

# Metodología para clasificación

- Tengo un problema de clasificación y quiero una función
  - ¿Cómo obtengo la función?
  - ¿Sobre qué instancias?
  - ¿Cómo medir?

## FASE 1: Preprocesamiento

Problemas:

- faltan datos
- en otros formatos
- vienen con ruido
- limpieza de datos

→ trabajar los datos de entrada para luego poderlos procesar

Suponemos

i) Conjunto de entrenamiento  $D = \{(x_i, y_i)\}$  con  $x_i \in \mathbb{R}^n$ ,  $y_i \in \mathbb{R}$

ej: TITANIC

### ① Separar conj. de entrenamiento y de evaluación

el modelo se arma  
o parte de este conj.

- Dividir el conjunto

### evaluación ( $\mathcal{T}_{\text{test}}$ )

→ solo para evaluar cuando tenemos

- lo quedo sin mirarlo,
- no se puede usar todo nada

- Necesito un conj. suficientemente grande

Caso:

$$80\% / 20\% \quad o \quad 70\% / 30\%$$

↓                    ↓  
entren.            evaluación

↓  
sirve para estimar mi performance del modelo generado

→ cuánto se equivoca mi clasificador.

→ tienen que tener similar

⇒ mezclar antes de dividir  
(train, test lo hace)

AURACIA

Entrenamiento se divide

en   
    ↑      entrenamiento ←  
    ↑      validación ← evaluación

para ajustar hiperparámetros

↑  
no sirve cuando tengo un dataset desbalanceado

\* Importante fijar la décima para mezclarlo ← nos evitamos que los resultados dependan del azar

↓  
solo al principio

} de esta forma se puede replicar

↓  
se ve a mejor de nuevo

↓  
se vuelve a mezclar

### ② Valores faltantes

¿qué hacemos con

opc A - cumplir la intención (si tengo muchos datos)

opc B - asignar un valor especial

opc C - asignar valor medio del conj. de entrenamiento

opc D - asignar según el método de aprendizaje

\* Importante entender el problema para elegir uno sol.

→ aplicar cambios en todo el dataset, usqz. cálculo de estadística aplican el conj. test

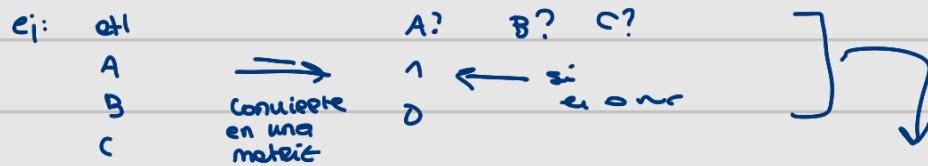
## ④ Atributos categóricos

↳ conj. disuelto, finito (<sup>concretos</sup>  
<sup>no numéricos</sup>)

opc1 - tenemos n etiquetas ..

opc2 -

Forma usual → transformar en frecuencias y saco la columna



- OneHotEncoder

## ④ Seleccionar atributos

quitar atributos que valen siempre lo mismo ó son muy efectivos

## Ingeniería de Atributos - TEXTOS

¿Cómo represento un texto?

Vectores con el diccionario de palabras ...

Método tradicional: Bag of Words (bow)

↳ no tiene en cuenta el orden  
ni la cont. de veces que aparece

⇒ utilizar cont. por palabra

Problema: hay palabras muy comunes en el español que se ven a repetir

⇒ tf idf -

tf =  $\frac{f}{n}$

j

## ESTANDARIZACIÓN DE ATRIBUTOS

Ajunto le encole ← facilita en algunos algoritmos

→ para datos cont.

