

# Experimentación y métricas de evaluación

Francisco Gómez Fernández (Pachi)

Métodos Numéricos  
Departamento de Computación  
Facultad de Ciencias Exactas y Naturales  
Universidad de Buenos Aires



**DEPARTAMENTO  
DE COMPUTACION**

---

Facultad de Ciencias Exactas y Naturales - UBA

# Clase de hoy

- ▶ Métricas de evaluación: *Precision/Recall* y *Accuracy*
- ▶ *Cross-validation* y *K-Fold cross-validation*.
- ▶ Problema a analizar: “Clasificación de noticias”
- ▶ Experimentación: ¿Qué experimentar y cómo?
- ▶ Variantes para mostrar resultados
- ▶ Consultas de TP2

# Motivación: detección de caras

- ▶ Objetivo: dada una imagen  $I$  decidir si contiene una cara o no
- ▶ Problema de clasificación binaria: 1 = es cara, 0 = no es cara
- ▶ Se quiere obtener un clasificador  $clf$  que puede verse como una función:  $clf : \mathbb{R} \times \mathbb{R} \rightarrow \{0, 1\}$

Ejemplo: se tienen tres imágenes  $I_1, I_2$  e  $I_3$



$$clf(I_1) = 1$$



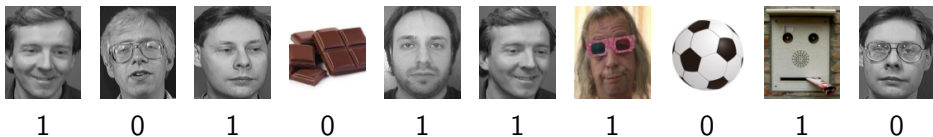
$$clf(I_2) = 0$$



$$clf(I_3) = 0$$

# Motivación: detección de caras

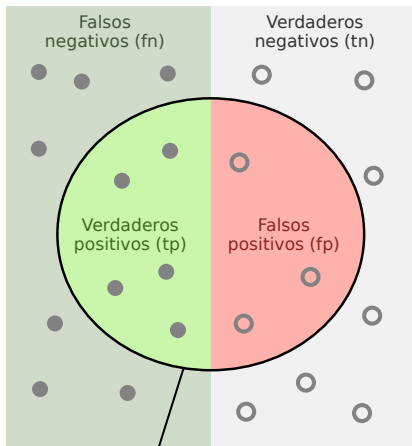
Ahora evalúo mi clasificador en 10 imágenes distintas.



- ▶ ¿Qué desempeño obtuvo mi clasificador?
- ▶ ¿Cómo sé si mi clasificador funciona bien?
  - ▶ ¿Qué significa que funcionó bien o mal?
- ▶ ¿Cómo mido el desempeño?
- ▶ Necesito definir alguna *métrica*
- ▶ ¿En qué conjunto evalúo mi métrica?

# Precision y Recall para clasificación binaria

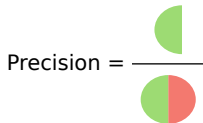
Elementos relevantes



Elementos recuperados

		Verdad	
		Si	No
Predicción	Si	tp	fp
	No	fn	tn

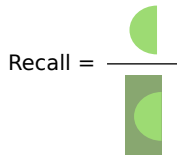
¿Cuántos de los elementos recuperados son **relevantes**?



$$\text{Precision} = \frac{\text{Verdaderos positivos}}{\text{Verdaderos positivos} + \text{Falsos positivos}}$$

$$\text{Precision} = \frac{tp}{tp+fp}$$

¿Cuántos elementos **relevantes** fueron recuperados?

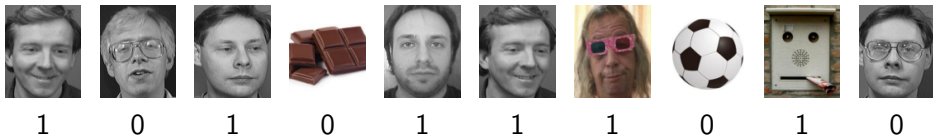


$$\text{Recall} = \frac{\text{Verdaderos positivos}}{\text{Verdaderos positivos} + \text{Falsos negativos}}$$

$$\text{Recall} = \frac{tp}{tp+fn}$$

# Motivación: detección de caras

Ahora calculemos estas métricas para el ejemplo anterior.



