ITBA
Instituto Tecnológicc
de Buenos Aires

18/OCTUBRE

SELECCIÓN Y EVALUACIÓN DE MODELOS

Selección de modelos

Supongamos que ya partimos los datos en train - test
(¿Por qué hacíamos esto?)

⇒ ¿Qué hacemos durante el desarrollo de modelos para **seleccionar modelos**?

Train-Validation

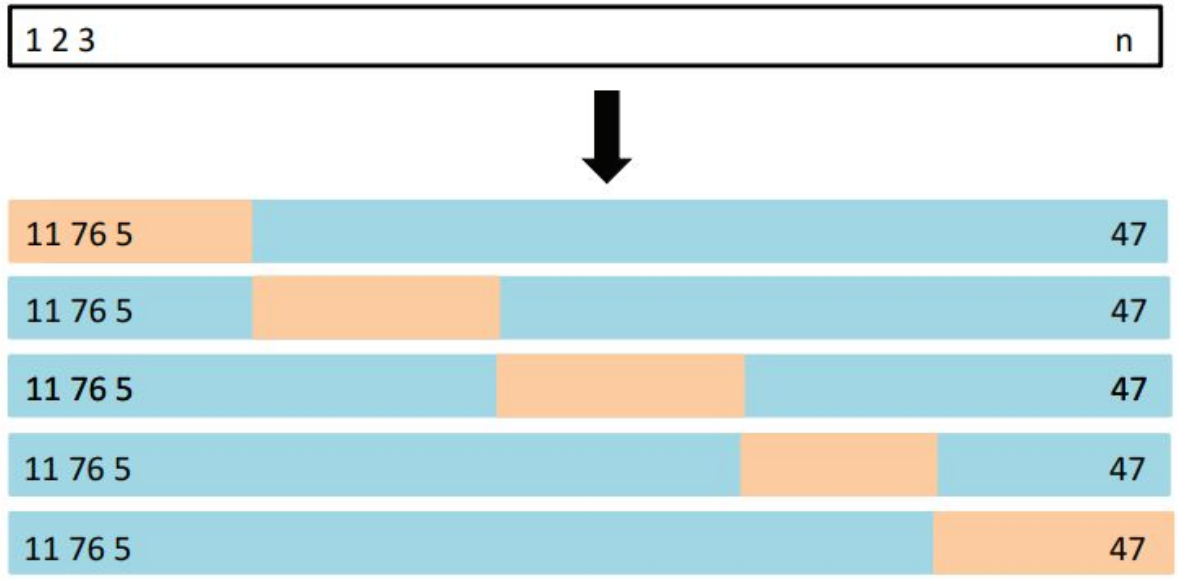
Entrenamos en unos datos (**train**) y evaluamos en otros (**validation/dev**)
(de acá en adelante, vamos a asumir que el conjunto de test ya fue separado)

Fuente: An Introduction to Statistical Learning (James *et al*, 2021)

k-Fold Cross-Validation

Cuando hay pocos datos, la **estimación en único validation set puede ser muy variable**
⇒ ¡Intentemos usar todos los datos!

Ejemplo: k-fold CV con k=5



→ *Para i=1,...,5:*

- *Entrenamos en los **datos restantes** y predecimos en **fold i***

→ *Calculamos el **error** $\mathbf{Err}_{\mathbf{CVk}}$ con la predicción de cada obs.*

$\mathrm{Err}_{\mathrm{CV}_k}$

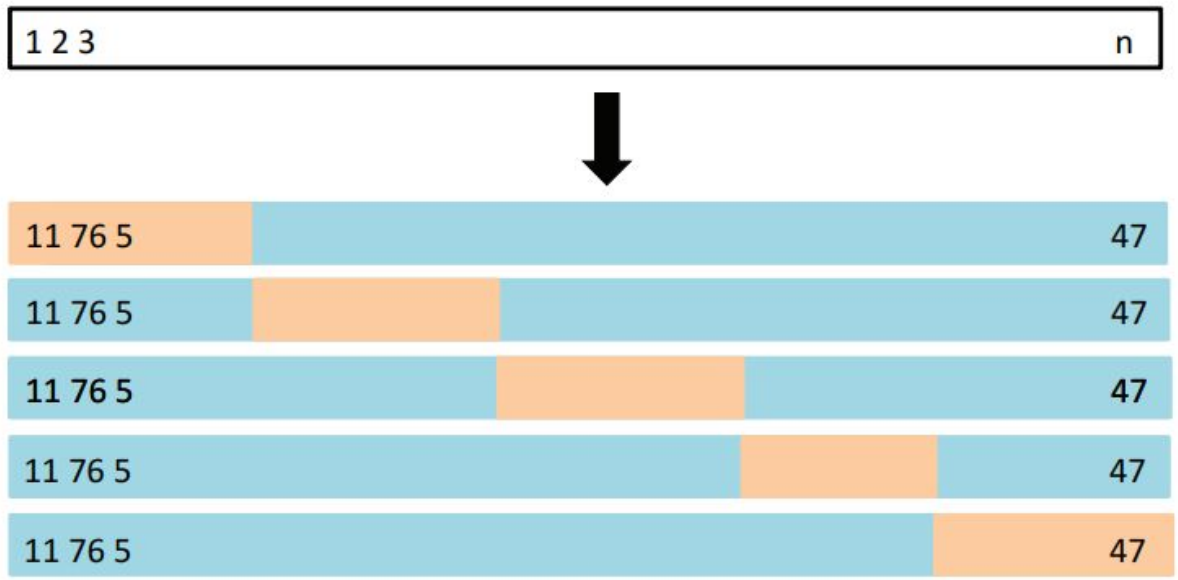
¿Es una estimación *confiable* del error de mi mejor modelo?

Fuente: An Introduction to Statistical Learning (James *et al*, 2021)

k-Fold Cross-Validation

Cuando hay pocos datos, la **estimación en único validation set puede ser muy variable**
⇒ ¡Intentemos usar todos los datos!

