


Clase 5 (A) - Categorización de textos

Marcelo Errecalde^{1,2}

¹Universidad Nacional de San Luis, Argentina 

²Universidad Nacional de la Patagonia Austral, Argentina 



Curso: Minería de Textos
Facultad de Informática - Universidad Nacional de La Plata
23 al 27 de Septiembre de 2019

Resumen

- 1 **Etapas del aprendizaje (supervisado) de clasificadores**
 - Etiquetado
 - Extracción de características
 - Entrenamiento (aprendizaje automático)
 - Evaluación y uso

Algunas *tareas* típicas del ACT

- Agrupamiento de documentos
- Extracción de información
- Análisis de asociaciones y correlaciones
- Categorización de textos

Categorización de textos

Dados

- Una colección de documentos \mathcal{D}
- Un conjunto de **categorías** $\mathcal{C} = \{c_1, \dots, c_{|\mathcal{C}|}\}$

Categorización de textos es la tarea de asignar los documentos en \mathcal{D} a las categorías en \mathcal{C}

Categorización de textos

Ejemplos:

Problema	Texto	Categorías (\mathcal{C})
detección de “spam”	e-mails	{si, no}
identificación de autores	documentos	autores
categorización de noticias	cables de noticias	secciones del periódico
WSD	palabras con su contexto	significados de la palabra
detección de pedófilos	conversación del chat	{si, no}
orientación política	blog	{oficialista, opositor}
Determinar género	twitter	{f, m}
análisis de opiniones	evaluación	{positiva, negativa}

Predicción numérica de textos

Cuando en lugar de asignar categorías a los textos se les asigna un **valor numérico**, se dice que la tarea es de **regresión**.

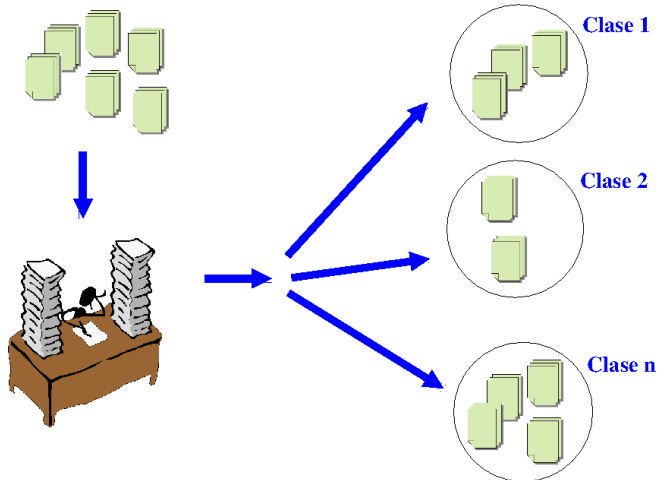
Ejemplos:

Problema	Texto	Valor (número)
estimación de rasgos de personalidad	composiciones escritas	nivel estimado del rasgo
predicción de mortalidad x enfermedad del corazón	mensajes de Twitter	tasa de mortalidad
variación del bienestar	mensajes de Twitter	puntuación en LS

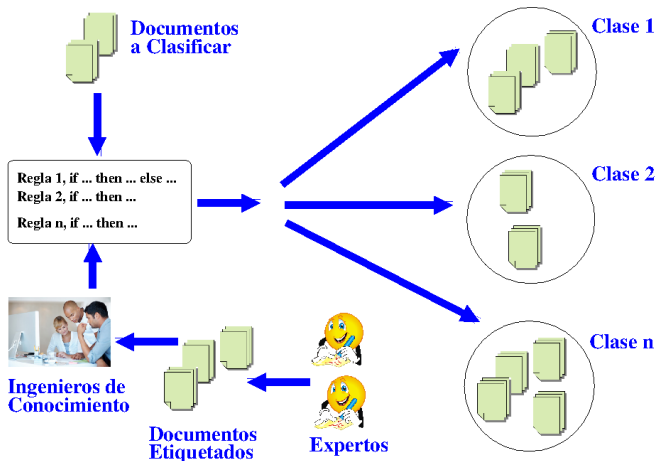
Enfoques para la clasificación de textos

- Categorización **manual**
- Sistemas basados en **reglas** (codificadas **manualmente**)
- Enfoques basados en **aprendizaje automático**

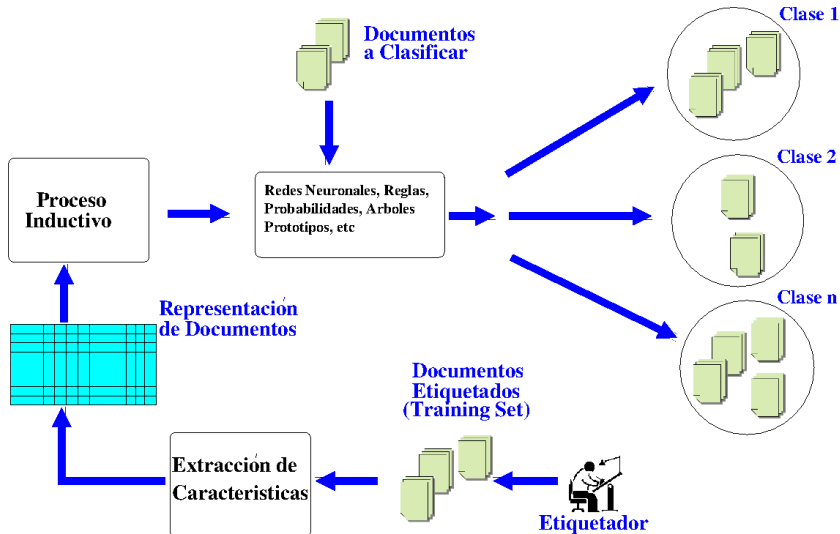
Clasificación Manual



Clasificación basada en reglas (manualmente codificadas)