►  $tp = 5$        $fp = 1$        $tn = 2$        $fn = 2$

►  $Precision = \frac{tp}{tp + fp} = \frac{5}{6} = 0,83$

“De los *recuperados*, qué porcentaje son *relevantes*”

►  $Recall = \frac{tp}{tp + fn} = \frac{5}{7} = 0,71$

“De los *relevantes*, qué porcentaje son *recuperados*”

- ¿Qué significa un valor de 1 en *precision* o *recall*?
  - Sistemas robustos: alto porcentaje de recall (o sensibilidad)
  - Sistemas precisos: alto porcentaje de precisión
- ¿Se puede prescindir de una o de la otra?

# Más métricas

## F-measures: métricas combinadas de Precision y Recall

- ▶ Media armónica:  $F_1 = 2 \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$
- ▶ Fórmula general:  $F_\beta = (1 + \beta^2) \frac{\text{precision} \times \text{recall}}{\beta^2 \text{precision} + \text{recall}}$
- ▶  $F_2$  enfatiza recall mientras que  $F_{0,5}$  enfatiza precision

Esta métrica sirve para establecer un compromiso entre *precision* y *recall*. Precision y Recall son dos medidas importantes que no necesariamente tienen la misma calidad para un mismo clasificador,

## Tasa de eficacia o exactitud

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

Mide el porcentaje de muestras bien clasificadas sobre el total.

- ▶ A favor: es fácil de entender y reportar
- ▶ En contra: puede ser engañosa. Ej: un 95 % parece muy bueno pero ¿y si hay 2 clases y el 98 % del total pertenece a una?

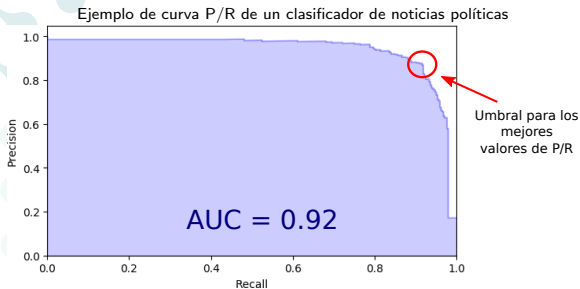
# Curvas de Precision/Recall (P/R) <sup>1</sup>

Muestra la relación entre ambas métricas para una clase particular.

- ▶ El clasificador, además del resultado de la clasificación, debe devolver un valor de confianza o probabilidad en la clasificación.

## ¿Cómo encontramos el mejor umbral de decisión?

- ▶ Se calculan pares de P/R para distintos umbrales de decisión
- ▶ Un área bajo la curva (AUC) grande representa altos valores de P/R
- ▶ Un sistema ideal con alto P/R devolverá muchos resultados y todos bien clasificados



<sup>1</sup>¡Ojo! no confundir con las curvas ROC



## Otras métricas (que no usaremos en el TP)

► True Positive Rate (TPR) =  $\frac{tp}{tp + fn} = \frac{tp}{P}$

También llamado *sensitivity* o *recall*.

Ej.: Porcentaje de pacientes *enfermos* correctamente diagnosticados.

► False Negative Rate (FNR) =  $\frac{fn}{tp + fn} = \frac{fn}{P}$

También llamado *miss rate*.

Ej.: Porcentaje de pacientes *enfermos* incorrectamente diagnosticados.

► True Negative Rate (TNR) =  $\frac{tn}{tn + fp} = \frac{tn}{N}$

También llamado *specificity*.

Ej.: Porcentaje de pacientes *sanos* correctamente diagnosticados.

► False Positive Rate (FPR) =  $\frac{fp}{tn + fp} = \frac{fp}{N}$

También llamado *porcentaje de falsas alarmas*.

