Repaso clase 1

Machine Learning

Datos / Experiencia → aprendizaje → **Modelo** → Evaluar (métricas)

aprendizaje
- ⇓ algoritmo ¿cómo aprender la tarea?
- ↓ familias
  - algoritmos de reglas (simbolistas) → Decision tree → clasif. / regresión
  - → algoritmos geométricos
  - → " vecindad → K-NN → clasif. / regresión
  - → K-Means → agrupamientos

Modelo ¿cómo hacer la tarea concreta?
- supervisado (target) → clasificación / regresión
- no supervisado (no target) → agrupamiento / asociación / reducción dimensión

Algoritmos de vecindad K- : distancias

Workflow

Business understanding → Data Understanding → Data Preparation → Modelling → Evaluation

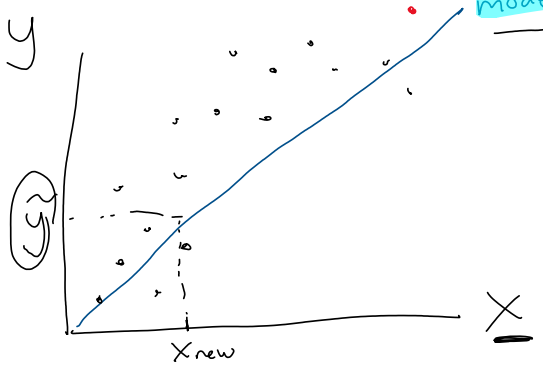
- Business understanding
  - definición tarea
  - ML Canvas
- Data Understanding
  - relación predictores / target
  - distribución y el tipo de predictores
    - ↳ data preparation plan
- Data Preparation
  - TODO A NUMÉRICO
    - → categóricas
    - → nulos
  - "traducir" el dataset a numeros
- Modelling
  - entrenar y fine-tune del algoritmo
- Evaluation
  - métricas
  - overfitting

Overfitting = memorización ≠ Generalización

X = predictor • : datos
y = target

Mismo dataset:

