Argentina programa 4.0

Universidad Nacional de San Martín

Módulo 2 - Ciencia de Datos

Argentina programa 4.0

Universidad Nacional de San Martín

Módulo 2 - Ciencia de Datos

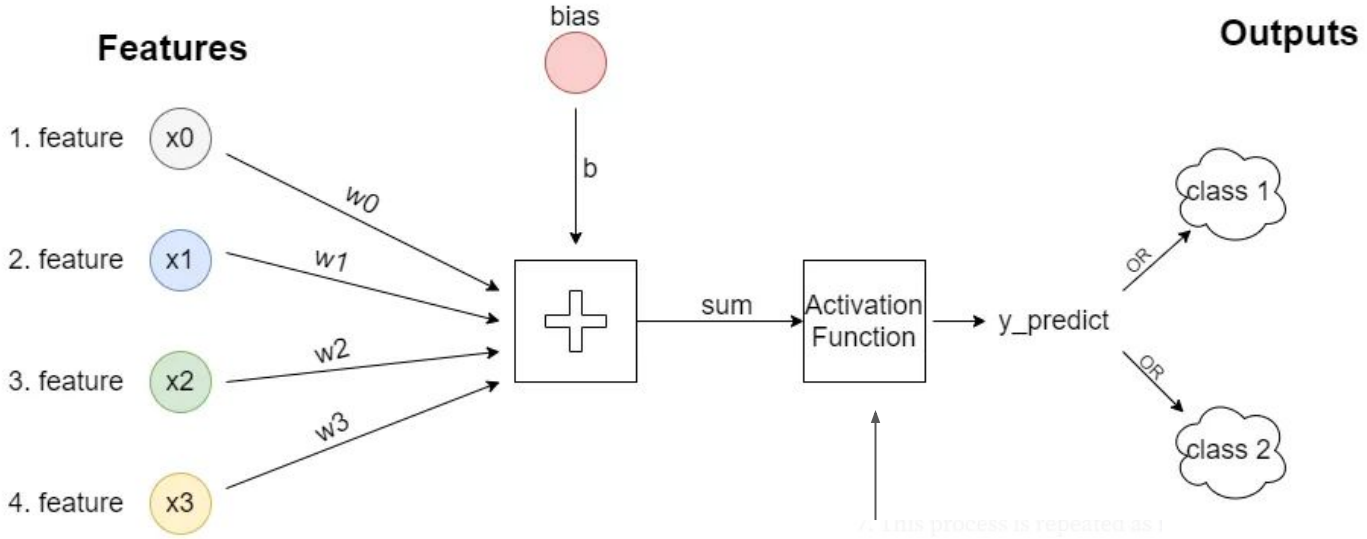
Semana 9. Clasificación

Repasando e Integrando

Weights (pesos):
coeficientes de cada variable (feature)

Bias: intercept

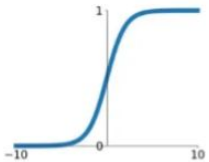
Sum:
(x0 * w0) + (x1 * w1) + (x2 * w2) + (x3 * w3) + (x4 * w4) + b



Predicción (y_predict):
activation_function(sum)

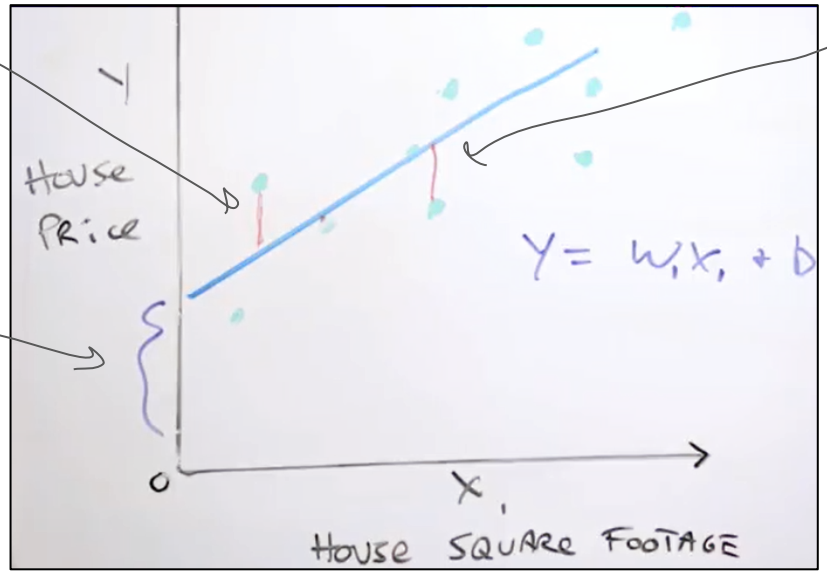
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