Ejemplo: k-fold CV con k=5



→ *Para i=1,...,5:*

- *Entrenamos en los **datos restantes** y predecimos en **fold i***

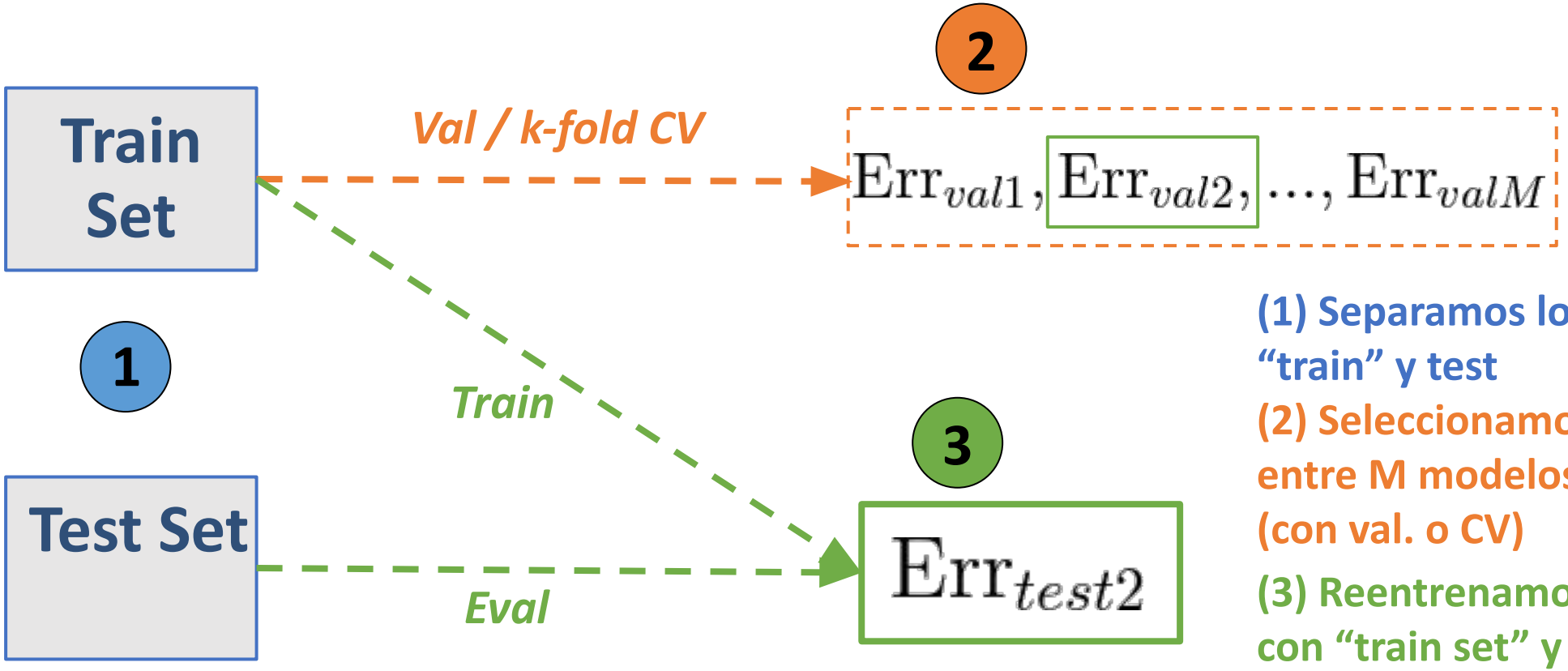
→ *Calculamos el **error** $\mathbf{Err}_{\mathbf{cvk}}$ con la predicción de cada obs.*

$\mathrm{Err}_{\mathrm{CV}_k}$

¿Es una estimación *confiable* del error de mi mejor modelo?
¡NO! solo nos interesa para encontrar el mínimo

Fuente: An Introduction to Statistical Learning (James *et al*, 2021)

Selección de modelos



(1) Separamos los datos en "train" y test

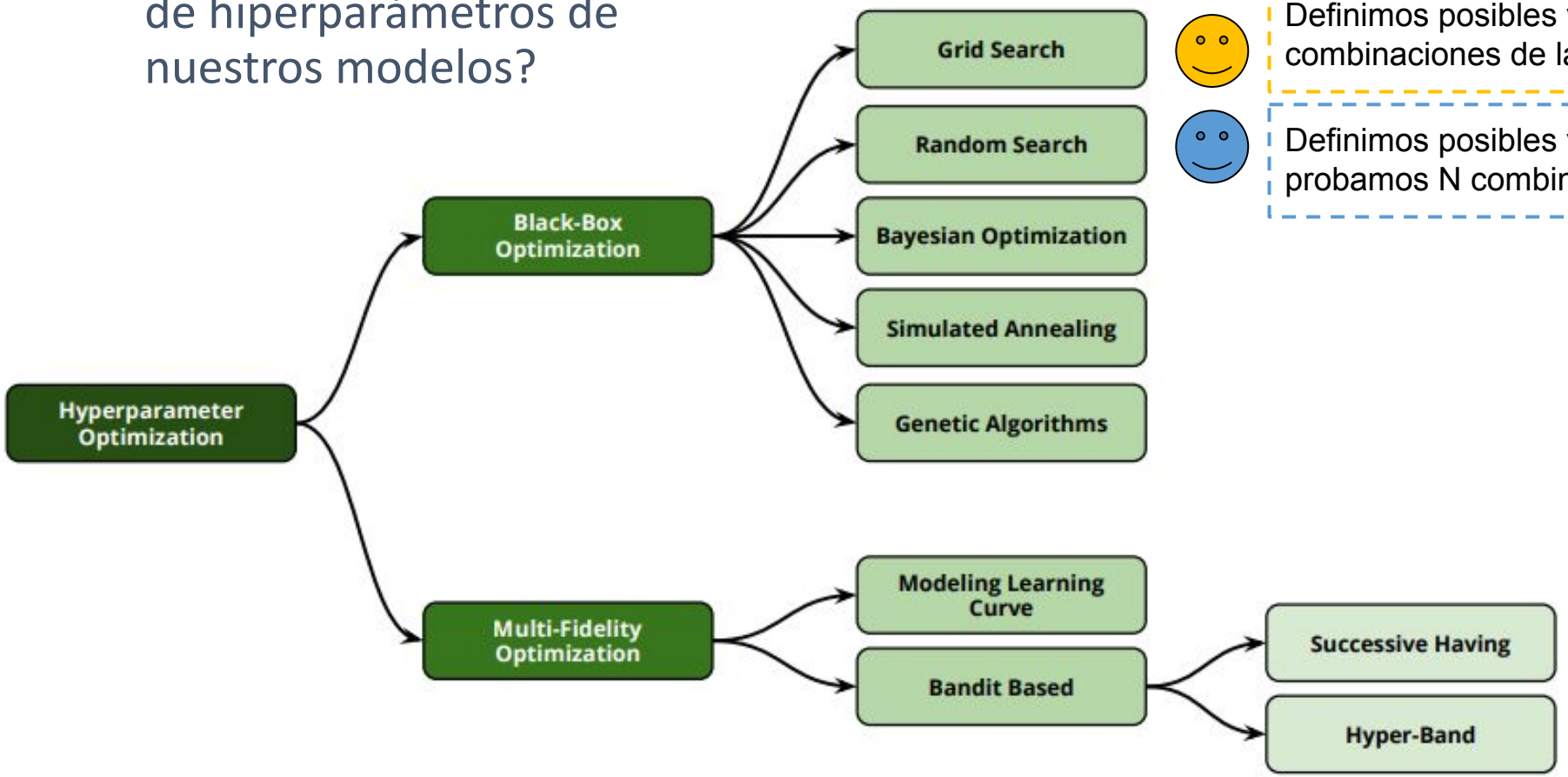
(2) Seleccionamos uno de entre M modelos según Err_{vali} (con val. o CV)