Ej.: Porcentaje de pacientes *sanos* incorrectamente diagnosticados.

$P = tp + fn$  (tamaño del conjunto de relevantes o casos positivos)

$N = fp + tn$  (tamaño del conjunto de irrelevantes o casos negativos)

## Motivación 2: clasificación de noticias

- ▶ Objetivo: dada una noticia  $Q$  se quiere decidir a qué tópico (entre  $N$  posibles) pertenece
- ▶ Problema de clasificación multiclase: 1 = deportes, 2 = cultura, ... ,  $N$  = política

### Procedimiento:

#### 1. Ante una nueva noticia (query)

id	section	topic	text	title
643	Procesados	????	"El presidente Mauricio ..."	"Panama Papers: Macri compli ..."

2. Procesar el texto (contenido) de la noticia
3. Comparar con noticias anteriores (clasificador ya entrenado)
4. Asignar una categoría (número entre 1 y  $N$ ) de acuerdo al resultado del clasificador

- ▶ ¿Como mido el desempeño de mi clasificador?
- ▶ ¿Me sirven las métricas anteriores?
- ▶ ¿En qué conjunto evalúo mi métrica?

# Precision y Recall para clasificación multiclase

Dada una clase  $i = 1 \dots N$ , se calcula para cada una:  $tp_i$ ,  $fp_i$ ,  $tn_i$  y  $fn_i$  de forma análoga al caso binario.

- ▶  $tp_i$  son las muestras que realmente pertenecían a la clase  $i$  y fueron exitosamente identificadas como tales.
- ▶  $fp_i$  son aquellas muestras que fueron identificadas como pertenecientes a la clase  $i$  cuando realmente no lo eran.
- ▶  $precision_i = \frac{tp_i}{tp_i + fp_i}$        $recall_i = \frac{tp_i}{tp_i + fn_i}$

# Precision y Recall para clasificación multiclase

- ▶ La *precision* en el caso de un clasificador multiclase, se define como el **promedio** de las *precision* para cada una de las clases.
- ▶ El *recall* en el caso de un clasificador multiclase, se define como el **promedio** del *recall* para cada una de las clases.
- ▶ El Accuracy mide el porcentaje de muestras bien clasificadas sobre el total

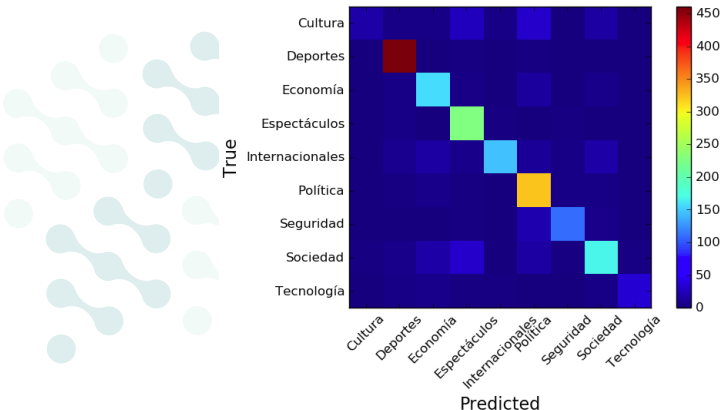
$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

- ▶ ¿Está bien promediar estos valores?
- ▶ Se suelen reportar por *clase*. Más si están desbalanceadas.

# Matriz de confusión

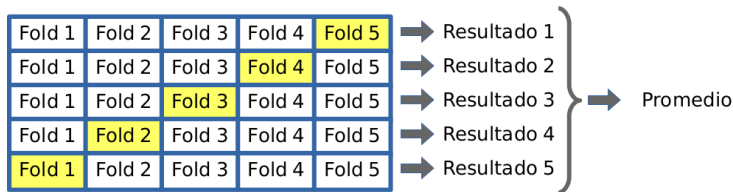
- ▶ La matriz de confusión muestra para cada par de clases  $c_i \neq c_j$ , cuántos documentos de  $c_i$  se asignaron incorrectamente a  $c_j$  o correctamente entre  $c_i$  y  $c_j$ .
- ▶ La matriz de confusión puede ayudar a identificar dónde se debe mejorar la precisión del sistema.