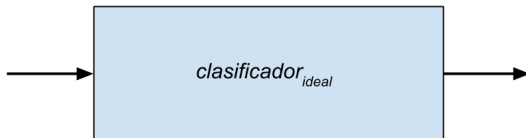


Sistemas de aprendizaje automático



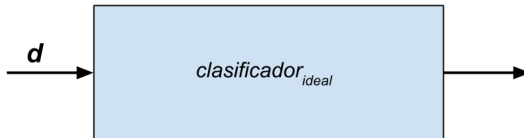
Aprendizaje automático

Idea intuitiva: intentar **reproducir** un proceso de clasificación correcto/ideal (*clasificador_{ideal}*),



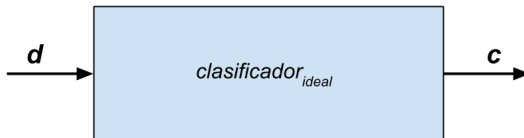
Aprendizaje automático

Idea intuitiva: ... que para cada **entrada** (documento a clasificar) **d**



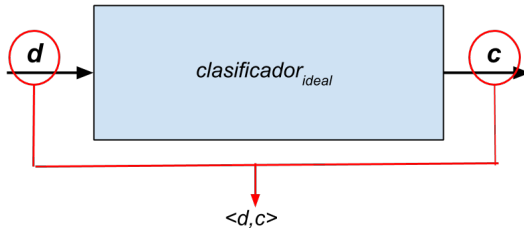
Aprendizaje automático

Idea intuitiva: ... que para cada **entrada** (documento a clasificar) **d** , genera una salida **c** (la clase de **d**)



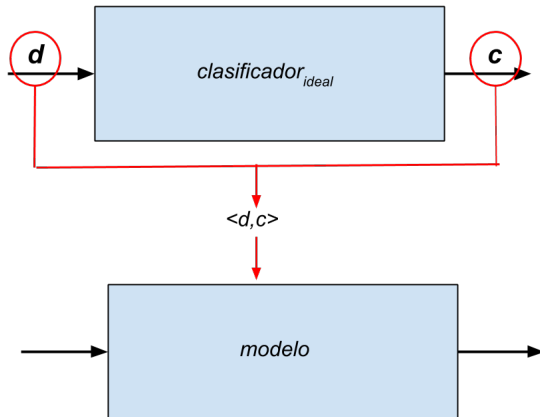
Aprendizaje automático

Idea intuitiva: ... usando ejemplos $\langle d, c \rangle$ del comportamiento de *clasificador*_{ideal},



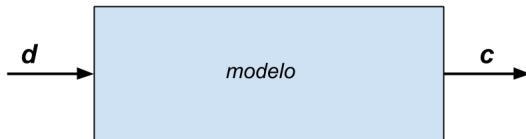
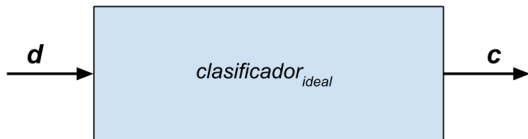
Aprendizaje automático

Idea intuitiva: ... usando **ejemplos** $\langle d, c \rangle$ del comportamiento de *clasificador*_{ideal}, para entrenar otro clasificador (*modelo*)



Aprendizaje automático

Idea intuitiva: ... cuyos comportamientos sean **tan parecidos** como sea posible.



Aprendizaje automático

Puntos **claves**:

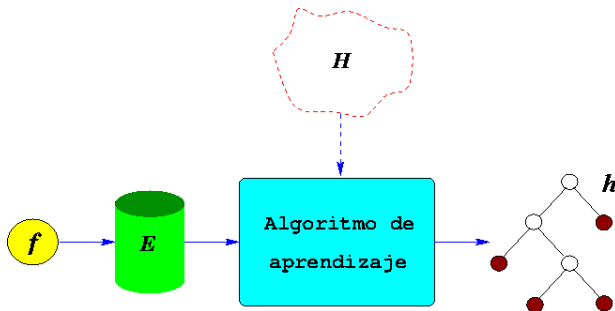
- las salidas (clasificaciones) de *clasificador_{ideal}* y *modelo* deberían coincidir respecto a los ejemplos de entrenamiento pero (y más importante),
- deberían coincidir sobre casos (documentos) no presentes en el conjunto de entrenamiento (**generalizar**)
- Este proceso, en matemática, se conoce como **aproximación de una función**

Aprendizaje de un clasificador

Idea: aproximar la función **ideal** de clasificación:

$$f : \mathcal{D} \mapsto \mathcal{C}$$

con un conjunto de entrenamiento E , de ejemplos $\langle d, c \rangle$, tal que $d \in \mathcal{D}$ es un **documento**, y $c \in \mathcal{C}$ es la **categoría** que f asigna a d .