(3) Reentrenamos el modelo con "train set" y lo evaluamos en test

IMPORTANTE: el criterio de partición es el mismo en (1) y en (2)

Optimización de Hiperparámetros (HPO)

¿Cómo recorremos el espacio de hiperparámetros de nuestros modelos?



Definimos posibles valores y probamos todas las combinaciones de la grilla

Definimos posibles valores (o distribuciones) y probamos N combinaciones al azar (sin reposición)

Fuente: Automated Machine Learning: State-of-The-Art and Open Challenges (Elshawi *et al*, 2019)

Evaluación de modelos

OK pero… ¿qué es el *error* de un modelo…?

¿Cómo medimos el rendimiento / performance de un modelo?
(ya sea en validation, CV, o en test)

Métricas de clasificación

$$y \in \{0, 1\} \quad \hat{y} \in \{0, 1\}$$

		True class		
		− or Null	+ or Non-null	Total
Predicted class	− or Null	True Neg. (TN)	False Neg. (FN)	N*
	+ or Non-null	False Pos. (FP)	True Pos. (TP)	P*
	Total	N	P	

*Supuesto: nuestro sistema **clasifica** entre 0 y 1*

***Accuracy** (tasa de aciertos)*

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}} = 1 - \text{Err} = \frac{1}{n} \sum I(y_i = \hat{y}_i)$$

Fuente: An Introduction to Statistical Learning (James *et al*, 2021)

Métricas de clasificación

Ejemplo: clasificación de default (paga / no paga)

		Default obs.		
		0	1	Total
Default pred.	0	4 988	3 784	8 772
	1	12	1 216	1 228
	Total	5 000	5 000	10 000

FP = …

FN = …

Acc = …

Métricas de clasificación

Ejemplo: clasificación de default (paga / no paga)

		Default obs.		
		0	1	Total
Default pred.	0	4 988	3 784	8 772
	1	12	1 216	1 228
	Total	5 000	5 000	10 000

FP = 12

FN = 3784

Acc = 62%

FPR = FP / N = 0.24%

(tasa de falsos pos.)

FNR = FN / P = 75.7%

(tasa de falsos neg.)

Métricas de clasificación

Ejemplo: clasificación de default (paga / no paga) pero un poco más realista…

		Default obs.		
		0	1	
Default pred.	0	9 644	252	9 896
	1	23	81	104
		9 667	333	10 000

FP = …

FN = …

Acc = …

FPR = …

FNR = …

Fuente: An Introduction to Statistical Learning (James *et al*, 2021)

Métricas de clasificación

Ejemplo: clasificación de default (paga / no paga) pero un poco más realista…

		Default obs.		
		0	1	
Default pred.	0	9 644	252	9 896
	1	23	81	104
		9 667	333	10 000

FP = 23
FN = 252
Acc = 97.3%

FPR = 0.24%
FNR = 75.7%

En problemas "desbalanceados" (con una clase mayoritaria) el accuracy es poco informativo
¿Cuánto es Acc si nuestro modelo siempre predice 0?

Fuente: An Introduction to Statistical Learning (James *et al*, 2021)

Métricas de clasificación

Ejemplo: clasificación de default (paga / no paga) pero un poco más realista…

		Default obs.		
		0	1	
Default pred.	0	9 644	252	9 896
	1	23	81	104
		9 667	333	10 000

FP = 23
FN = 252
Acc = 97.3%

FPR = 0.24%
FNR = 75.7%

En problemas "desbalanceados" (con una clase mayoritaria) el accuracy es poco informativo
¿Cuánto es Acc si nuestro modelo siempre predice 0?
⟶ Acc = (9644 + 23 + 0) / 1000 = 96.67%

Fuente: An Introduction to Statistical Learning (James *et al*, 2021)

Métricas de clasificación

La mayor parte de las veces nos va a interesar evaluar **otras métricas** en lugar de la tasa de aciertos

⇒ hoy vamos ver algunas y más adelante vamos a ver cómo elegir

Métricas de clasificación

		Obs		
		0	**1**	
Pred	**0**	TN	FN	N*
	1	FP	TP	P*
		N	P	

$$\mathrm{TPR} = \mathrm{TP}/\mathrm{P} = 1 - \mathrm{FNR}$$

Recall / Sensibilidad

$$\mathrm{TNR} = \mathrm{TN}/\mathrm{N} = 1 - \mathrm{FPR}$$

Especificidad

$$\mathrm{PPV} = \mathrm{TP}/\mathrm{P}^*$$

Precision
(Positive predictive value)

		Default obs.		
		0	1	
Default pred.	0	9 644	252	9 896
	1	23	81	104
		9 667	333	10 000

Acc = 97.3%
Recall = 24.3% = 1 - 0.757
Specificity = 99.8% = 1 - 0.0024
Precision = 77.9%

Métricas de clasificación

		Obs		
		0	**1**	
Pred	**0**	TN	FN	N*
	1	FP	TP	P*
		N	P	

$$\mathrm{TPR} = \mathrm{TP}/\mathrm{P} = 1 - \mathrm{FNR}$$

Recall / Sensibilidad

$$\mathrm{TNR} = \mathrm{TN}/\mathrm{N} = 1 - \mathrm{FPR}$$

Especificidad

$$\mathrm{PPV} = \mathrm{TP}/\mathrm{P}^*$$

Precision
(Positive predictive value)

		Default obs.		
		0	1	
Default pred.	0	9 644	252	9 896
	1	23	81	104
		9 667	333	10 000

¿Qué modelo optimiza el recall?

¿Qué modelo optimiza la especificidad?