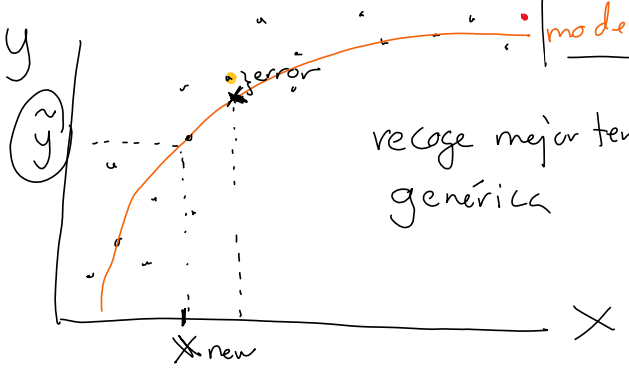
Bias - variance trade-off



Algoritmo: $Y = \beta_0 + \beta_1 X$
Underfitting

Métrica: error medio

↳ error datos → se equivoca mucho
=
↳ error nuevo dato → se equivoca mucho

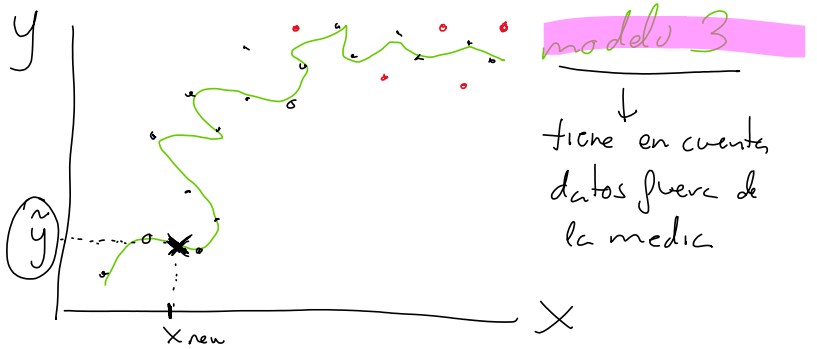


recoge mejor tendencia genérica

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X^2$

se equivoca un poco
=
se equivoca un poco

predicción exacta para los datos que ve



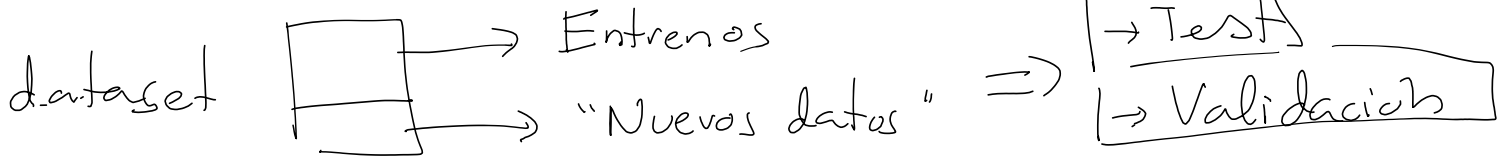
↳ tiene en cuenta datos fuera de la media

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X^2 + \beta_3 X^3$

Overfitting
no se equivoca
≠
sí que se equivoca bastante

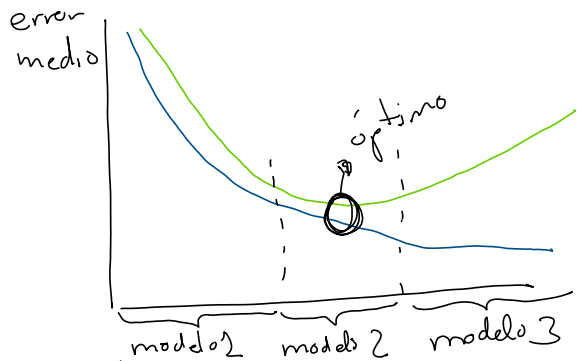
Objetivo: métrica en datos nuevos y antiguos debe ser igual y lo más alta posible

"Nuevos datos": no vamos a esperar al futuro, partimos los datos disponibles



Métrica

error medio → objetivo : lo más pequeño posible



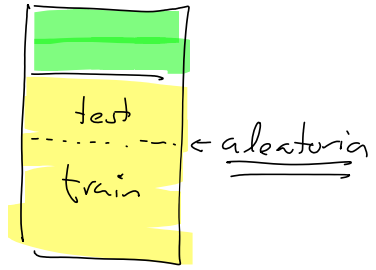
datos "antiguos"

datos "nuevos"

Tabla resumen

Validation: siempre a parte (2K-3K filas) y se parece a la realidad

development: test (>2K/3K filas) } aleatoria
train



⟹

Filas	
>60K	Random Holdout
60-20K	K-fold
<20K	Bootstrap

Note: ¿cómo escoger validation?

target → estacional: ciclo (Black Friday, venta helados, ...) ⟹ datos ciclo anterior

→ trend/tendencia: (precio vivienda) ⟹ datos más recientes

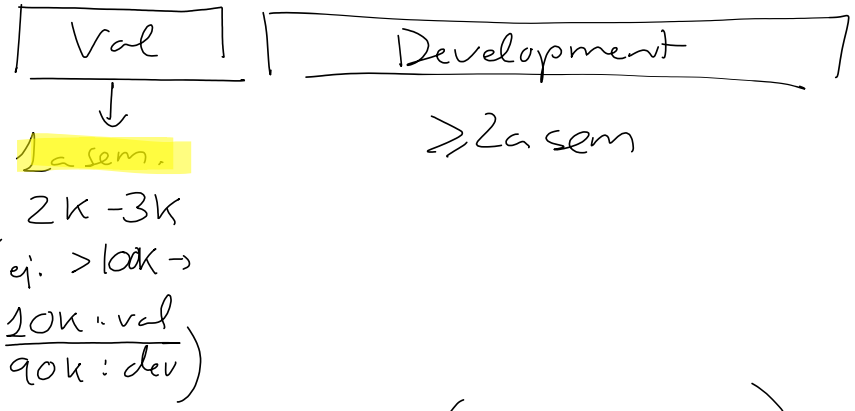
Ejemplo: nuevo lanzamiento producto (iphone 16) new

Business understanding: modelo que prediga quién es más probable que compre el iphone 16 en la primera semana de lanzamiento, de entre mi base de clientes.