Aprendizaje de un clasificador (+ formal)

Dados

- Una **función de clasificación** o *clasificación objetivo* desconocida:

$$f : \mathcal{D} \rightarrow \mathcal{C}$$

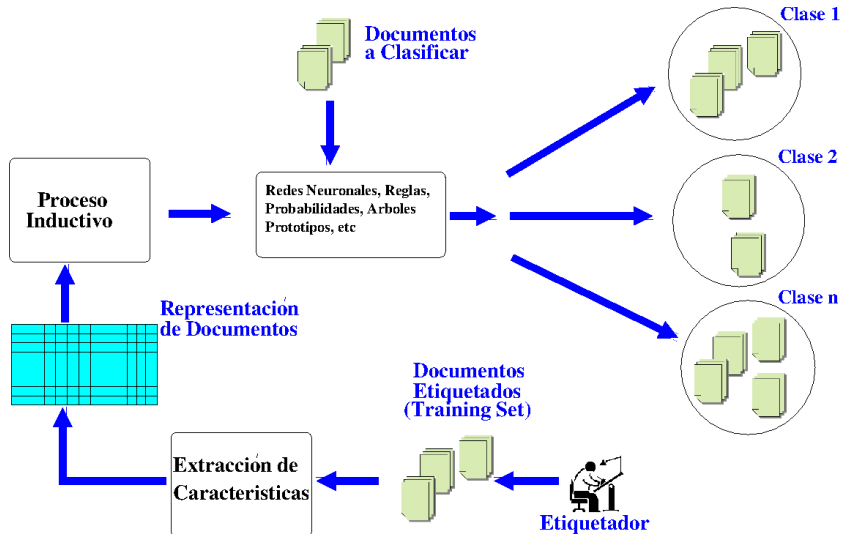
- Un **conjunto de entrenamiento** E , tal que cada ejemplo es una instancia rotulada con una de las posibles clases, $\langle d, f(d) \rangle$ donde $d \in \mathcal{D}$ y $f(d) \in \mathcal{C}$

Tarea: **estimar** c , es decir, encontrar una función:

$$h : \mathcal{D} \mapsto \mathcal{C}$$

denominada *hipótesis clasificadora* o *clasificador*, tal que $h(d) = f(d)$ para todo $d \in \mathcal{D}$.

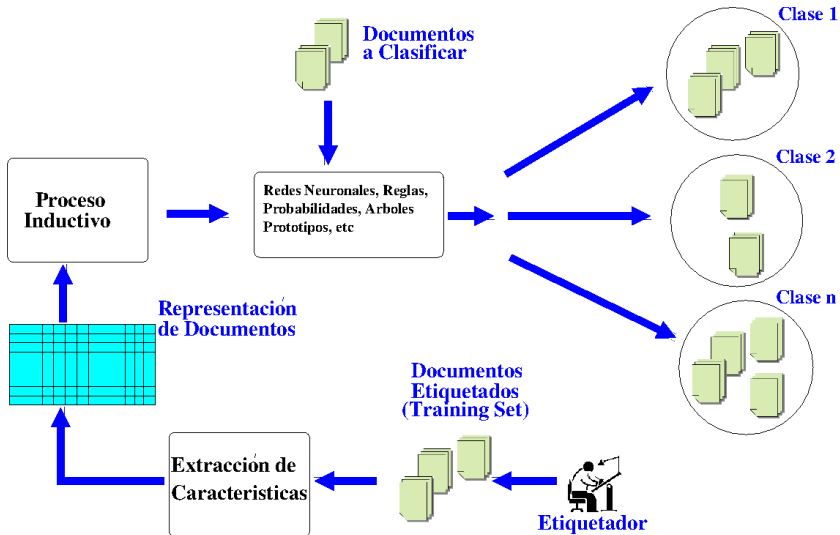
Etapas del aprendizaje (supervisado) de clasificadores



Etapas del aprendizaje (supervisado) de clasificadores

- Etiquetado
- Extracción de características
- Entrenamiento
- Uso y evaluación

Etapas del aprendizaje (supervisado) de clasificadores



Etiquetado



Etiquetado

Proceso de **asignar la clase/categoría (o valor numérico) correcto a cada documento** del conjunto de entrenamiento.

Este proceso varía en complejidad de acuerdo al tipo de valor **a predecir** y los datos disponible sobre el problema.