Métricas de clasificación

		Obs		
		0	**1**	
Pred	**0**	TN	FN	N*
	1	FP	TP	P*
		N	P	

$$\mathrm{TPR} = \mathrm{TP}/\mathrm{P} = 1 - \mathrm{FNR}$$ ***Recall / Sensibilidad***

$$\mathrm{TNR} = \mathrm{TN}/\mathrm{N} = 1 - \mathrm{FPR}$$ ***Especificidad***

$$\mathrm{PPV} = \mathrm{TP}/\mathrm{P}^*$$ ***Precision*** *(Positive predictive value)*

		Default obs.		
		0	1	
Default pred.	0	9 644	252	9 896
	1	23	81	104
		9 667	333	10 000

¿Qué modelo optimiza el recall?

Predigo siempre 1 → (252+81)/333 = 1

¿Qué modelo optimiza la especificidad?

Predigo siempre 0 → (9644+23)/9667 = 1

⇒ hay un **trade-off** entre *capturar 1s* (**recall**) y *capturar 0s* (**especificidad**)

Métricas de clasificación

Métricas de la matriz de confusión que pueden reemplazar a accuracy (más adelante vamos a ver cómo elegir)

F-score (combinación entre precision y recall)

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R} \qquad \mathrm{F}_1 = \frac{2PR}{P+R}$$

$\beta = 2$ ***+ recall***

$\beta = 0.5$ ***+ precision***

Función de costos

(a veces podemos definir los costo relativos/absolutos de acertar y/o errar)

$$C = C_{\mathrm{TN}}\mathrm{TN} + C_{\mathrm{TP}}\mathrm{TP} + C_{\mathrm{FN}}\mathrm{FN} + C_{\mathrm{FP}}\mathrm{FP}$$

Métricas de clasificación

En general los modelos nos devuelven un **puntaje** o ***score***, no una clase

$$\hat{p} \in [0, 1]$$

$$\hat{y} = \underset{y}{\operatorname{argmax}} \widehat{p(y|x)}$$

$$(\hat{y} = 1 \iff \hat{p} > 0.5)$$

Esto no necesariamente es óptimo!

- porque el score puede estar *descalibrado,* y/o
- porque ese umbral puede no optimizar nuestra métrica

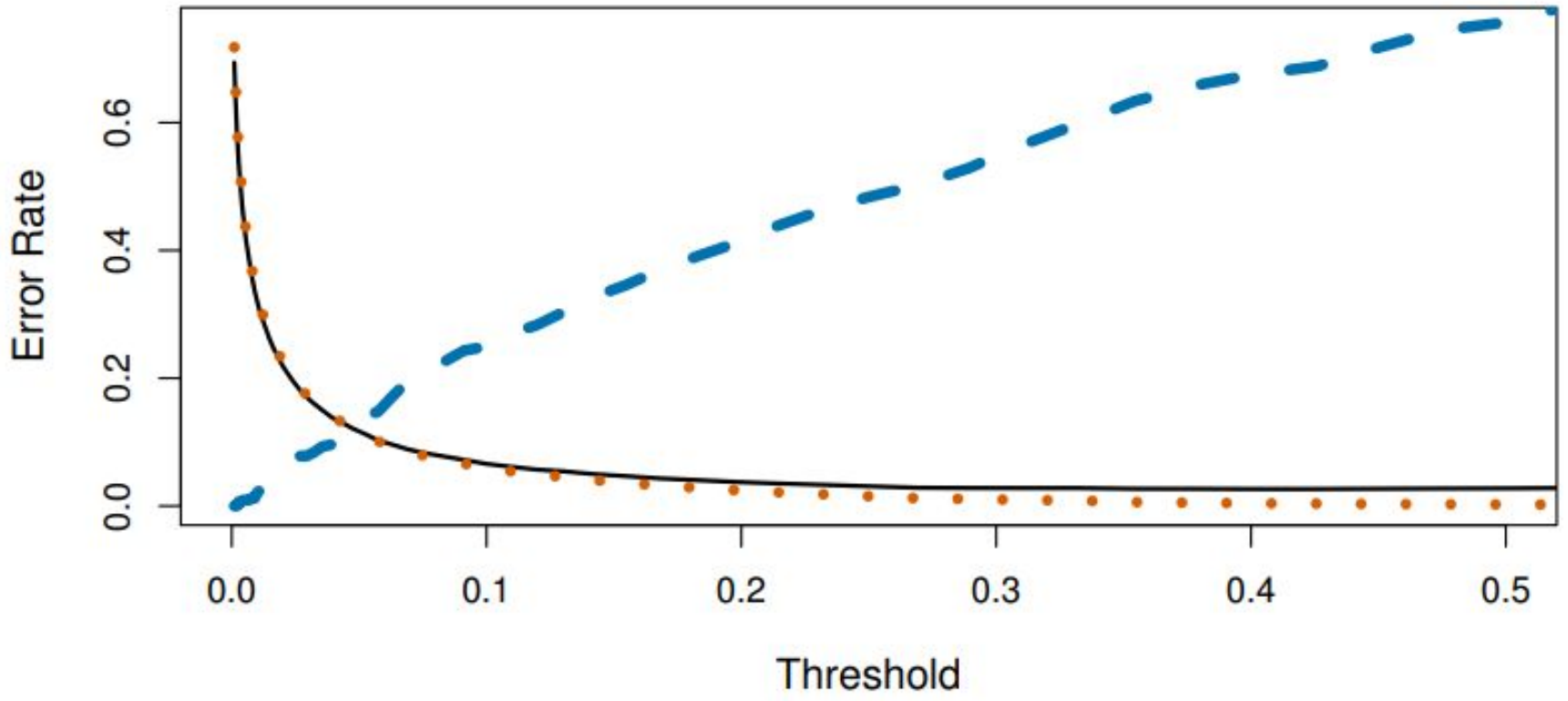
⇒ Tenemos que elegir un **umbral** que optimice la métrica que nos interesa

Fuente: An Introduction to Statistical Learning (James *et al*, 2021)

Métricas de clasificación

$$\hat{y} = \begin{cases} 1 & \text{si } \hat{p} \geq \theta \\ 0 & \text{si } \hat{p} < \theta \end{cases}$$

Por ejemplo, bajar el umbral puede reducir Accuracy pero mejorar Recall → y entonces, quizás, la métrica que nos interese



