


Clase 4: Categorización de textos

Enfoques Clásicos y Neuronales a la Minería de Texto

Marcelo Errecalde^{1,2}

¹Universidad Nacional de San Luis, Argentina 

²Universidad Nacional de la Patagonia Austral, Argentina 



Resumen

1 Categorización de textos

2 Etapas del aprendizaje (supervisado) de clasificadores

- Etiquetado
- Extracción de características
- Entrenamiento (aprendizaje automático)
- Evaluación de un clasificador

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

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

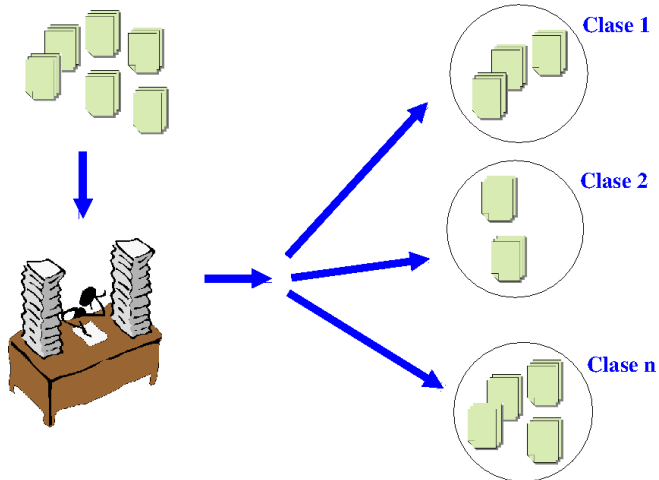
Ejemplos:

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}

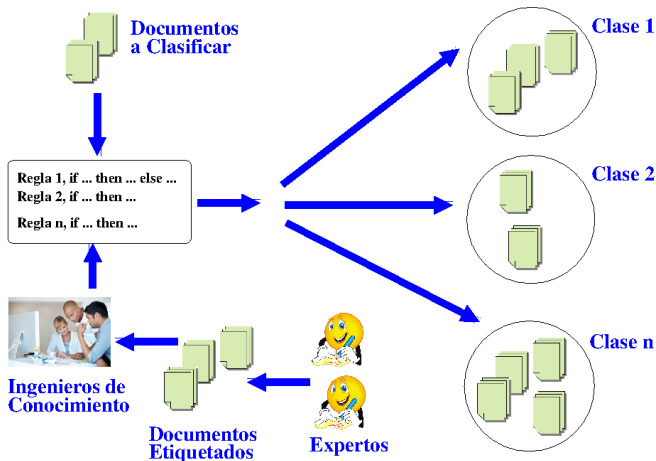
Enfoques para la clasificación de textos

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

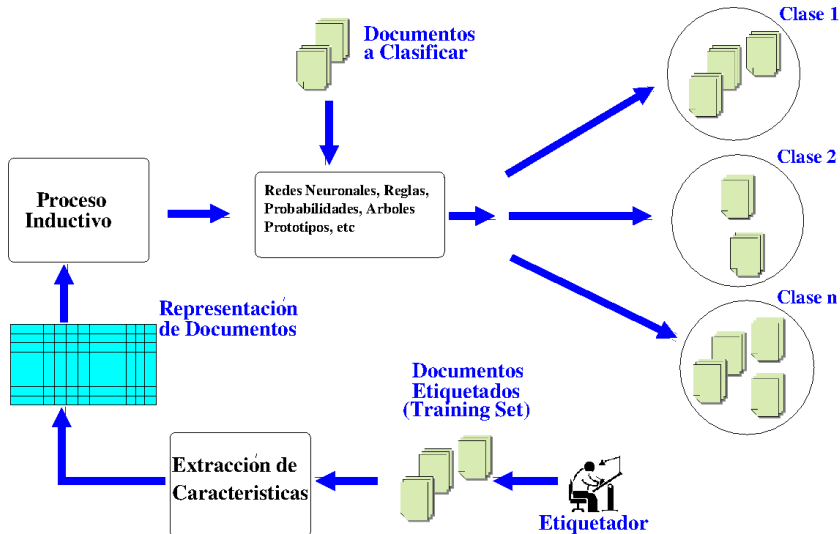
Clasificación Manual



Clasificación basada en reglas (manualmente codificadas)

