real
```

```
@attribute nombre string
```

```
@data
```

```
sol,SI,28,82.3,"Pepe Pinto"
```

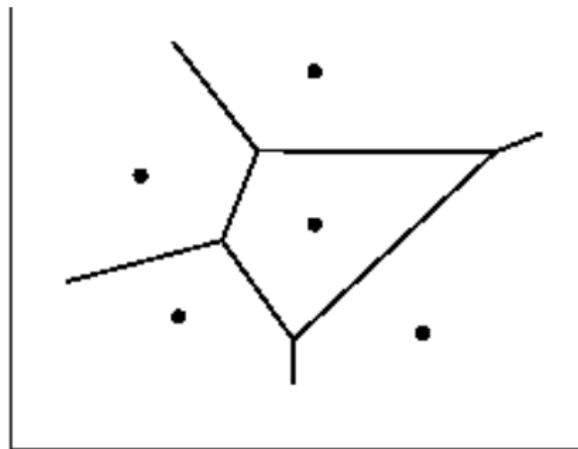
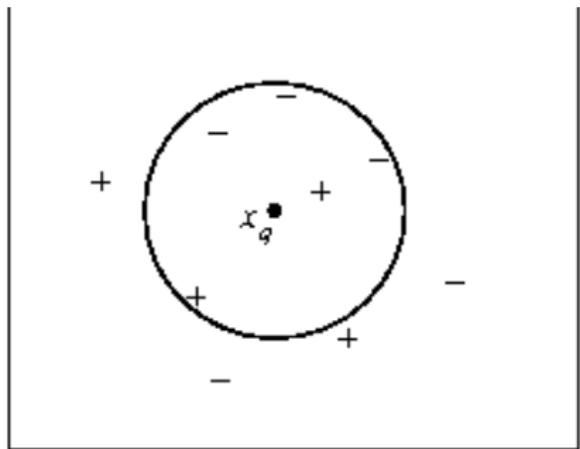
```
cas,NO,42,95.4,"Pedro Tome"
```

```
sep,NO,38,62.8,"Sonia Lario"
```

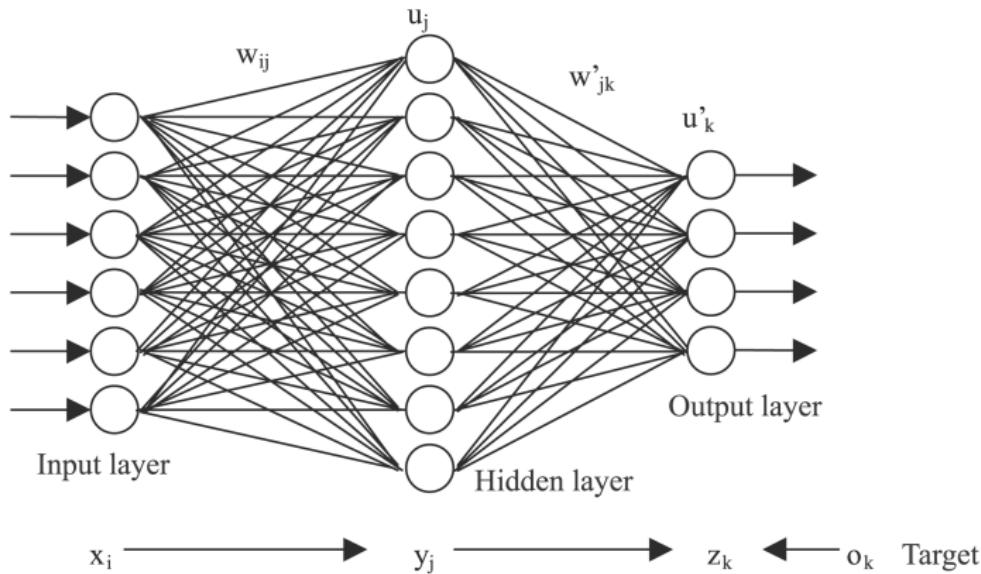
La *salida* del proceso de MD

- Reglas de clasificación
- Clusters (grupos)
- Árboles de decisión
- Redes neuronales
- Reglas de asociación
- Reglas relacionales (ILP)
- Reglas difusas
- Ecuaciones de regresión
- Árboles de regresión
- K-NN y CBR (Case-based reasoning)
- Modelos Bayesianos