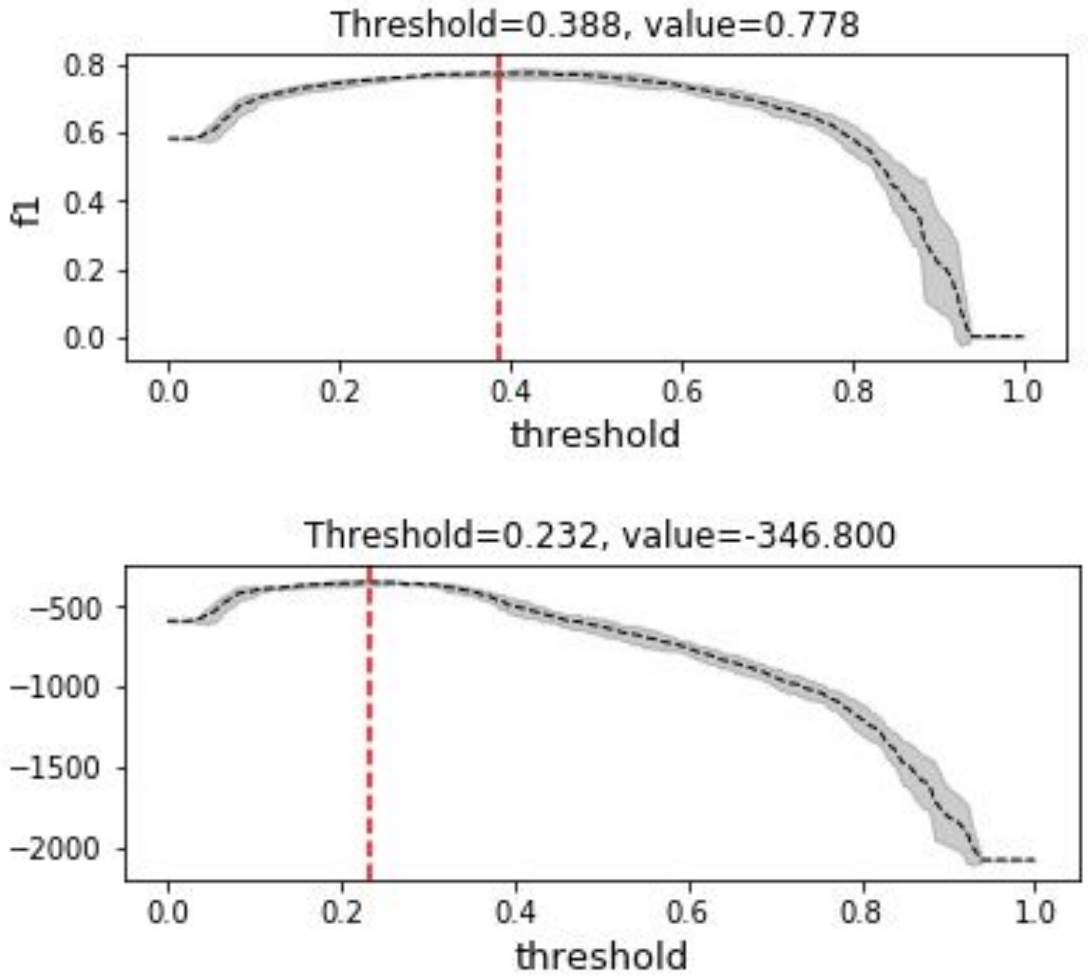
FNR = 1 - Recall

Error rate = 1 - Accuracy

FPR = 1 - Specificity

Fuente: An Introduction to Statistical Learning (James *et al*, 2021)

Métricas de clasificación



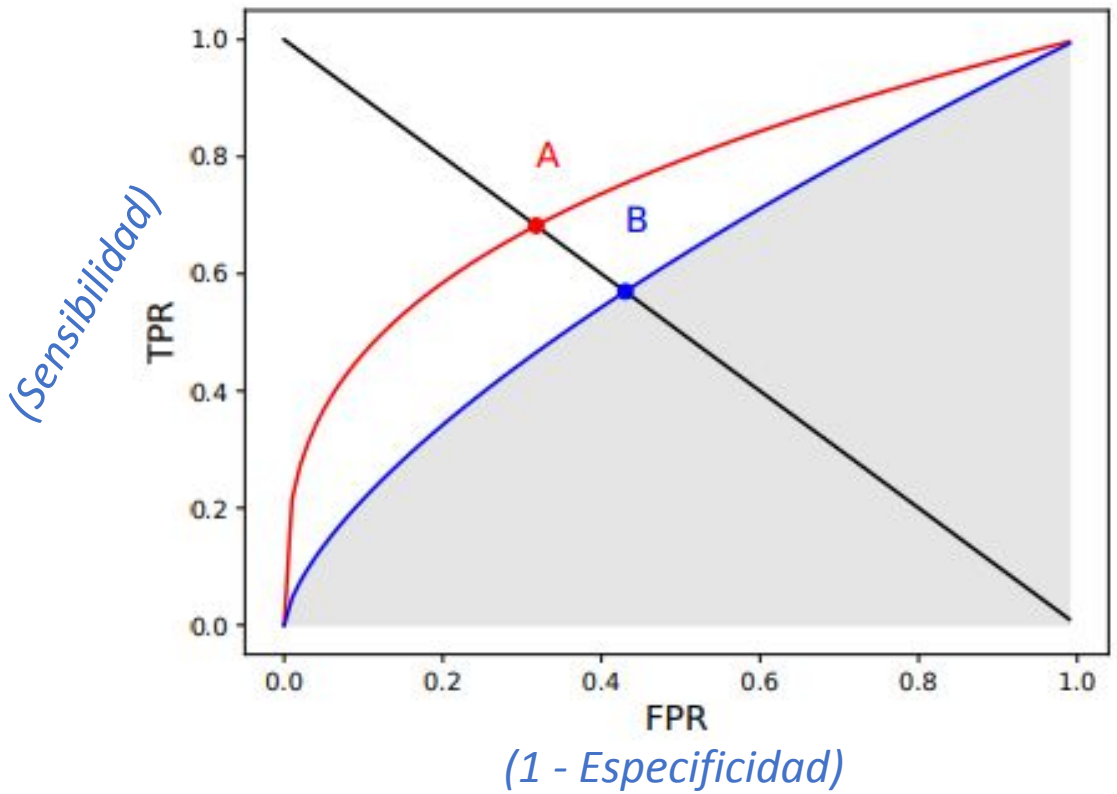
$$\hat{y} = \begin{cases} 1 & \text{si } \hat{p} \geq \theta \\ 0 & \text{si } \hat{p} < \theta \end{cases}$$

Por ejemplo, bajar el umbral puede reducir Accuracy pero mejorar Recall ⟶ ***y entonces, quizás, la métrica que nos interese***

Métricas de clasificación

$$y \in \{0, 1\} \quad \hat{p} \in [0, 1]$$

A veces nos interesa evaluar *globalmente* los scores i.e. para cualquier umbral posible



AUROC
(*Area Under the ROC Curve*)