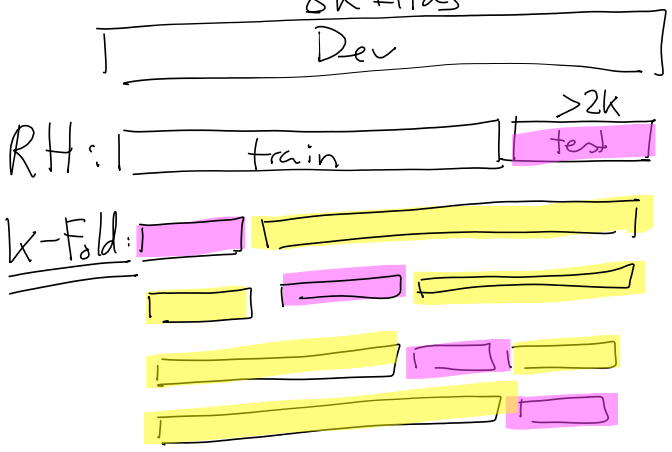
Validation: parecer a realidad

Test: misma distribucion que train ⇒

total ventas de iphone 15



¿Partición development?



Si fila 1 esta en versión 1 test, no puede estar en otras versiones de test

mismo algoritmo

modelo 1
modelo 2
modelo 3
modelo 4

→

overfitting (memorizacion)

train1 = 80%	test1 = 78%
train2 = 79%	test2 = 78%
train3 = 82%	test3 = 80%
train4 = 78%	test4 = 78%

overfitting: 2pp, 1pp, 2pp, 0pp → avg = 1,25pp

¿Cómo escoger modelo?

→ overfitting y performance medio

Data preparation → todo sea numérico

Nulos

1. No eliminar filas, porque cambia la distribución

   sí elimino si:

   I. Nulo en el target
   II. Afecta a una cantidad pequeña < 0,5% (1%)
   III. Fila en concreto tiene mayoría de predictores con nulos (> 70%)

2. Más lío en eliminar predictores

   ↳ eliminar atributo/predictor si > 95% filas con nulos

Tratamientos

I. Según categórica / numérica
II. Según algoritmo

Técnicas:
- moda (categ.) / media (numérica) / mediana (numérica) } prio 3
- rlarle categórica nueva (valor fuera de rango) } prio 2
- valores agrupados/cercanos
  - moda/media/mediana subset
  - time series: día anterior
  - ML: vecindad K-NN
- otras variables } prio 1
- data engineering

funciona bien en algoritmos simbolistas
- categórica: nueva etiqueta
- numérica: valor fuera de rango
  (boolean) → −1
  1/0

No funciona bien en geométricos o vecindad

Transformar categóricas:

1. Ajuste de tipos: Python lee mal el atributo (ej: IDs)

¿es booleano? (2 etiq. (nulo / no nulo)) — sí → True/False 1/0

no (etiq.) → ¿tiene orden?

sí → 1. Intrínseco → Label Encoding

→ respuesta encuesta

muy mal ..... muy bien

1 5

→ versiones app:

	vers.	=
Android: 'Baklava'	16	1
'Vanilla Ice Cream'	15	2
'Upside down cake'	14	3

no → ¿tengo tiempo? → sí → 2. Fuentes externas

Weather 01/01/18 → temp. mm lluvia

ciudad: población 1913

matrícula: antigüedad

no → ¿< 100 etiquetas?

sí → One hot encoding

no → reducir # etiquetas → agrupar etiq. minoritarias en "Otros"

Frecuency encoding

Conteo del número de veces que aparece la etiqueta en el dataset

ID	Ciudad
1	Madrid
2	Valencia
3	Valencia
4	Madrid
5	Madrid

⇒

ID	FE
1	3
2	2
3	2
4	3
5	3