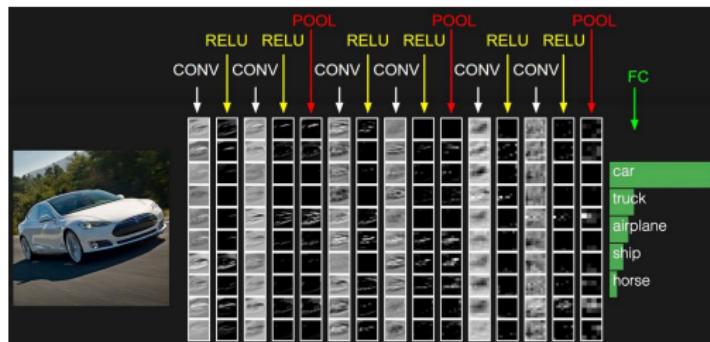
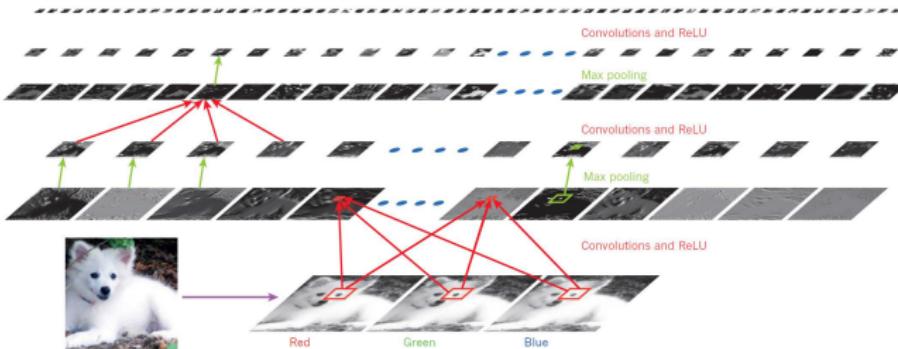
Un clasificador muy simple: k -NN



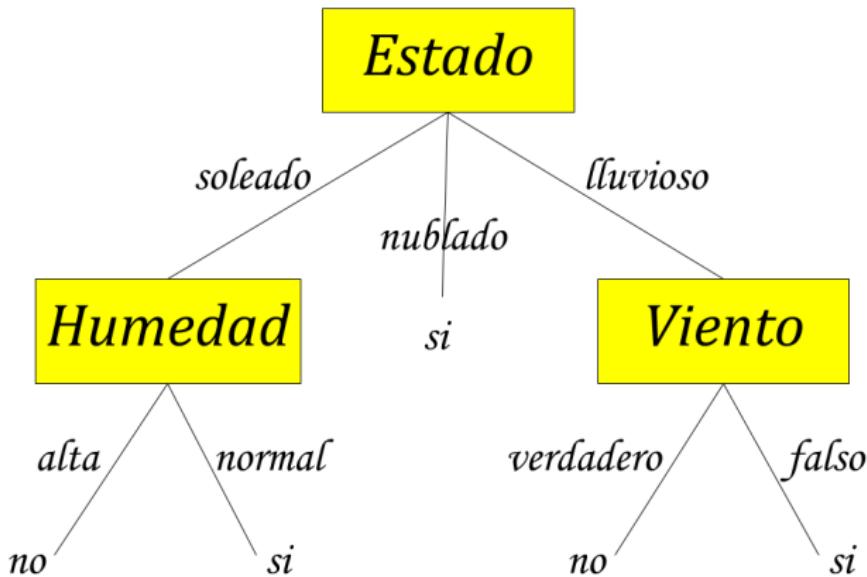
Otro clasificador muy usado: redes neuronales (NN)



... con renovados bríos en arquitecturas “profundas”



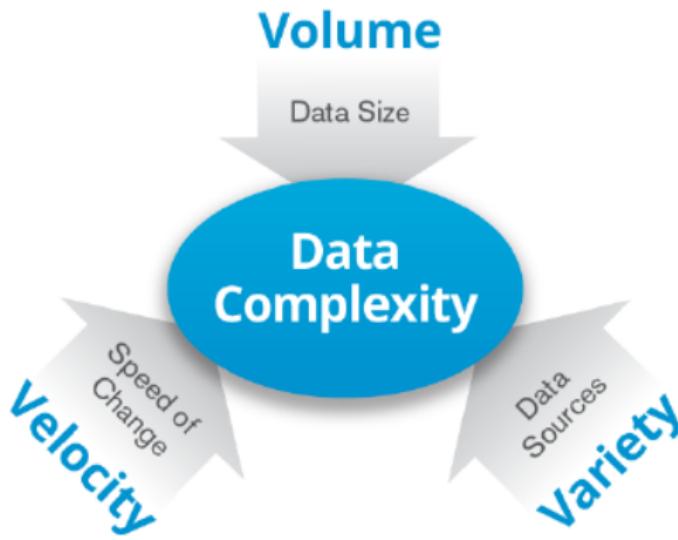
Árboles de decisión



Big Data (Analytics)

Se lo puede ver como aquellos **procesos KDD** donde los datos presentan una **mayor complejidad** debido a las **3 V's**

- Volúmen
- Velocidad
- Variedad



Big Data (Analytics)

- **Volúmen**: Grandes cantidades de datos (en el orden de terabytes hasta zettabytes).