Dificultad del etiquetado

Ejemplos:

- Identificación de **autor**, determinación del **género** y **edad**, detección de **pedófilos** (en inglés), ratings de **opiniones** (**sencillos**, usual/ **(semi)-automático**)
- evaluación de productos (sólo comentarios, sin ratings), orientación política (**no tan sencillos**, usual/ **manual**)
- rasgos de personalidad, nivel de bienestar, estilos de aprendizaje (**complejos**, **manual** y **especializado**)

Etiquetado: “cuestionarios on-line”

Etiquetado masivo: *Mechanical Turk* de Amazon

The screenshot shows the Amazon Mechanical Turk website. At the top, there's a navigation bar with links for 'Your Account', 'HITS', and 'Qualifications'. Below this, a yellow banner states: 'Mechanical Turk is a marketplace for work. We give businesses and developers access to an on-demand, scalable workforce. Workers select from thousands of tasks and work whenever it's convenient. 249,685 HITS available. View them now.' The main content area is split into two columns. The left column, titled 'Make Money by working on HITS', describes HITS as individual tasks and lists benefits for workers: 'Can work from home', 'Choose your own work hours', and 'Get paid for doing good work'. It includes a flow diagram: 'Find an interesting task' (with a list of task examples like 'transcribe audio', 'verify image quality', etc.) -> 'Work' (with a gear icon) -> 'Earn money' (with a dollar sign icon). A 'Find HITS Now' button is at the bottom of this section. The right column, titled 'Get Results from Mechanical Turk Workers', describes the process for requesters: 'Ask workers to complete HITS - Human Intelligence Tasks - and get results using Mechanical Turk. Get Started.' It lists benefits for requesters: 'Have access to a global, on-demand, 24 x 7 workforce', 'Get thousands of HITS completed in minutes', and 'Pay only when you're satisfied with the results'. It includes a flow diagram: 'Fund your account' (with a plus icon) -> 'Load your tasks' (with a list icon) -> 'Get results' (with a star icon). A 'Get Started' button is at the bottom of this section. The footer contains links for 'FAQ', 'Contact Us', 'Careers at Mechanical Turk', 'Developers', 'Press', 'Policies', 'Blog', and 'Service Health Dashboard'. The system clock in the bottom right corner shows '11:20 p.m. 10/05/2013'.

Amazon Mechanical Turk
https://www.mturk.com/mturk/welcome

amazonmechanicalturk
Artificial Intelligence

Already have an account?
Sign in as a Worker | Requester

Your Account HITS Qualifications

Introduction | Dashboard | Status | Account Settings

Mechanical Turk is a marketplace for work.
We give businesses and developers access to an on-demand, scalable workforce.
Workers select from thousands of tasks and work whenever it's convenient.
249,685 HITS available. [View them now.](#)

Make Money by working on HITS
HITS - Human Intelligence Tasks - are individual tasks that you work on. [Find HITS now.](#)
As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work

[Find HITS Now](#)
 or [learn more about being a Worker](#)

Get Results from Mechanical Turk Workers
Ask workers to complete HITS - Human Intelligence Tasks - and get results using Mechanical Turk. [Get Started.](#)
As a Mechanical Turk Requester you:

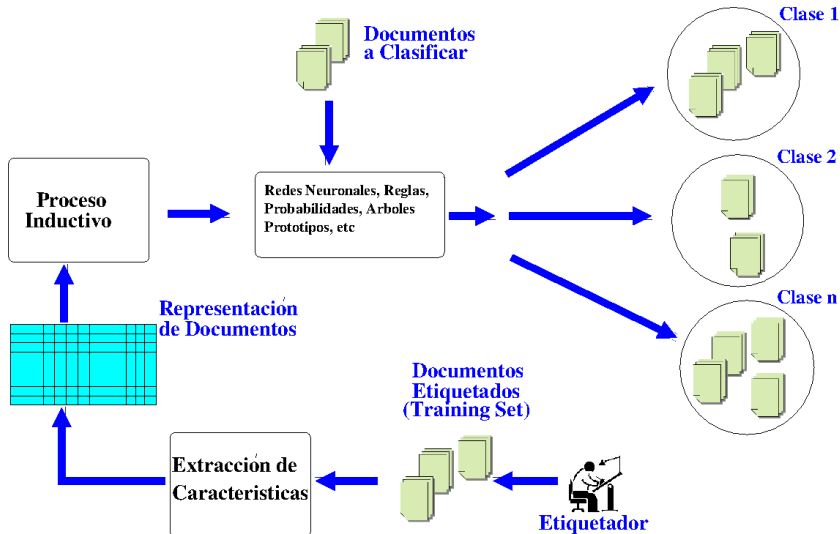
- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITS completed in minutes
- Pay only when you're satisfied with the results

[Get Started](#)

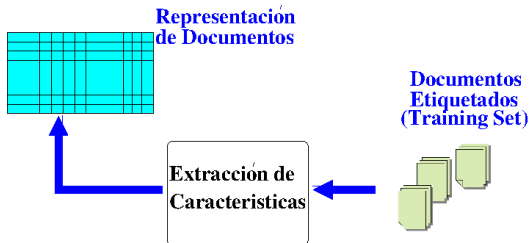
FAQ | Contact Us | Careers at Mechanical Turk | Developers | Press | Policies | Blog | Service Health Dashboard

11:20 p.m. 10/05/2013

Etapas del aprendizaje (supervisado) de clasificadores



Extracción de características



Extracción de características

Extracción de características

Etaapa encargada de tomar los documentos/textos **crudos** y generar una **representación** adecuada para el módulo de análisis (aprendizaje).