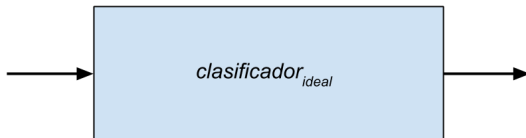


Sistemas de aprendizaje automático



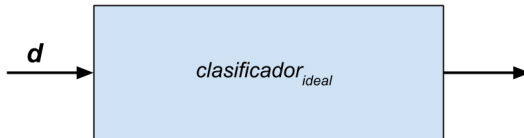
Aprendizaje automático

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



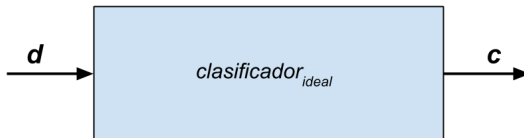
Aprendizaje automático

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



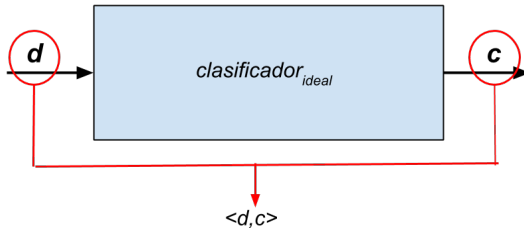
Aprendizaje automático

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



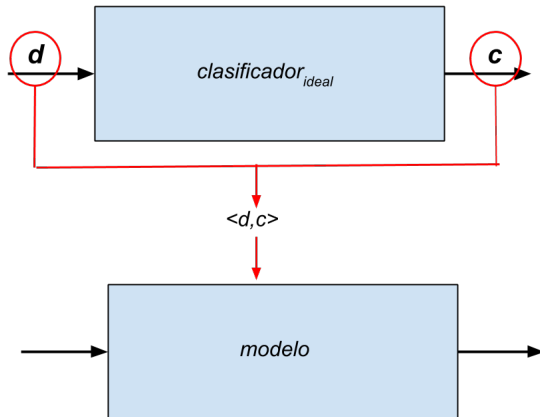
Aprendizaje automático

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



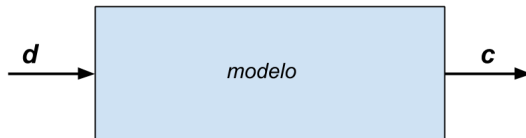
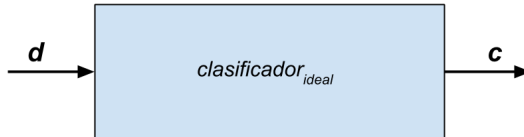
Aprendizaje automático

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**:

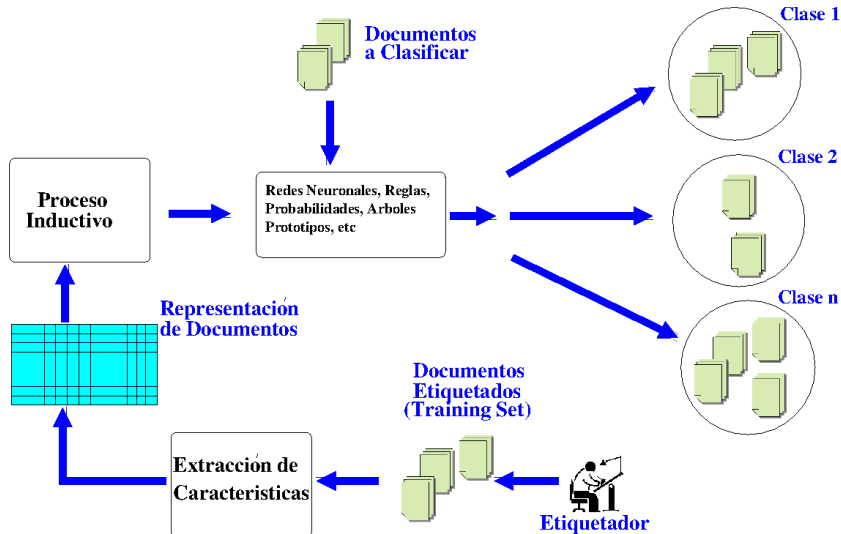
- las salidas (clasificaciones) de *clasificador_{ideal}* y *modelo* deberían coincidir respecto a los ejemplos de entrenamiento pero (y más importante),