- **Velocidad**: Flujos de datos llegando a gran velocidad con restricciones de tiempo y espacio para su procesamiento.
- **Variedad**: Datos provienen de diferentes fuentes de datos y en formatos muy variados (datos de transacciones, datos estructurados (ej. tablas de bases de datos), datos semiestructurados (ej. XML) y datos no estructurados como texto, imágenes, videos, audios, etc.

Estas características plantean **grandes desafíos** a las **técnicas clásicas de MD**, ya que se deben **analizar y combinar** estos flujos de datos **enormes y veloces** prácticamente en **tiempo real**

¿Por qué usaremos Python?

- Se ha tornado el **lenguaje universal** para muchas aplicaciones de ciencia de datos.

¿Por qué usaremos Python?

- Se ha tornado el **lenguaje universal** para muchas aplicaciones de ciencia de datos.
- Combina el **poder** de lenguajes de programación de propósito general con la **facilidad de uso** de lenguajes scripts como MATLAB o R.

¿Por qué usaremos Python?

- Se ha tornado el **lenguaje universal** para muchas aplicaciones de ciencia de datos.
- Combina el **poder** de lenguajes de programación de propósito general con la **facilidad de uso** de lenguajes scripts como MATLAB o R.
- Python tiene innumerables bibliotecas para la carga y visualización de datos, estadísticas, procesamiento del lenguaje natural, procesamiento de imágenes, etc.

¿Por qué usaremos Python?

- Se ha tornado el **lenguaje universal** para muchas aplicaciones de ciencia de datos.
- Combina el **poder** de lenguajes de programación de propósito general con la **facilidad de uso** de lenguajes scripts como MATLAB o R.
- Python tiene innumerables bibliotecas para la carga y visualización de datos, estadísticas, procesamiento del lenguaje natural, procesamiento de imágenes, etc.
- Herramientas como **Jupyter Notebooks** permiten interactuar directamente con el código proveyendo una rápida iteración y fácil interacción que caracteriza a los procesos de aprendizaje automático y análisis de datos.

¿Por qué usaremos Python?

- Se ha tornado el **lenguaje universal** para muchas aplicaciones de ciencia de datos.
- Combina el **poder** de lenguajes de programación de propósito general con la **facilidad de uso** de lenguajes scripts como MATLAB o R.
- Python tiene innumerables bibliotecas para la carga y visualización de datos, estadísticas, procesamiento del lenguaje natural, procesamiento de imágenes, etc.
- Herramientas como **Jupyter Notebooks** permiten interactuar directamente con el código proveyendo una rápida iteración y fácil interacción que caracteriza a los procesos de aprendizaje automático y análisis de datos.
- Permite la creación de complejas interfaces de usuario gráficas y servicios Web y la integración a sistemas existentes.

Bibliotecas esenciales y herramientas

- **Scikit-Learn**: proyecto de código abierto ampliamente usado en la industria y la academia. La biblioteca para aprendizaje automático en Python más destacada.

Bibliotecas esenciales y herramientas

- **Scikit-Learn**: proyecto de código abierto ampliamente usado en la industria y la academia. La biblioteca para aprendizaje automático en Python más destacada.
- **Jupyter Notebook**: ambiente interactivo para ejecución de código Python en el navegador.

Bibliotecas esenciales y herramientas

- **Scikit-Learn**: proyecto de código abierto ampliamente usado en la industria y la academia. La biblioteca para aprendizaje automático en Python más destacada.
- **Jupyter Notebook**: ambiente interactivo para ejecución de código Python en el navegador.
- **NumPy**: soporte para **arreglos multi-dimensionales**, operaciones de álgebra lineal, transformada de Fourier, generador de números pseudo-random, etc

Bibliotecas esenciales y herramientas

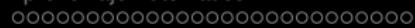
- **Scikit-Learn**: proyecto de código abierto ampliamente usado en la industria y la academia. La biblioteca para aprendizaje automático en Python más destacada.
- **Jupyter Notebook**: ambiente interactivo para ejecución de código Python en el navegador.