Ejemplo de clasificación de noticias:



# Validación y cross-validation

- ▶ Evaluar el modelo en los datos de entrenamiento puede darnos una impresión errónea.
- ▶ Separa los datos al azar para evitar tomar patrones en las divisiones en dos partes. Ejemplo:
  - ▶ Entrenamiento  $(100 - p)\%$       Validación  $p\%$  (con  $p = 20\%$ )
- ▶ ¿Y si al azar no funciona tan bien? → k-Fold Cross Validation
  1. Desordenar los datos
  2. Separar en  $K$  folds del mismo tamaño
  3. Para  $i = 1 \dots K$ :  
Entrenar sobre todos los folds menos el  $i$  y validar sobre el  $i$



2

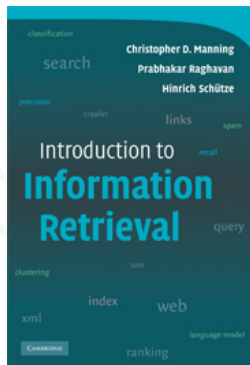
<sup>2</sup>Diapo adaptada de la clase de Aprendizaje Automático.

# Lectura recomendada

## An Introduction to Information Retrieval

Manning, Raghavan y Schütze. Año 2009.

Disponible online: <http://www.informationretrieval.org/>



- ▶ Capítulo 8: “*Evaluation in information retrieval*”.
- ▶ Capítulo 14.3: “*k nearest neighbor*”.
- ▶ Capítulo 14.5: “*Classification with more than two classes*”.

# El problema: motivación



canchallena.com > Tenis > Sebastián Torok > US Open

US OPEN

Lunes 14 de septiembre de 2015 | 07:17

## No es una utopía: Djokovic puede alcanzar el récord de Grand Slam

Llegó a 10 títulos de los Grandes, siete menos que los que reúne su vencido, Roger Federer; sin embargo, acumula méritos propios para aventurar que podría alcanzar semejante hito

Por [Sebastián Torok](#) | canchallena.com

### Objetivo

En función del contenido (texto) de las noticias decidir qué categoría (deportes, espectáculos, tecnología, etc.) le corresponde.

### Objetivo *deseable*

Poder devolver un valor de probabilidad o confianza en la predicción/clasificación de cada noticia.



# El problema: una *propuesta* de solución

## Datos con noticias etiquetadas

id	section	topic	text	title
3066966	Deportes	Deportes	"NUEVA YORK. - Novak ..."	"No es una utopía: ..."
3065926	"El Mundo"	Internacionales	"Los bomberos tratan de ..."	"Declaran el estado de ..."
.	.	.	.	.
.	.	.	.	.
2496990	Tecnología	Tecnología	"La plataforma de video ..."	"La tiranía digital del ..."

## Datos nuevos a categorizar o clasificar

id	section	topic	text	title
6431364	Procesados	????	"El presidente Mauricio ..."	"Panama Papers: la respuesta de Macri ..."

## Clasificación supervisada

- ▶ Usar el conjunto de datos de etiquetados para entrenar un *predictor* que nos permita predecir las etiquetas de nuevos datos con etiqueta desconocida
- ▶ Se asume que los nuevos datos sobre los cuales no se entrenó el predictor, tienen el mismo origen o fueron producidos por la misma fuente (desconocida) que estamos tratando de modelar

# El problema: una propuesta de solución

## Paso previo

Utilizando el conjunto de entrenamiento, entrenar un clasificador que permita predecir el tópico de las noticias.