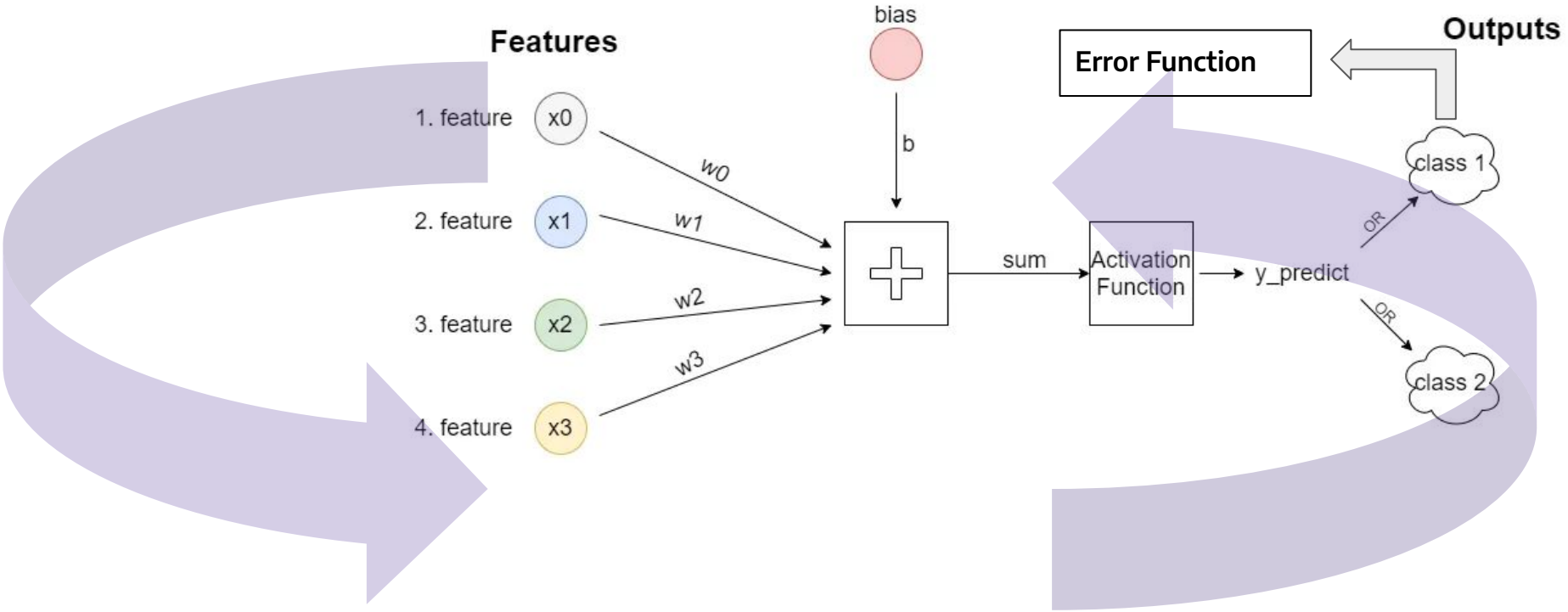
Vocabulario, jerga (conceptos)

Machine Learning	En otras áreas
$y = w_1 * x_1 + b$	$y = m * a + b$ (recta)
Bias	Ordenada al origen (recta), Intercept
Weights	Pendiente (recta), Coeficientes
Loss	Error

