

Experimentación y métricas de evaluación

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

Menú del día

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

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



$$clf(I_1) = 1$$



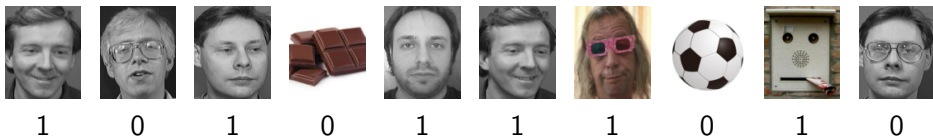
$$clf(I_2) = 0$$



$$clf(I_3) = 0$$

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?

Tasa de eficacia o exactitud

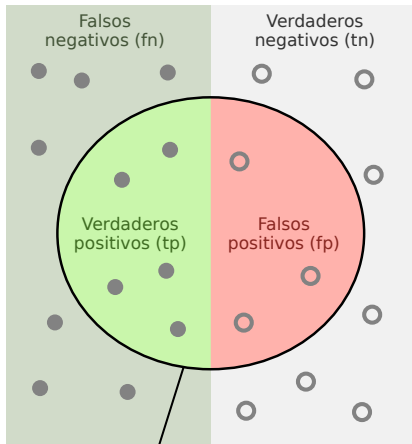
$$\text{Accuracy} = \frac{\# \text{correctos}}{\# \text{muestras}}$$

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?

Precision y Recall para clasificación binaria

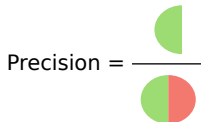
Elementos relevantes



Elementos recuperados

		Verdad	
		Si	No
Predicción	Si	tp	fp
	No	fn	tn

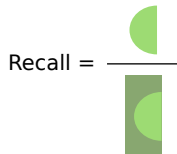
¿Cuántos de los elementos recuperados son **relevantes**?



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

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

¿Cuántos elementos **relevantes** fueron recuperados?

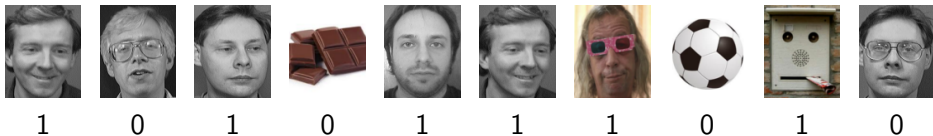


$$\text{Recall} = \frac{\text{Verdaderos positivos}}{\text{Verdaderos positivos} + \text{Falsos negativos}}$$

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

Motivación: detección de caras

Ahora calculemos estas métricas para el ejemplo anterior.



$$\text{tp} = 5 \quad \text{fp} = 1 \quad \text{tn} = 2 \quad \text{fn} = 2$$

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

“De los recuperados, qué porcentaje son relevantes”

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

“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

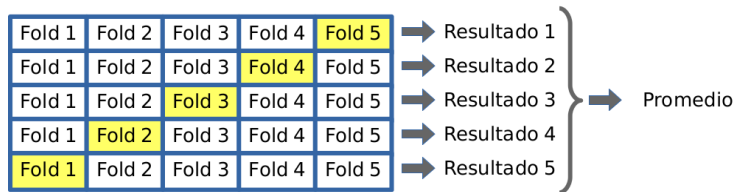
Esta métrica sirve para establecer un compromiso entre *precision* y *recall*. Precision y Recall son dos medidas importantes que no necesariamente tienen la misma calidad para un mismo clasificador.

Ejemplos de aplicaciones

- ▶ Al usar un navegador web un usuario común quiere tener un muy buen primer resultado
- ▶ Al recuperar los datos de un disco rígido podemos estar interesados en en búsquedas de alta recuperación (recall)
- ▶ Aplicaciones médicas

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



Clasificación de dígitos

Volvamos a nuestro problema de reconocer dígitos.

- ▶ ¿Como mido el desempeño de mi clasificador?
- ▶ ¿Me sirven las métricas anteriores? (Sumemos la Kappa de Cohen)
- ▶ ¿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

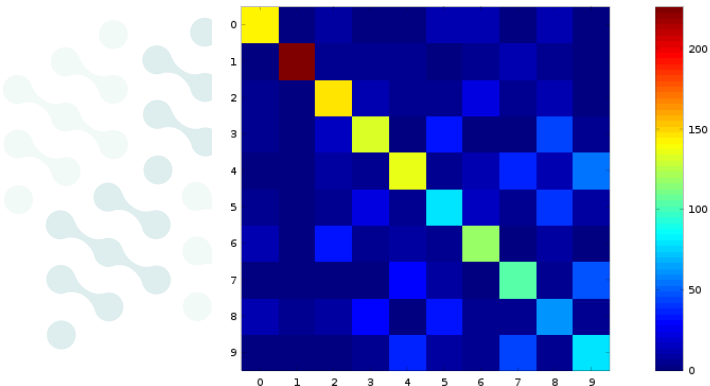
- ▶ La *precision* en el caso de un clasificador multiclase, se define como el **promedio** de las *precision* para cada una de las clases.
- ▶ El *recall* en el caso de un clasificador multiclase, se define como el **promedio** del *recall* para cada una de las clases.