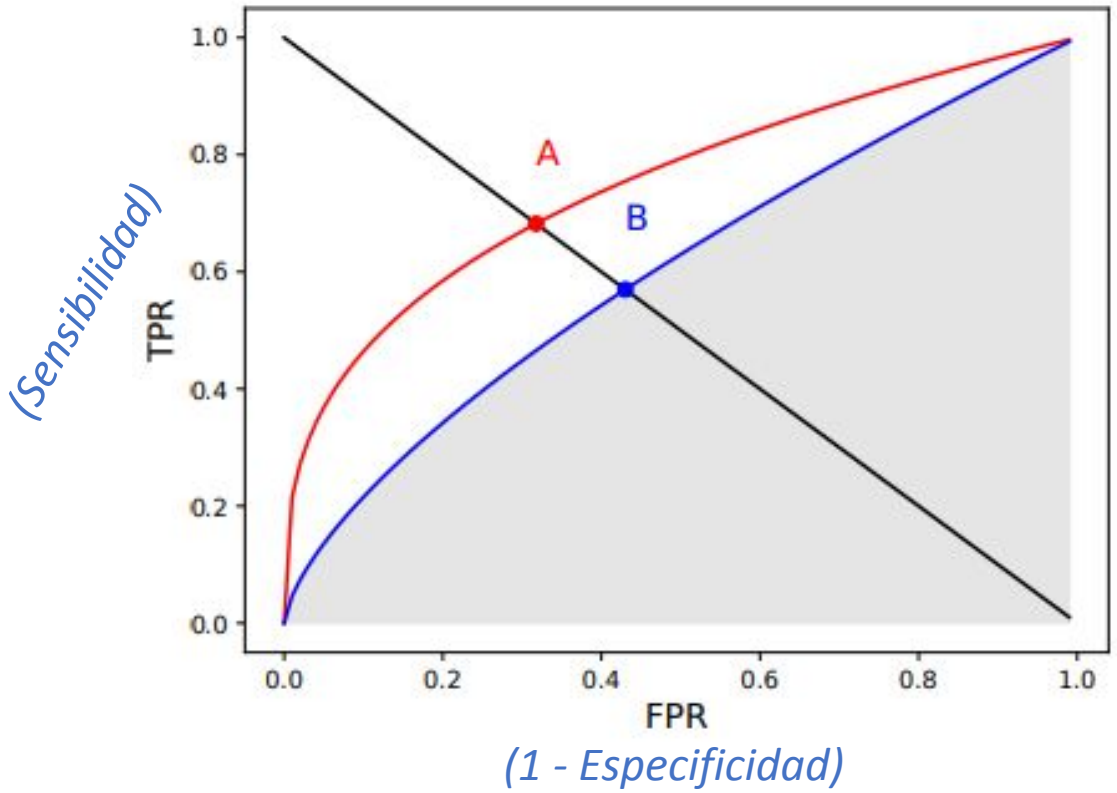
- Mide la **capacidad global de discriminar** entre 0s y 1s
- Se puede interpretar como la probabilidad de que el score de una obs. positiva seleccionada al azar sea más alto que el de una negativa

Fuente: Probabilistic Machine Learning (Murphy, 2021)

Métricas de clasificación

$$y \in \{0, 1\} \quad \hat{p} \in [0, 1]$$

A veces nos interesa evaluar *globalmente* los scores i.e. para cualquier umbral posible



AUROC
(*Area Under the ROC Curve*)

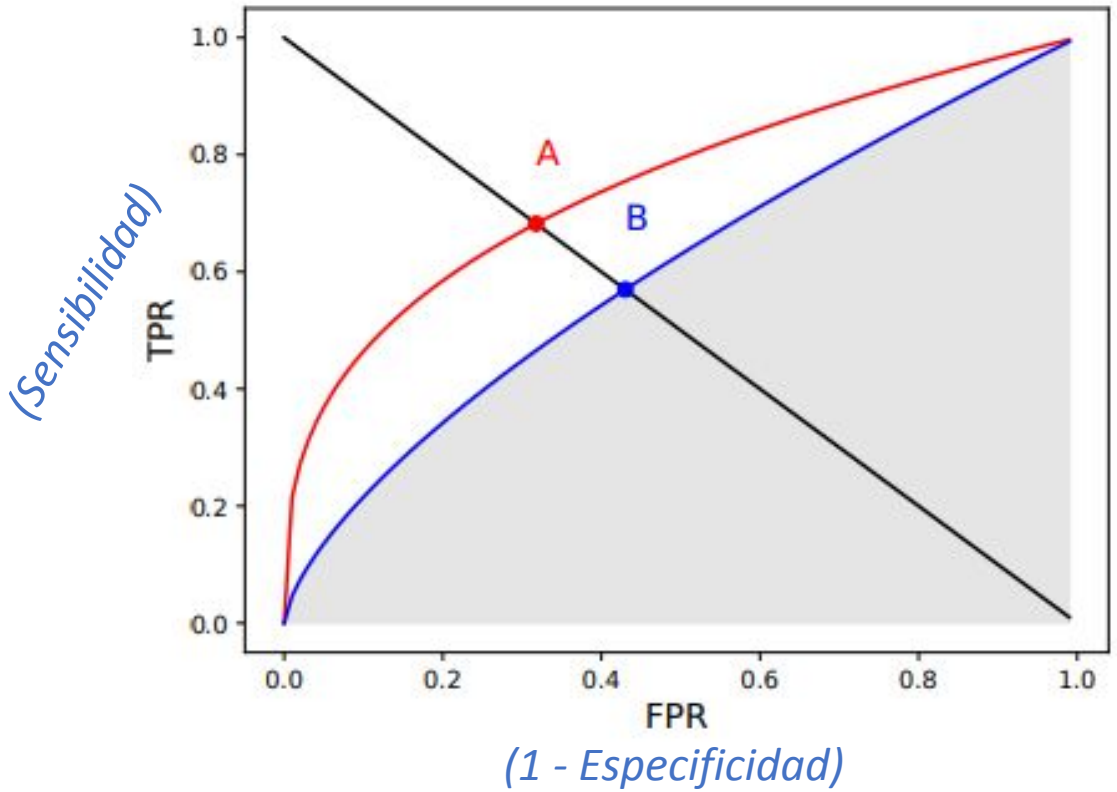
¿A qué valor tiende el AUROC de un score random?

Fuente: [Probabilistic Machine Learning (Murphy, 2021)]

Métricas de clasificación

$$y \in \{0, 1\} \quad \hat{p} \in [0, 1]$$

A veces nos interesa evaluar *globalmente* los scores i.e. para cualquier umbral posible



AUROC
(*Area Under the ROC Curve*)

¿A qué valor tiende el AUROC de un score random?