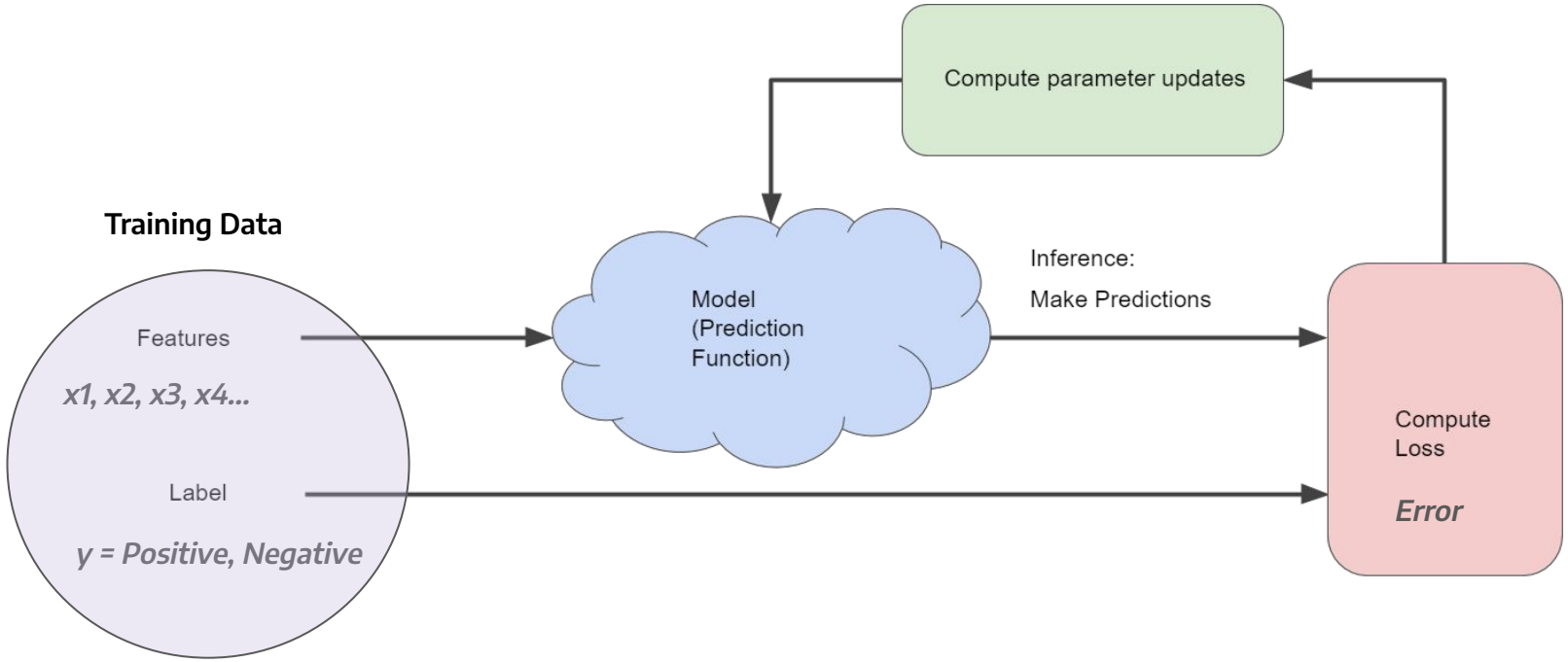


En este mundo hay mucha confusión ...