## Ante una nueva noticia (query)

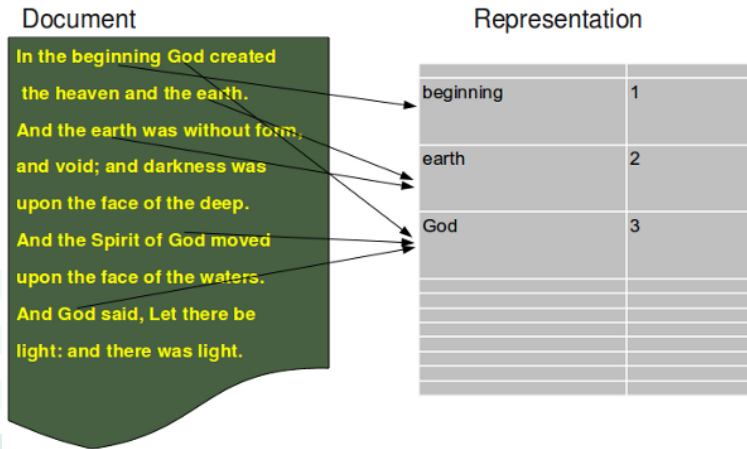
id	section	topic	text	title
6431364	Procesados	????	"El presidente Mauricio ..."	"Panama Papers: la respuesta de Macri ..."

1. Procesar el texto (contenido) de la noticia
2. Comparar con noticias anteriores (clasificador ya entrenado)
3. Asignar una categoría de acuerdo al resultado del clasificador
4. Devolver la *confianza* de la clasificación

## Métodos empleados

- ▶ Bag of words
- ▶ N-grams + Stemming
- ▶ Term frequency-inverse document frequency (tf-idf)

# Bag of words



- ¿Virtudes? ¿Problemas?

# N-grams

- ▶ La probabilidad de un “N-grama” está dada por los  $N - 1$  anteriores
- ▶ 
$$P(g_n | g_{n-1}, g_{n-2}, \dots, g_{n-N+1}) = \frac{\#(g_n, g_{n-1}, g_{n-2}, \dots, g_{n-N+1})}{\#(g_{n-1}, g_{n-2}, \dots, g_{n-N+1})}$$
- ▶ Bigramas, trigramas, etc.
- ▶ En nuestro caso, se puede hacer sobre palabras o sobre caracteres
  - ▶ Otras aplicaciones: detección de idioma (letras), cadenas de proteínas, fonemas (en el contexto de procesamiento del habla), features visuales (en el contexto del procesamiento de imágenes), etc..
- ▶ ¿Virtudes? ¿Problemas?

# Stemming


- ▶ ¿Tiene sentido considerar separadamente palabras como *investigadora*, *investigador*, *investigadoras*, *investigadores*, *investigar*, *investigaron*, *investigación*, etc.?
- ▶ Stemming: reemplazar palabras por su tema o “palabra madre” (*word stem*).  
Se utiliza antes de aplicar, por ejemplo, *bag of words*.
- ▶ ¿Virtudes? ¿Problemas?

# Tf-idf

- ▶ tf-idf: cuán importante es una palabra en un documento respecto del conjunto de todos los documentos
- ▶ Dados  $d$  un documento en  $D$  y  $t$  un término,
- ▶ term frequency:
$$\text{tf}(t) = \frac{\text{\#apariciones de } t \text{ en } d}{\text{\#términos en } d}$$
- ▶ inverse document frequency:
$$\text{idf}(t) = \log \frac{|D|}{\text{\#documentos que contienen a } t}$$
- ▶  $\text{tf-idf}(t) = \text{tf}(t) \cdot \text{idf}(t)$
- ▶ ¿Virtudes? ¿Problemas?

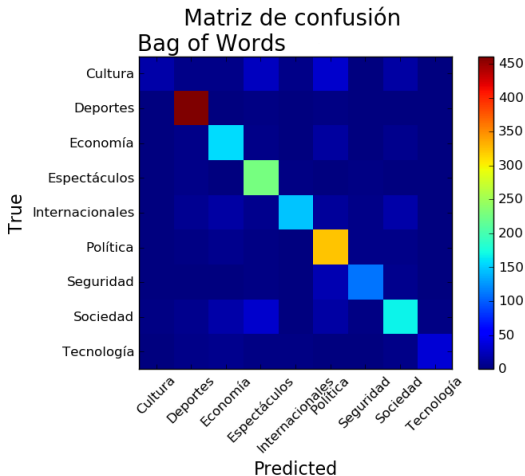
# ¿Qué experimentar y cómo?