0.5 (para cualquier punto de corte, me equivoco igual ⟶ FPR ≈ FNR ⟶ diagonal imaginaria en el gráfico)

Fuente: Probabilistic Machine Learning (Murphy, 2021)

Métricas de regresión

Gráfico de residuos

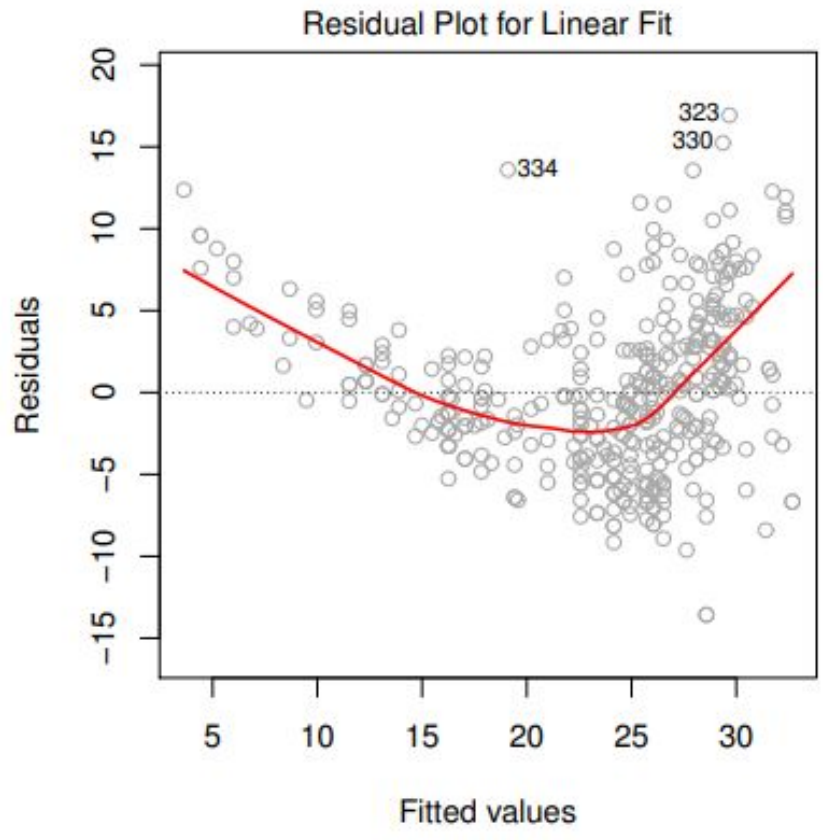
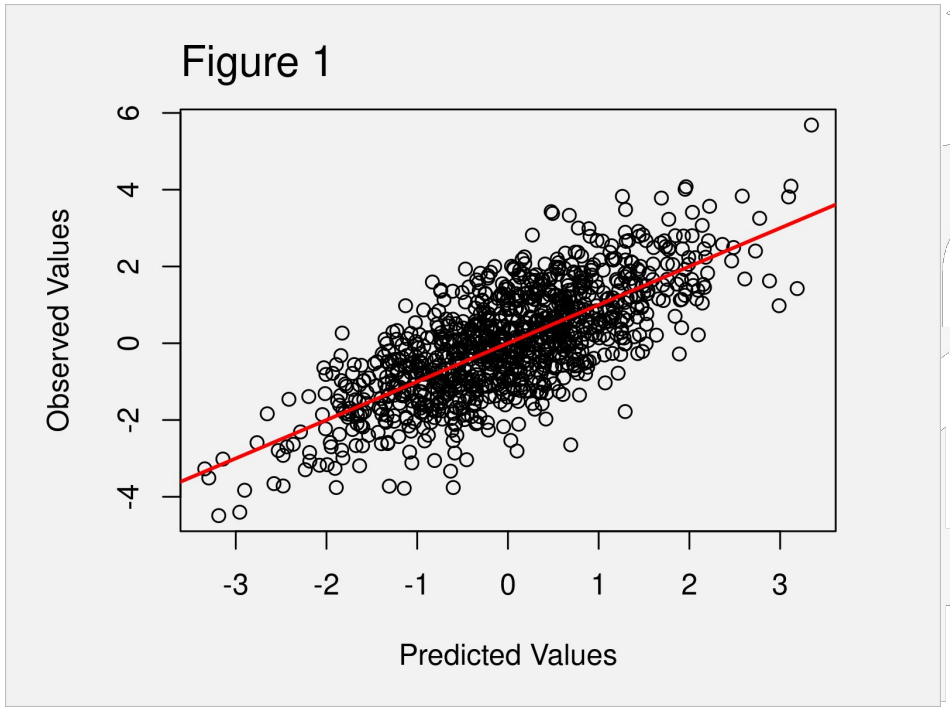


Gráfico de correlación



Fuente: An Introduction to Statistical Learning (James *et al*, 2021)

Métricas de regresión

$$\mathrm{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Root mean squared error
(error cuadrático)

$$\mathrm{MAE}(y, \hat{y}) = \frac{1}{n}\sum_{i=0}^{n}|y_i - \hat{y}_i|$$

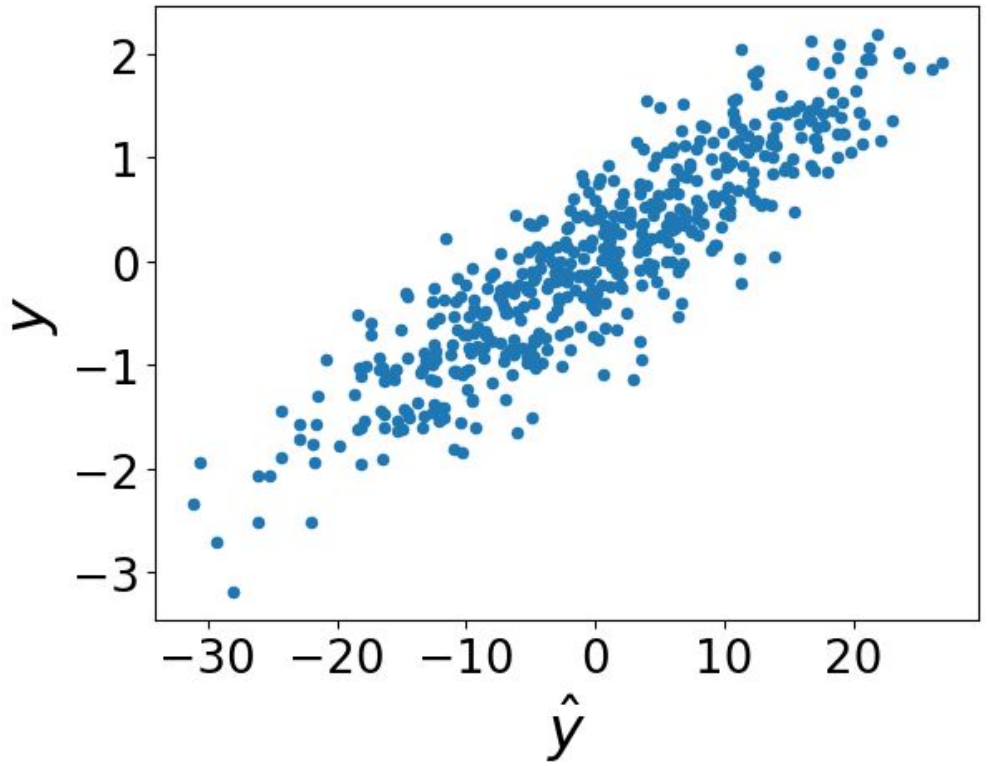
Mean absolute error
(error absoluto)