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**

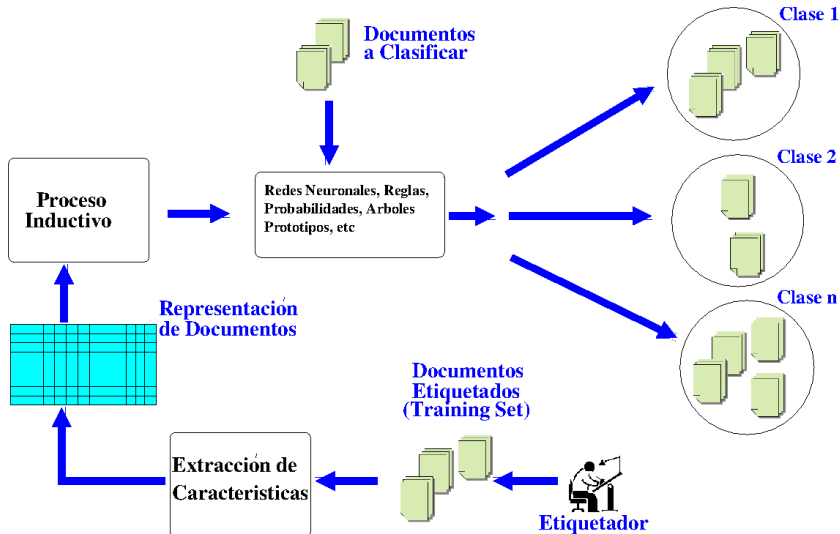
Etapas del aprendizaje (supervisado) de clasificadores