Sub-etapas

Surgen **como parte** del procesamiento de los documentos.

- Pre-procesamiento
- Representación/ponderación (indexado)
- Reducción de dimensionalidad

Características estáticas vs dinámicas (Layton)

Características *estáticas*

Se eligen **antes** del procesamiento de los documentos.

- basadas en **caracteres**
- basadas en **palabras**
- **sintácticas**
- **estructurales**
- específicas del **contenido**

Características *dinámicas* (variables)

Surgen **como parte** del procesamiento de los documentos.

- el modelo Bag of Words (BOW)
- *n*-gramas de palabras
- *n*-gramas de caracteres

Pre-procesamiento

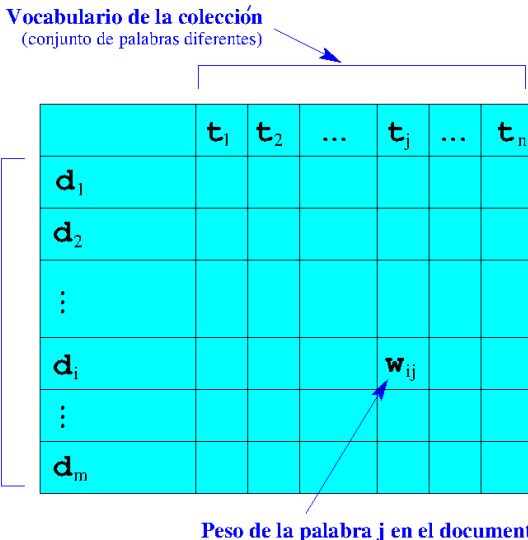
Algunas técnicas usuales:

- 1 **Partición** y eliminación de “**palabras-vacías**”
- 2 Truncado y lematización
- 3 Etiquetado de **partes de la oración** (Part of Speech (POS) Tagging)
- 4 Análisis sintáctico (**parsing**)
- 5 Desambiguación del Significado de las Palabras
- 6 Extracción de n-gramas
- 7 Reconocimiento de Entidades Nombradas

Representación vectorial de documentos: visión general

Vocabulario de la colección

(conjunto de palabras diferentes)



	t_1	t_2	...	t_j	...	t_n
d_1						
d_2						
\vdots						
d_i				w_{ij}		
\vdots						
d_m						

Todos los documentos
(un vector por documento)

Peso de la palabra j en el documento i

Reducción de Dimensionalidad

Selección de características

Se elige un **subconjunto** de las características más **informativas**.

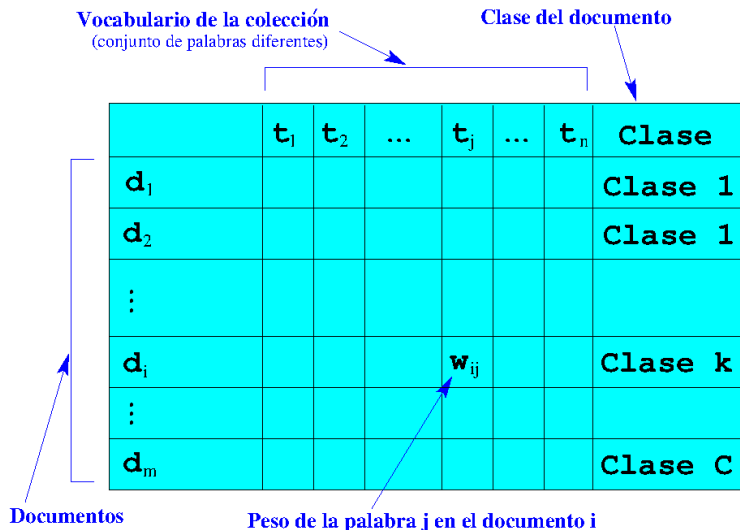
- Umbral de la frecuencia de documento
- Ganancia de Información

Transformación del espacio de características

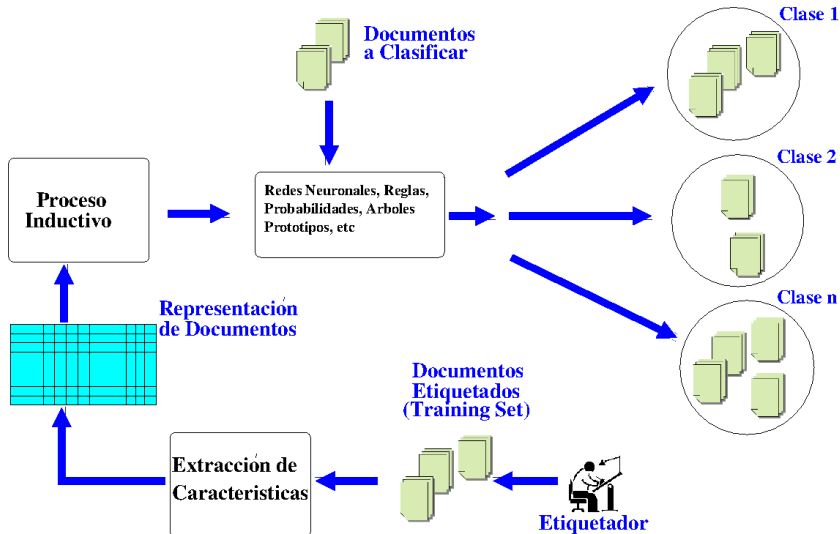
Se obtiene un **nuevo** conjunto de características, de menor dimensionalidad.