- **NumPy**: soporte para **arreglos multi-dimensionales**, operaciones de álgebra lineal, transformada de Fourier, generador de números pseudo-random, etc
- **SciPy**: matrices **ralas**, optimización de funciones matemáticas, procesamiento de señales, distribuciones estadísticas, etc

Bibliotecas esenciales y herramientas

- **Scikit-Learn**: proyecto de código abierto ampliamente usado en la industria y la academia. La biblioteca para aprendizaje automático en Python más destacada.
- **Jupyter Notebook**: ambiente interactivo para ejecución de código Python en el navegador.
- **NumPy**: soporte para **arreglos multi-dimensionales**, operaciones de álgebra lineal, transformada de Fourier, generador de números pseudo-random, etc
- **SciPy**: matrices **ralas**, optimización de funciones matemáticas, procesamiento de señales, distribuciones estadísticas, etc
- **matplotlib**: la principal biblioteca para **gráficas científicas y visualización** de datos en Python.

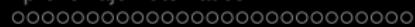
Bibliotecas esenciales y herramientas

- **Scikit-Learn**: proyecto de código abierto ampliamente usado en la industria y la academia. La biblioteca para aprendizaje automático en Python más destacada.
- **Jupyter Notebook**: ambiente interactivo para ejecución de código Python en el navegador.
- **NumPy**: soporte para **arreglos multi-dimensionales**, operaciones de álgebra lineal, transformada de Fourier, generador de números pseudo-random, etc
- **SciPy**: matrices **ralas**, optimización de funciones matemáticas, procesamiento de señales, distribuciones estadísticas, etc
- **matplotlib**: la principal biblioteca para **gráficas científicas y visualización** de datos en Python.
- **pandas**: biblioteca Python para **manipulación de datos y análisis**. Organizada alrededor de los DataFrame (tablas).



¿Por qué usaremos Weka?

- Ambiente para MD **muy intuitivo** y **simple** de usar.



¿Por qué usaremos Weka?

- Ambiente para MD **muy intuitivo** y **simple** de usar.
- **No requiere** de habilidades previas de **programación**.

¿Por qué usaremos Weka?

- Ambiente para MD **muy intuitivo** y **simple** de usar.
- **No requiere** de habilidades previas de **programación**.
- Permite fácil y rápida **experimentación** en pocos y sencillos pasos.

¿Por qué usaremos Weka?

- Ambiente para MD **muy intuitivo** y **simple** de usar.
- **No requiere** de habilidades previas de **programación**.
- Permite fácil y rápida **experimentación** en pocos y sencillos pasos.
- Tal vez no tan robusto para el desarrollo como **R** o **RapidMiner**, pero de **uso difundido** en aplicaciones **industriales** y **científicas**.

¿Por qué usaremos Weka?

- Ambiente para MD **muy intuitivo** y **simple** de usar.
- **No requiere** de habilidades previas de **programación**.
- Permite fácil y **rápida experimentación** en pocos y sencillos pasos.
- Tal vez no tan robusto para el desarrollo como **R** o **RapidMiner**, pero de **uso difundido** en aplicaciones **industriales** y **científicas**.
- **Actualización** e **incorporación** de nuevas funcionalidades aceptable (data streams, deep learning, etc)

¿Por qué usaremos Weka?

- Ambiente para MD **muy intuitivo** y **simple** de usar.
- **No requiere** de habilidades previas de **programación**.
- Permite fácil y **rápida experimentación** en pocos y sencillos pasos.
- Tal vez no tan robusto para el desarrollo como **R** o **RapidMiner**, pero de **uso difundido** en aplicaciones **industriales** y **científicas**.
- **Actualización** e **incorporación** de nuevas funcionalidades aceptable (data streams, deep learning, etc)
- Para los **programadores**, soporta una interface directa con **Java**, uno de los lenguajes de programación más utilizados a nivel mundial.

¿Por qué usaremos Weka?

- Ambiente para MD **muy intuitivo** y **simple** de usar.
- **No requiere** de habilidades previas de **programación**.
- Permite fácil y **rápida experimentación** en pocos y sencillos pasos.
- Tal vez no tan robusto para el desarrollo como **R** o **RapidMiner**, pero de **uso difundido** en aplicaciones **industriales** y **científicas**.
- **Actualización** e **incorporación** de nuevas funcionalidades aceptable (data streams, deep learning, etc)
- Para los **programadores**, soporta una interface directa con **Java**, uno de los lenguajes de programación más utilizados a nivel mundial.
- Provee herramientas para tareas no soportadas por **Scikit-Learn** (ej: reglas de asociación)