Etapas del aprendizaje (supervisado) de clasificadores

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

Etapas del aprendizaje (supervisado) de clasificadores



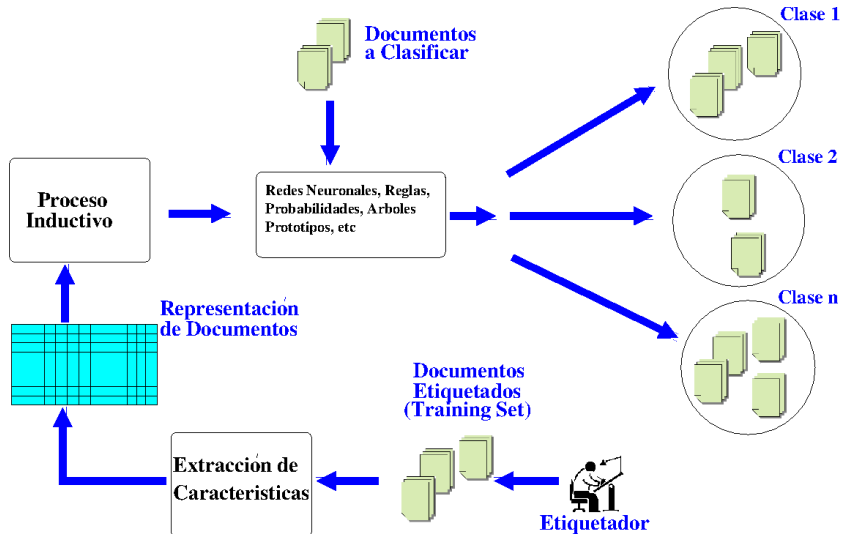
Etiquetado

**Documentos
Etiquetados
(Training Set)**

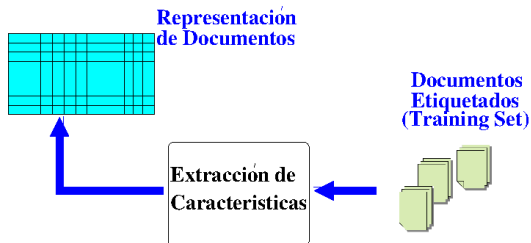


Etiquetador

Etapas del aprendizaje (supervisado) de clasificadores



Extracción de características



Representación vectorial de documentos: visión general

Vocabulario de la colección

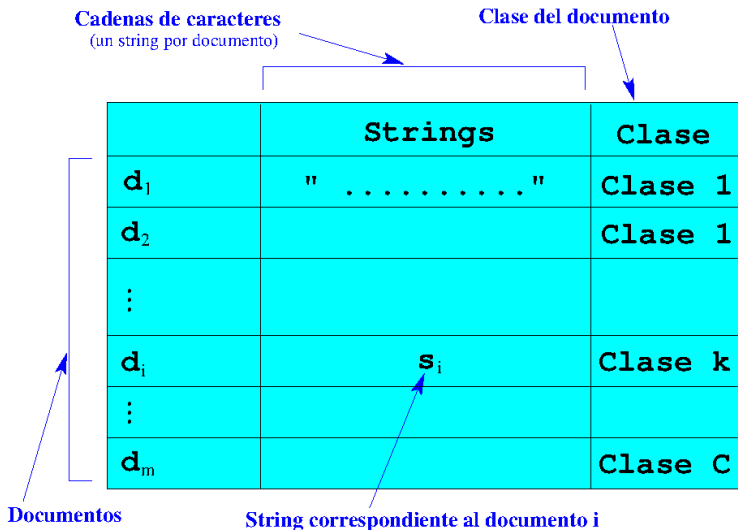
(conjunto de palabras diferentes)

	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

Lista de strings (con la clase)



Representación “Bolsa de Palabras” (con la clase)

Vocabulario de la colección
(conjunto de palabras diferentes)

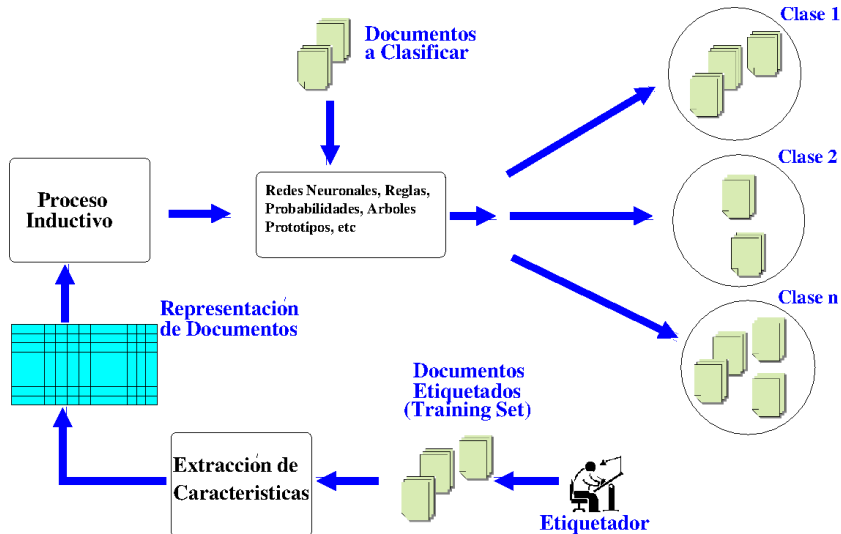
Clase del documento

	t_1	t_2	...	t_j	...	t_n	Clase
d_1							Clase 1
d_2							Clase 1
\vdots							
d_i				w_{ij}			Clase k
\vdots							
d_m							Clase C