- Indexado de semántica latente (en inglés, **LSI**)
- Agrupamiento de características (clustering)

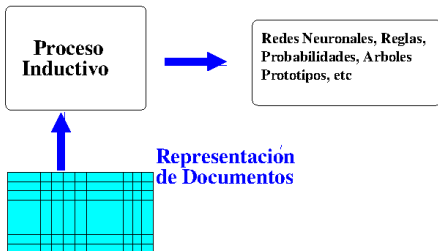
Representación “Bolsa de Palabras” (con la clase)



Etapas del aprendizaje (supervisado) de clasificadores



Aprendizaje automático

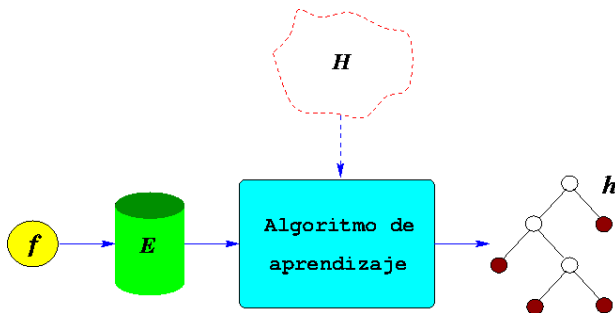


Aprendizaje de un clasificador

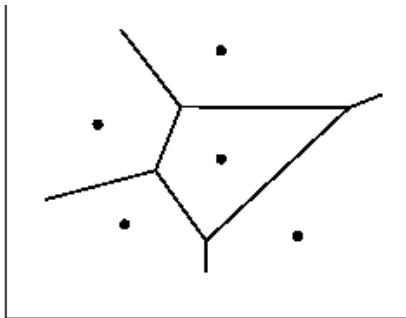
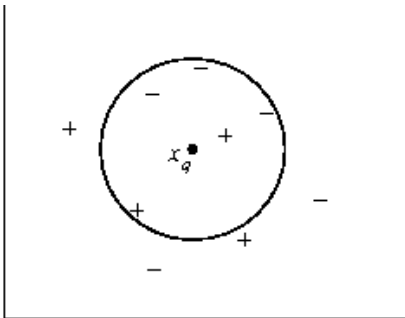
Idea: aproximar la función **ideal** de clasificación:

$$f : \mathcal{D} \mapsto \mathcal{C}$$

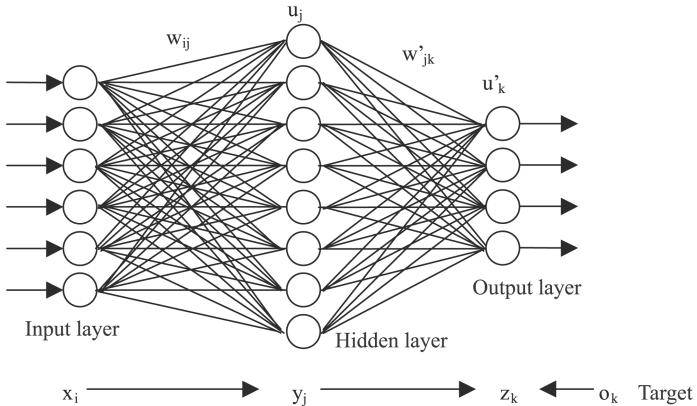
con un conjunto de entrenamiento E , de ejemplos $\langle \vec{x}, f(\vec{x}) \rangle$



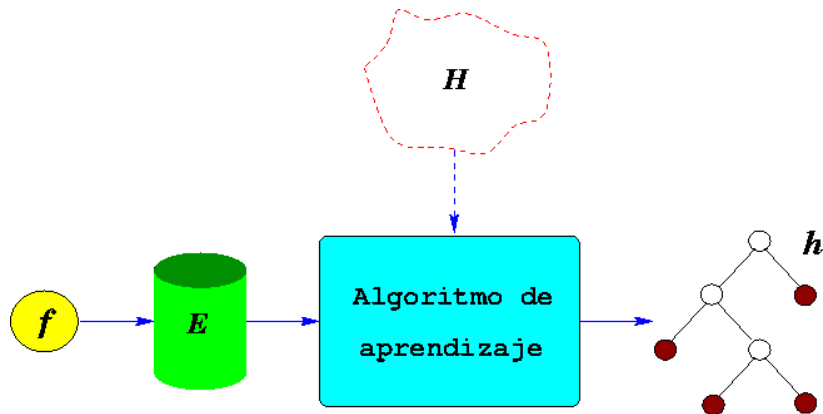
Un clasificador muy simple: k -NN

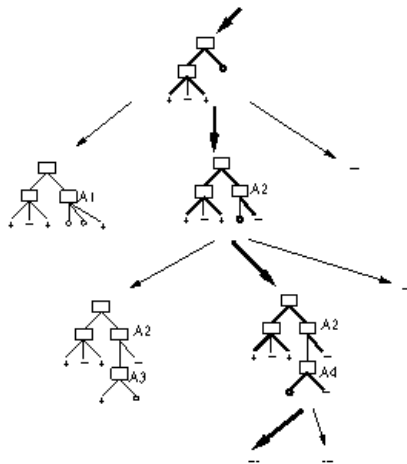


Otro clasificador muy usado: redes neuronales (NN)