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

Matriz de confusión

- ▶ Es una matriz $C \in \mathbb{R}^{p \times p}$ (p es la cantidad de clases), donde C_{ij} indica la cantidad de elementos para los que el algoritmo predijo la clase i , cuando en realidad la respuesta correcta era j .
- ▶ Es una forma de visualizar el desempeño del algoritmo. Puede ayudar a identificar dónde se debe mejorar la precisión del sistema.

Ejemplo de clasificación de dígitos:



Precision y recall en la matriz de confusión

¿Cómo puedo calcular el precision y recall de una clase particular i a partir de la matriz de confusión?

		truth _{<i>i</i>}				
pred <i>i</i>	tn	tn	tn	fn	tn	
	tn	tn	tn	fn	tn	
	tn	tn	tn	fn	tn	
	fp	fp	fp	tp	fp	
	tn	tn	tn	fn	tn	

		truth	
pred	$\mathcal{L} \setminus \{i\}$	<i>i</i>	
	$\mathcal{L} \setminus \{i\}$	tn	fn
<i>i</i>		fp	tp

\mathcal{L} es el conjunto de etiquetas posibles.

¿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?

Resultados

- Corremos el clasificador sobre los datos, usando kNN ($k = 5$) con PCA ($\alpha = 4$) y obtenemos 61 % de accuracy.

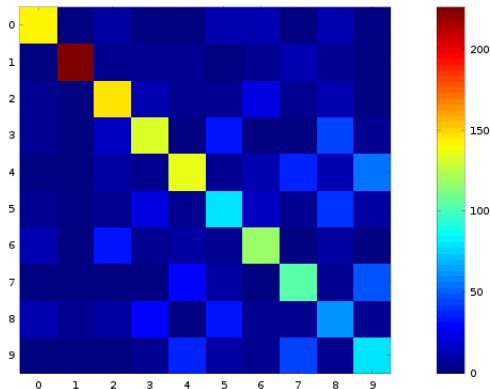
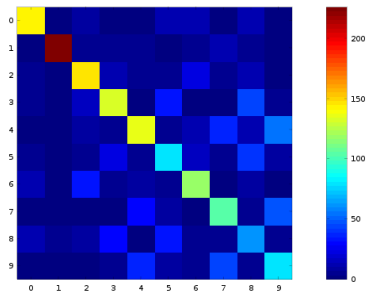


Figura: Matriz de confusión para $k = 5$, $\alpha = 4$.

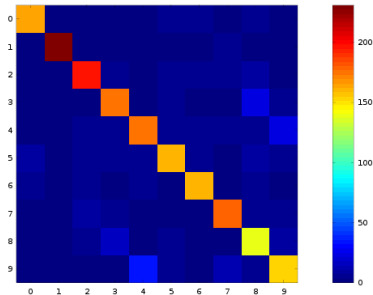
- ¿Algo interesante para destacar?

Resultados - Problemas

- ▶ Muchos nueves, cuatros y siete se confunden.
- ▶ Algo similar sucede con tres, cinco y ocho.
- ▶ Esas categorías son muy parecidas.
Aumentando el alpha de PCA debería mejorar.



(a) Matriz de confusión para $k=5$, $\alpha=4$. Accuracy = 61 %.



(b) Matriz de confusión para $k=5$, $\alpha=8$. Accuracy = 86.9 %.

Resultados - Parámetros

Vimos que los parámetros parecen ser muy influyentes en el desempeño del modelo. ¿Cómo elegimos los mejores valores?

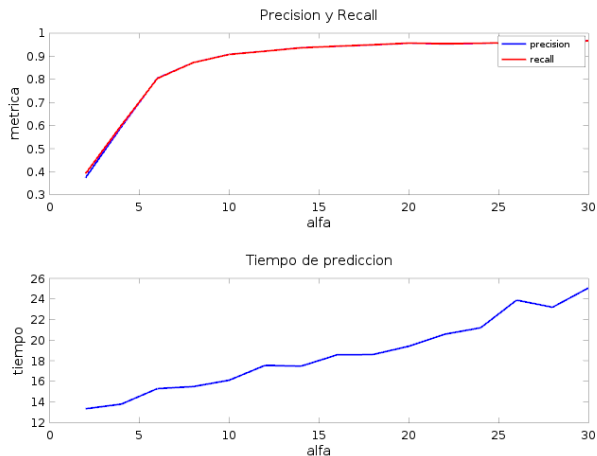


Figura: (a) Precision y Recall en función del α de PCA. (b) Tiempo de predicción para 2000 imágenes en función de α .

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.