Documentos

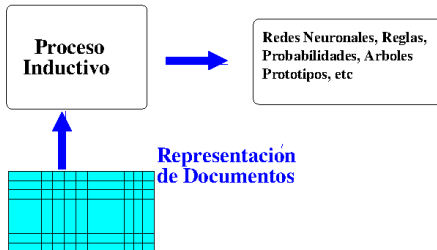
Peso de la palabra j en el documento i

Etapas del aprendizaje (supervisado) de clasificadores



Entrenamiento (aprendizaje automático)

Aprendizaje automático

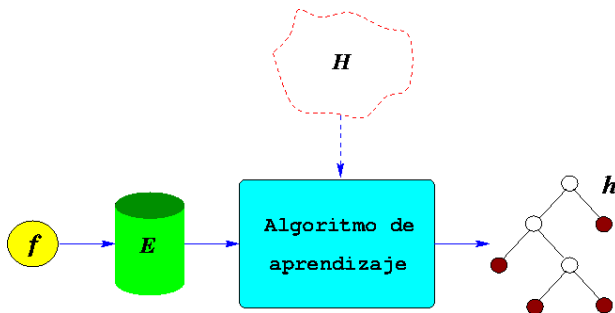


Aprendizaje de un clasificador

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

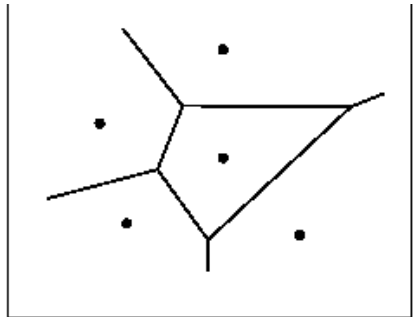
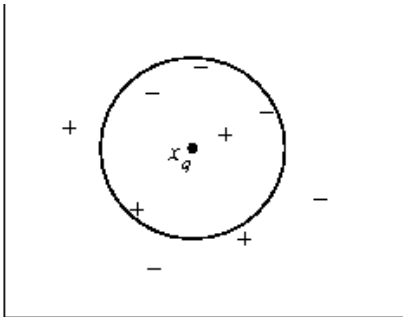
$$f : \mathcal{D} \mapsto \mathcal{C}$$

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



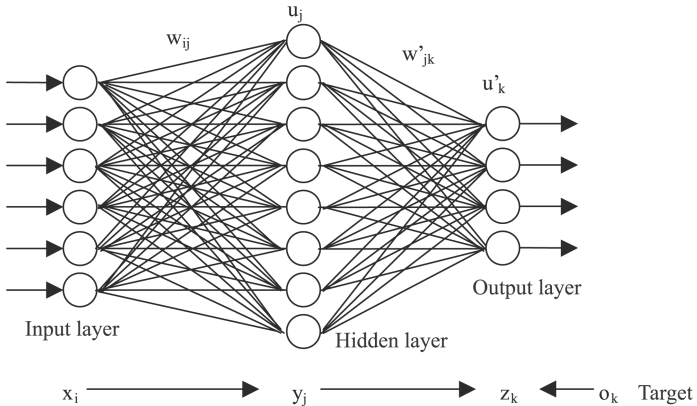
Entrenamiento (aprendizaje automático)