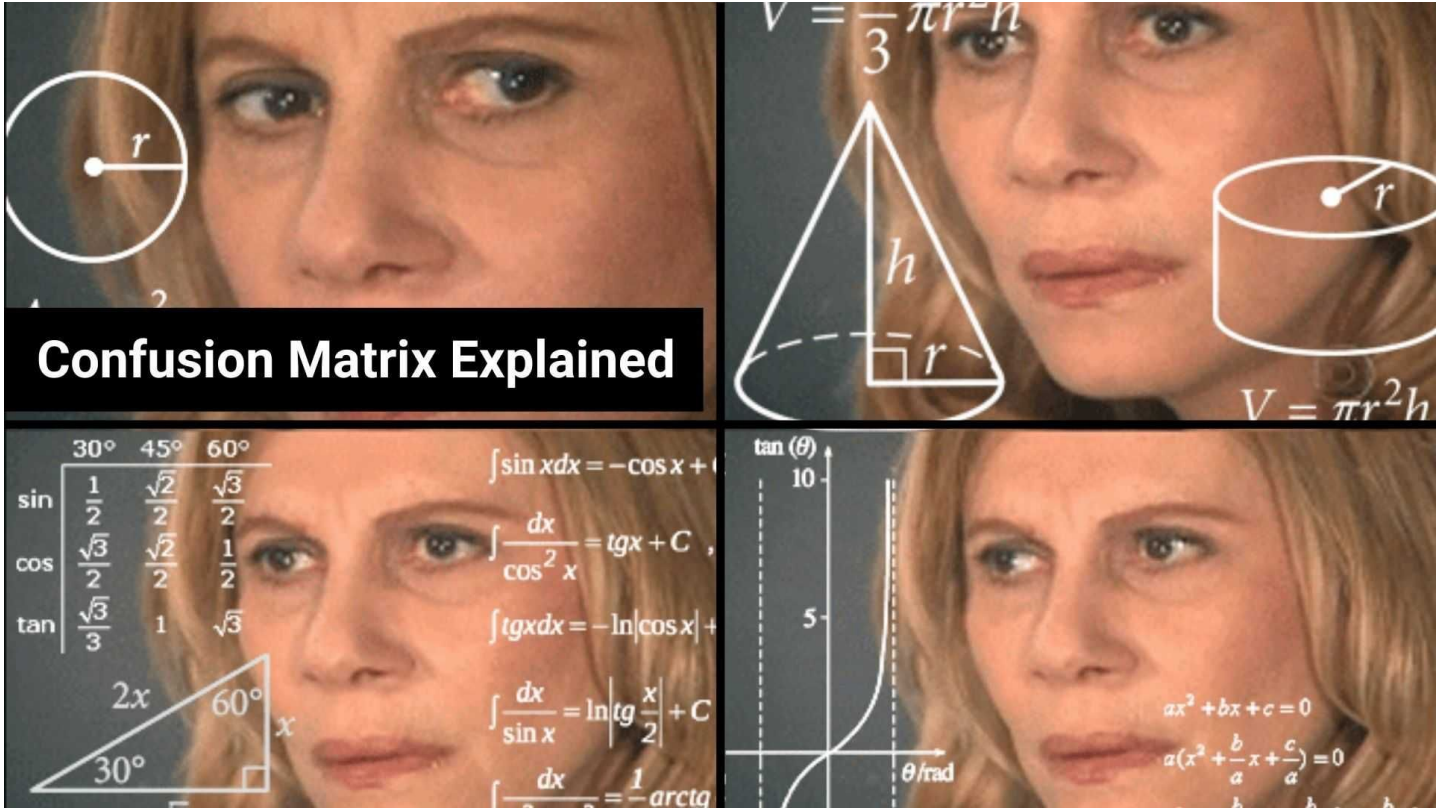
Repaso e Integración: función de error



Iterando para reducir el error



Sobre la confusión y sus matrices



De donde viene la matriz de confusión?

La ***"matriz de confusión"*** se usa para **evaluar algoritmos** de clasificación/predicción, y ver en qué casos **confunde** clases. He ahí la confusión. Si hay 6 perros reales y el algoritmo predice o clasifica bien solamente 3, se está ***confundiendo*** en otros 3 casos.

Se la llama también **matriz de error**, aunque también muestra los **aciertos**! Con total validez se la podría llamar "Matriz de Equivocaciones y Aciertos del algoritmo".

Aunque sería un poco largo, demasiado claro y poco perverso.

En este mundo hay mucha confusión ...

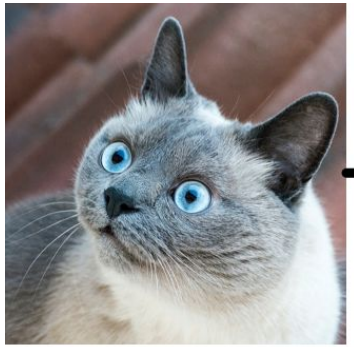
Por qué matriz si es una tabla?

Por qué la matriz se llama “de confusión” si **muestra también los aciertos del algoritmo**?

https://en.wikipedia.org/wiki/Confusion_matrix

“En este mundo hay mucha confusión, suenan los tambores de la rebelión ...” – Manu Chao

Un modelo de clasificación fácil



Binary Image Classifier

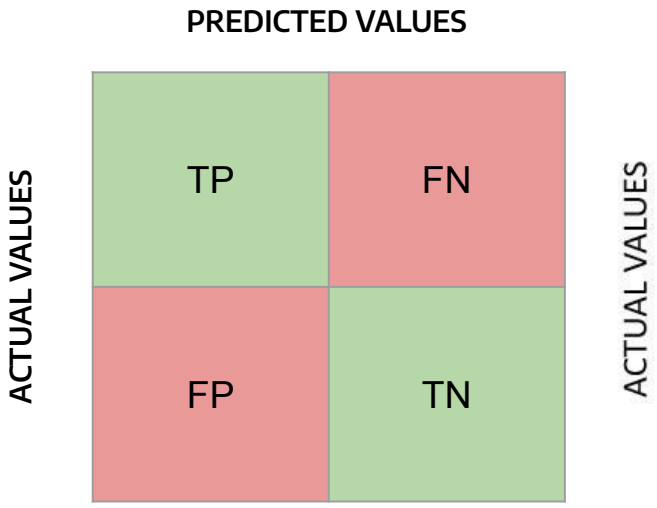
“cat”