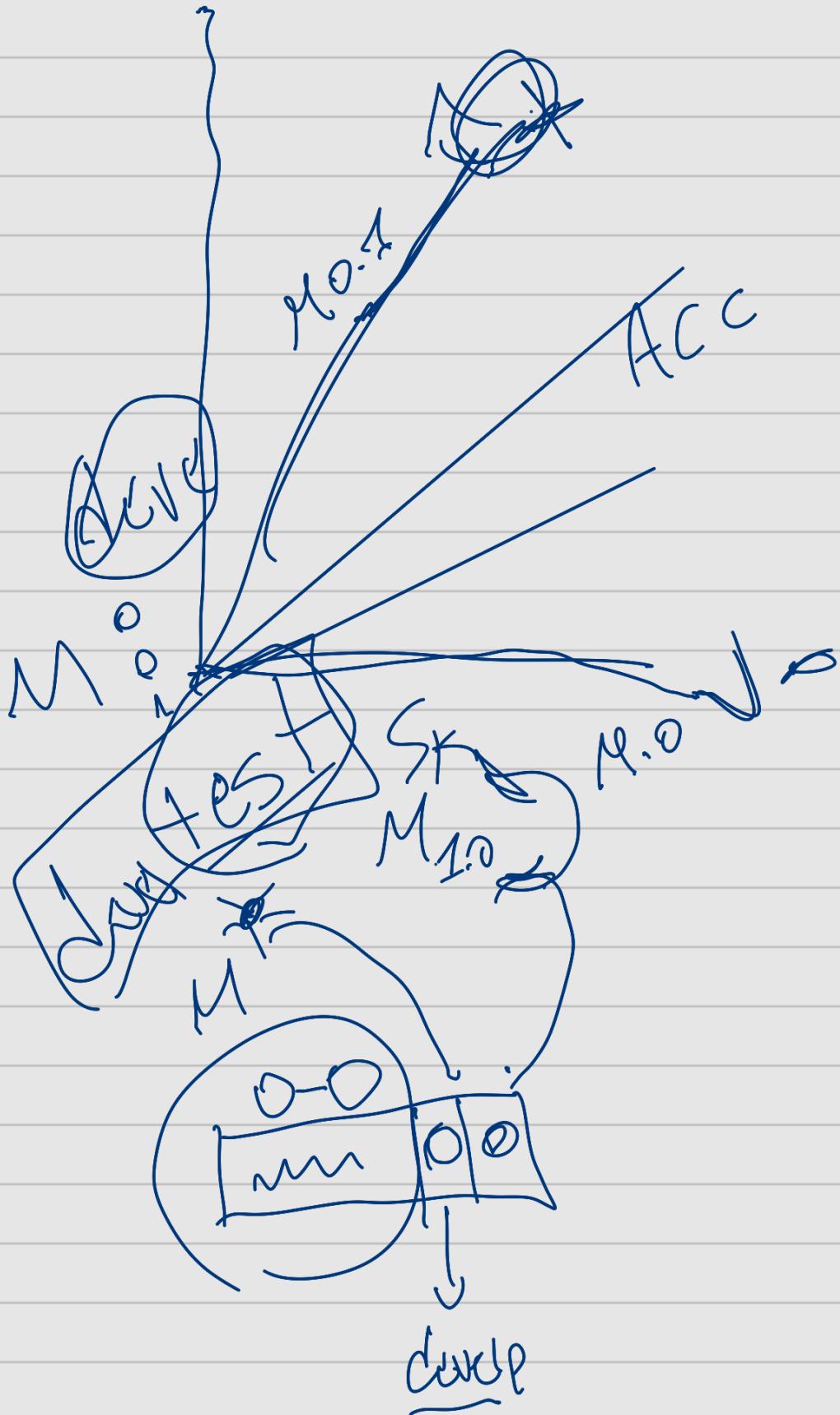
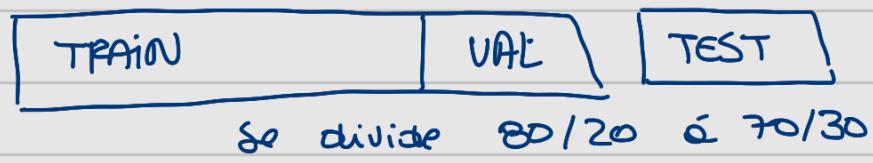
→ en test aplicamos lo calculado en el entrenamiento

## FASE 2 - División del conjunto de datos

Conj. de ent., testeos, [y validación]

↑  
se usa el ~~entrenamiento~~!!

↑  
se usa para ajustar



• Evaluaciones



Conjunto de datos desbalanceados  
→ OverSampling

→ Undersampling → reducir la cantidad de instancias de la clase mayoritaria

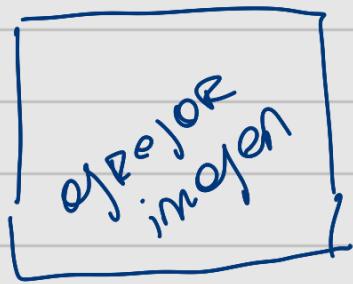
- 
- Cluster

## Selección de atributos

- busco sobre: id's, atributo repetitivo
- ⇒ • Método básico: —
- $\chi^2$  (chi squared)
  - Ganancia de información

Métodos wrappers

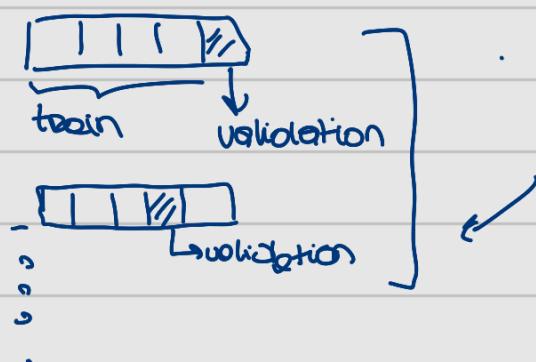
## FASE 3 - ENTRENAMIENTO



1)

2)

### Cross-validation



Tomo cada combinación de  
particiones  
la repito  
y luego elijo cual  
es el mejor

## FASE 4 - VALIDACIÓN

Accuracy → maldito: si lo doy en lo mejor  
del balanceador  
o en muy bien  
'juega'

TP ← true poss

positivos  
que predice  
como verdadero

¿Qué tan buen predictor es en el mundo real?

⇒ EVALUACIÓN DE HIPÓTESIS

• quiero evitar el azar

hipótesis

modelo

$$\text{error}_s(h) = \frac{1}{n} \sum d(y_i, h(x))$$

(para tener un buen estimador necesito muchos datos)

## Precisión y Recuperación

- la precisión es que tanto sacusto cuando digo TRUE
- Recall → de todos los + que hay e cuantos le sacaste
- Permite ver que pasa y que prefiiero hacer.  
por ejemplo → \*Recall cuando quiero levantar cosas aunque no les han puesto



Un bueno cuando es proporcional

$$+ \text{pre} \implies - \text{recall}$$

$$+\text{recall} \implies - \text{pre}.$$

\* Es dependiente de lo close que se elige

Mejorada - F (balance entre pre y recall)

↑ con esto logramos evaluar mejor

④ Depende del problema, uno preferir

## (\*) MATRIZ DE CONFUSIÓN

→ se estima que la diagonal sea más marcada

MACRO ADU

→ promedio de

para que sea bien en la mayoría y minoría

MICRO ADU

→

## Teer Multiclasses

→ calculo ocurrencia centro...

## Multietiquetas

L

'¿cómo se si mi resultado es bueno?'

- tener línea base ← método muy sencillo que practice

- línea máxima ← 'techo' combinado con el trabajo de humano

- comparan con lo que yo habré