Un clasificador muy simple: k -NN

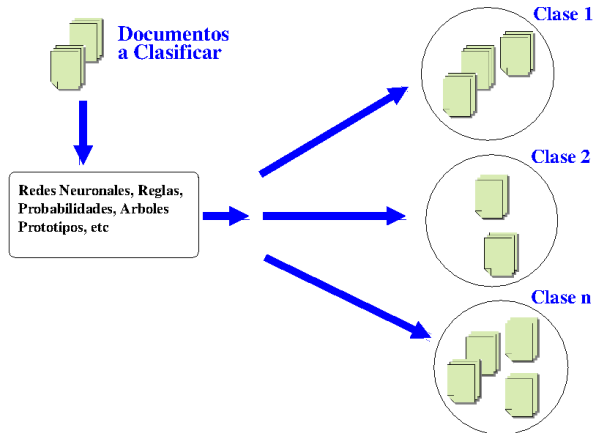


Entrenamiento (aprendizaje automático)

Otro clasificador muy usado: redes neuronales (NN)



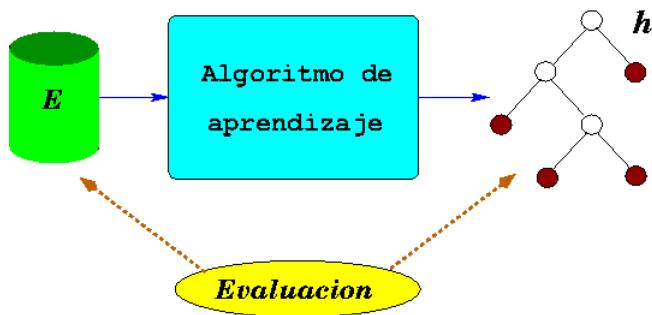
Evaluación y uso



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

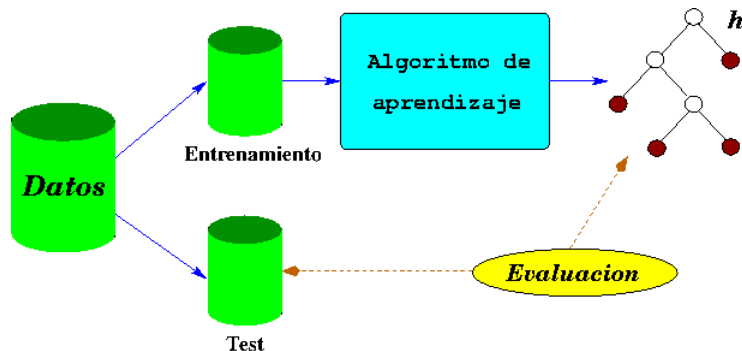
Entrenamiento y evaluación sobre el mismo conjunto



Problemas:

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

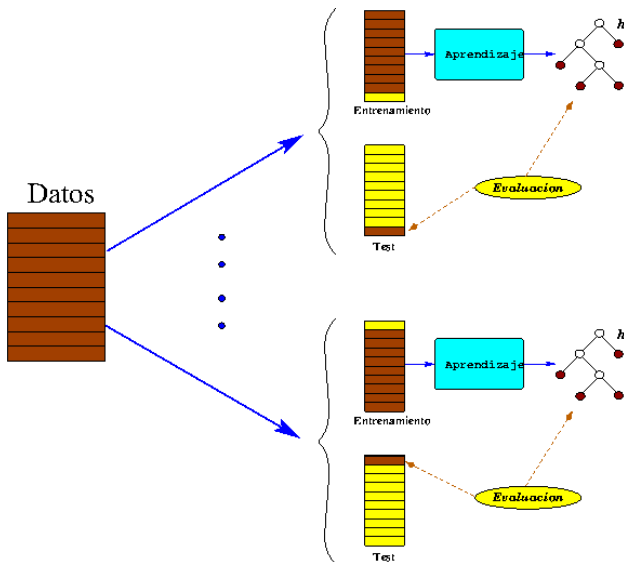
Entrenamiento y evaluación sobre conjuntos separados



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

Combinando π y ρ

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

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

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 $0 \leq \beta \leq +\infty$

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, se verá de que manera:

- **Cargar** un conjunto de datos **etiquetado** en scikit-learn

Entrenamiento y evaluación de un clasificador

A continuación, 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)

Entrenamiento y evaluación de un clasificador

A continuación, 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