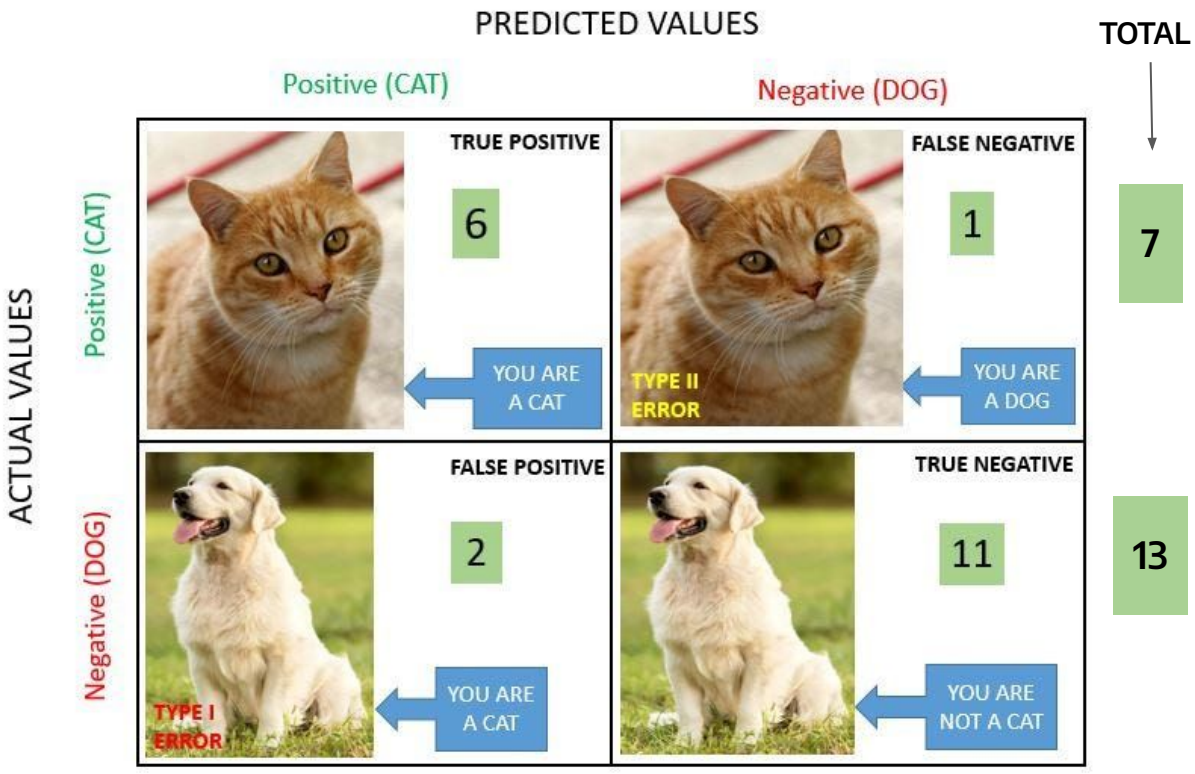
“dog”