$$\mathrm{MAPE}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{\max(\epsilon, |y_i|)}$$

Mean absolute percentage error
(error porcentual)

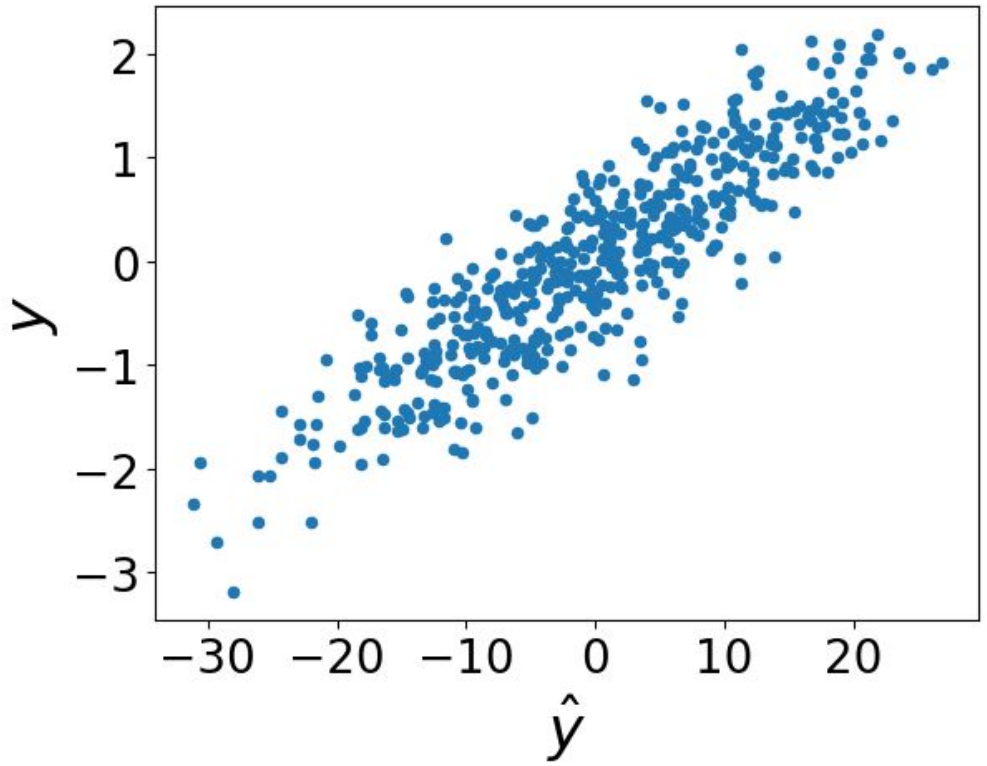
Métricas de regresión

¿Cómo nos darían las métricas anteriores en este ejemplo?
(bien / mal / regular / kcyo)



Métricas de regresión

¿Cómo nos darían las métricas anteriores en este ejemplo?
(bien / mal / regular / kcyo)



¿Y si usamos el viejo y querido R cuadrado?
Veamos… →

Métricas de regresión

R cuadrado

(A) $$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

(B) $$R^2(y, \hat{y}) = \text{pearson}(y, \hat{y})^2$$

"(A) y (B) son iguales …"
"0 < R^2(A) < 1 …"

- Podemos interpretar R^2 (A) como la comparación entre nuestro modelo (*y_hat*) y un modelo baseline (*y_mean*)
- R^2 está basado en MSE

Métricas de regresión

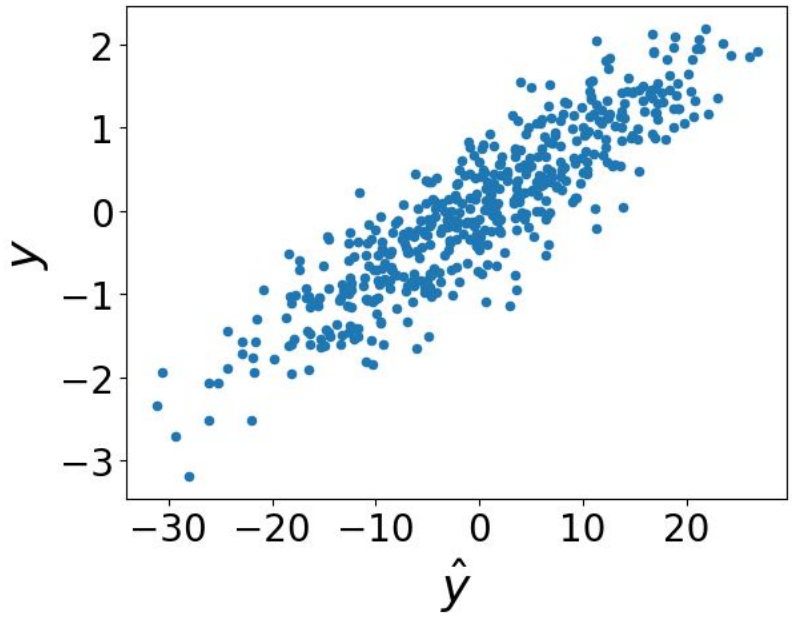
R cuadrado

(A) $$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

(B) $$R^2(y, \hat{y}) = \text{pearson}(y, \hat{y})^2$$

- Podemos interpretar R^2 (A) como la comparación entre nuestro modelo (*y_hat*) y un modelo baseline (*y_mean*)
- R^2 está basado en MSE

"(A) y (B) son iguales …"
"0 < R^2(A) < 1 …"
en los datos de entrenamiento de MCO!



R^2(B) = 0.80
R^2(A) = -108