## Ciclo de desarrollo y elaboración de experimentos

1. Entender el problema y sus **objetivos**
2. Proponer una **solución** y elaborar hipótesis o conjeturas que la demuestren, expliquen o justifiquen.
3. Visualizar los resultados preliminares.
  - ▶ ¿Qué medidas de “performance” podré usar?
  - ▶ ¿Qué es performance?
  - ▶ ¿Qué mido? ¡Qué miedo! 

# Resultados

- Corremos el clasificador sobre los datos y obtenemos 82,4 % de accuracy. Nada mal.

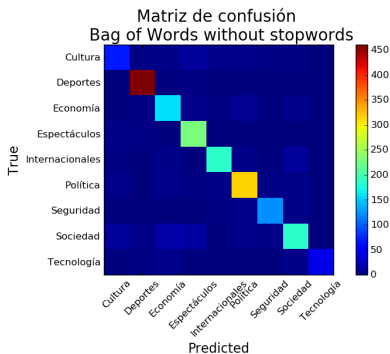
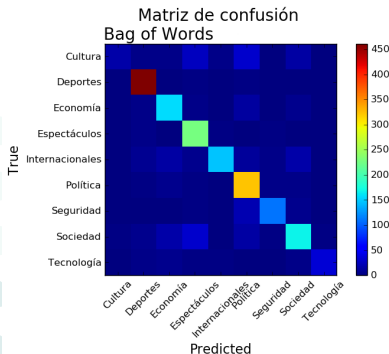


- ¿Algo interesante para destacar?



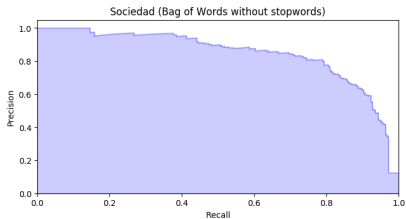
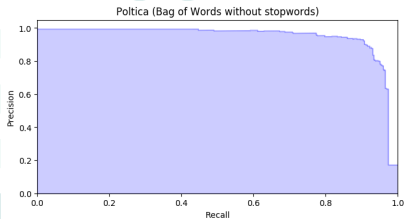
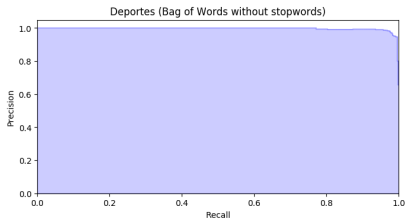
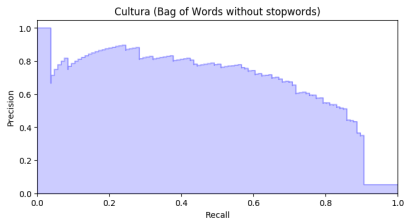
# Resultados - Problemas

- ▶ Muchas instancias de “Cultura” son clasificadas como otras categorías, principalmente como “Política”
- ▶ Algo similar sucede con “Sociedad”
- ▶ Tal vez esas categorías tienen muchas palabras en común.  
¿Y si sacamos las stopwords<sup>3</sup> del texto? Accuracy: 88,8 %



<sup>3</sup> Las palabras más usadas en un lenguaje, como por ejemplo artículos y preposiciones, no aportan información relevante al tipo de texto pero pueden influir negativamente en la clasificación.

# Resultados - Curvas P/R



► ¿Y ahora qué podemos decir?

# Resumen

- ▶ La experimentación no es sólo reportar resultados. En base a los resultados se gana entendimiento y se repiensa el problema y esto permite iterar nuevamente con experimentos.
- ▶ Es importante elegir una manera adecuada para mostrar los resultados. Ciertas características pueden quedar ocultas detrás de medidas mentirosas.
- ▶ Siempre recordando los límites en términos de tiempo que hay en los TPs.