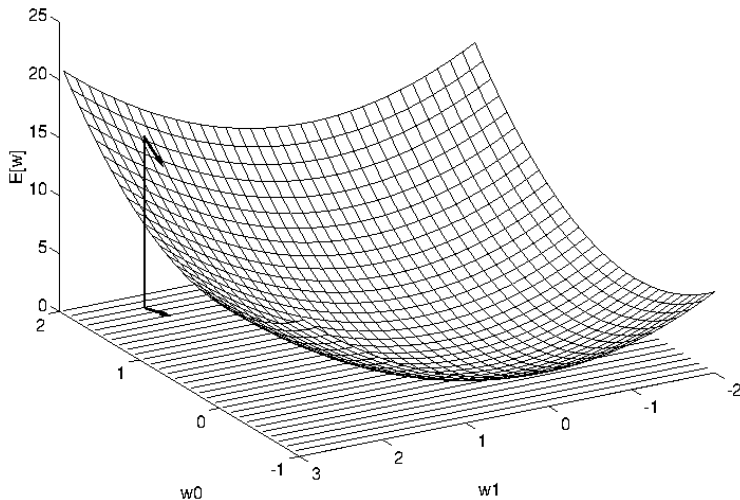


Aprendizaje de una hipótesis: esquema general

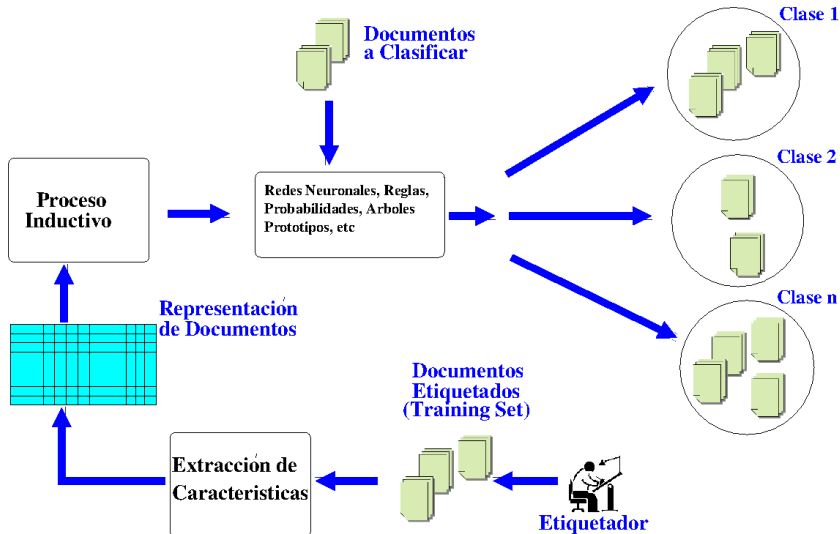




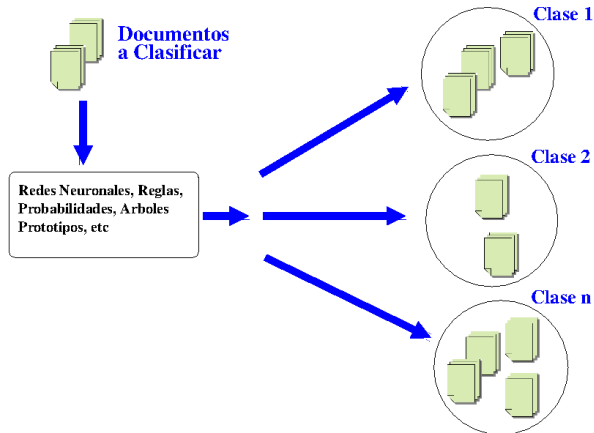
Búsqueda en el espacio de hipótesis en NN



Etapas del aprendizaje (supervisado) de clasificadores



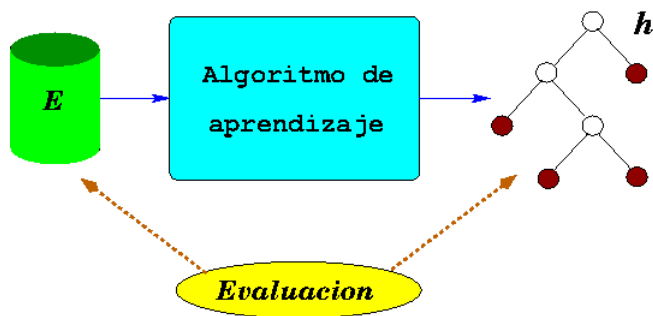
Evaluación y uso



Algunas alternativas para evaluar una hipótesis

- El conjunto E se usa para entrenamiento y evaluación
- Separar la evidencia en un **conjunto de entrenamiento** y un **conjunto de test (prueba)**.
- Validación cruzada

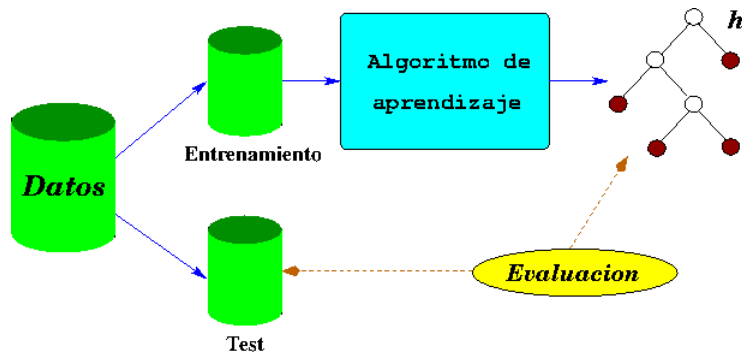
Entrenamiento y evaluación sobre el mismo conjunto



Problemas:

- sobreajuste (**overfitting**)
- subajuste (**underfitting**)

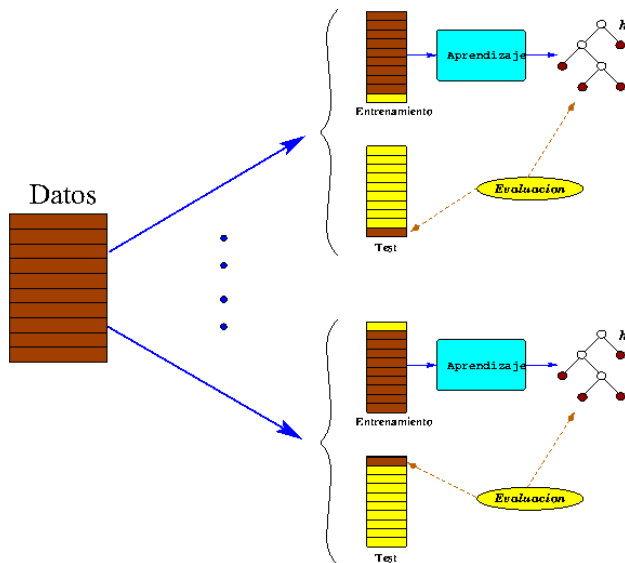
Entrenamiento y evaluación sobre conjuntos separados



Permite detectar el sobreajuste cuando la hipótesis arroja resultados mucho mejores para el conjunto de entrenamiento que el de test. **Problemas:**

- Resultados muy dependientes de la partición
- Escasez de datos

Evaluación mediante validación cruzada (*cross validation*)



Medidas de evaluación de clasificadores

Un método usual para medir las bondades de un clasificador, es considerar la **exactitud (accuracy)** del modelo, que mide esencialmente el **porcentaje de aciertos** de la hipótesis aprendida.

Esta medida se obtiene fácilmente a partir de la **matriz de confusión**.

Si se deben categorizar textos en n clases, corresponderá una matriz de confusión M de $n \times n$.

Matriz de confusión

Cada componente $M_{i,j}$ es el número de casos en que la hipótesis h predijo el valor i y el valor real era j .

Ejemplo: Identificación de Autoría

<i>Estimado ($h(x)$)</i>	<i>Real ($f(x)$)</i>			
		Borges	Cortázar	Arlt
	Borges	71	3	1
	Cortázar	8	7	1
	Arlt	4	2	3

La exactitud se calcula dividiendo el número de casos en la diagonal (**aciertos**) por el número total de casos testeados:

$$acc_T(h) = \frac{71 + 7 + 3}{71 + 3 + 1 + 8 + 7 + 1 + 4 + 2 + 3} = \frac{81}{100} = 0,81$$

Otras medidas de evaluación

Precisión (precision) y alcance (recall)

<i>Estimado ($h(x)$)</i>	<i>Real ($f(x)$)</i>			
		Borges	Cortázar	Arlt
	Borges	71	3	1
	Cortázar	8	7	1
	Arlt	4	2	3

$$\pi_{Borges} = \frac{71}{71 + 3 + 1} = 0,947$$

$$\rho_{Borges} = \frac{71}{71 + 8 + 4} = 0,855$$

Combinando π y ρ

- Rara vez precision y recall son consideradas en forma aislada
- Alternativas: medidas combinadas como la “ F -measure” (medida F):

$$F = \frac{2\pi\rho}{\pi + \rho}$$

- La medida previa es un caso particular (F_1) de la función F_β :

$$F_\beta = \frac{(\beta^2 + 1)\pi\rho}{\beta^2\pi + \rho}$$

para algún $0 \leq \beta \leq +\infty$

- Usualmente $\beta = 1$ (igual peso a π y ρ)

Entrenamiento y evaluación de un clasificador

A continuación de esta teoría, se verá de que manera:

- **Cargar** un conjunto de datos **etiquetado** en scikit-learn
- **Entrenar** uno o más clasificadores mediante distintos métodos de aprendizaje (SVM, Bayes “Ingenuo”, etc)
- **Evaluar** los resultados obtenidos