Positive | 1

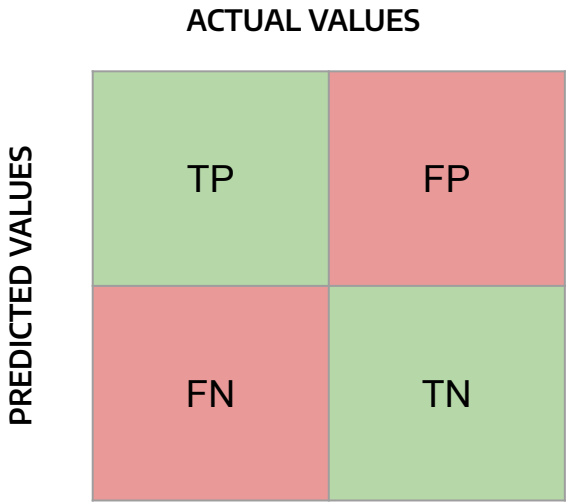
Negative | 0

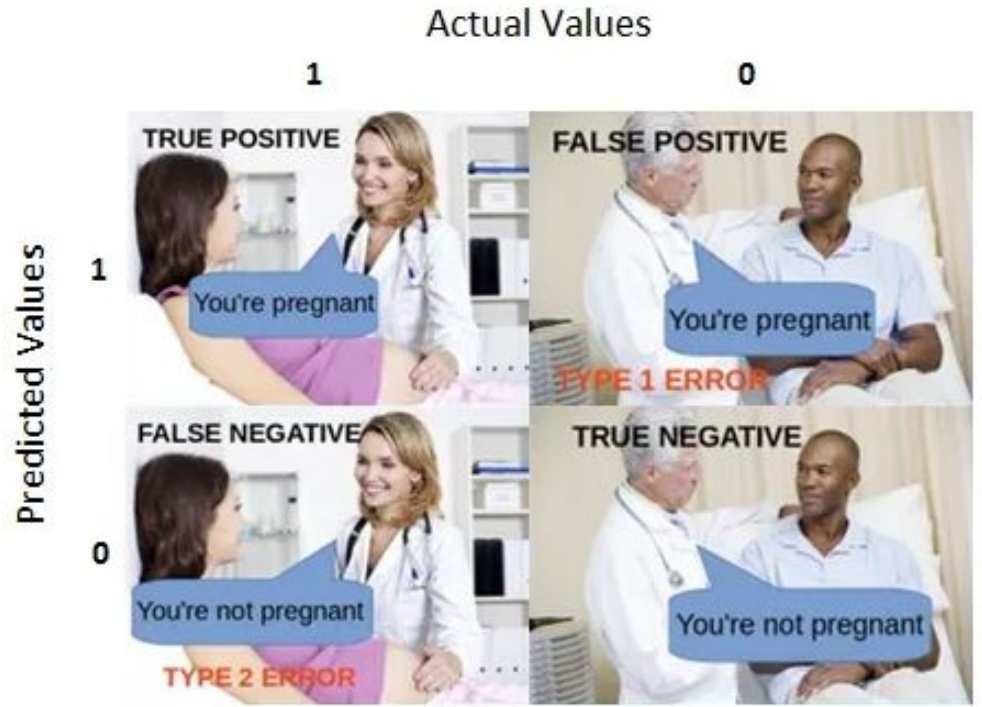
Estructura de una matriz de confusión



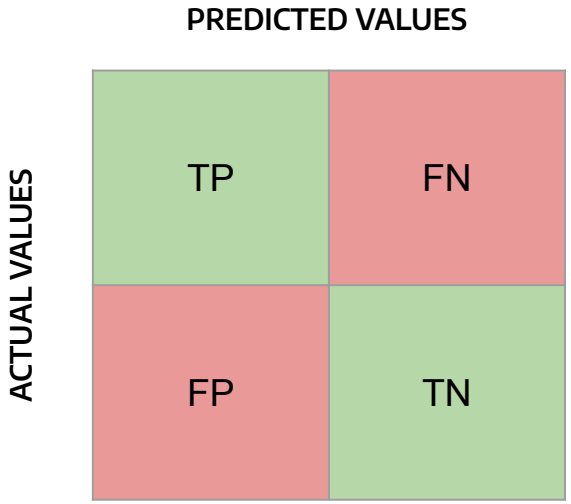


Otro caso fácil (porque meme)





Spam or not spam?



		1	0
	n=200	Predicted to be: SPAM	Predicted to be: NOT SPAM
1	Actually is : SPAM	120	30
0	Actually is : NOT SPAM	10	40

Qué queremos maximizar?

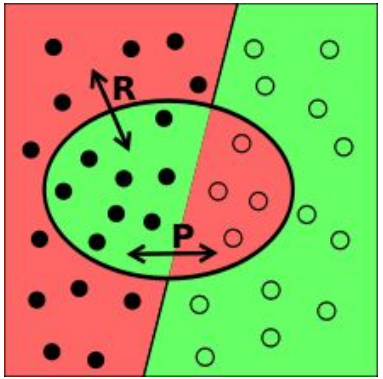
Precision: Fracción de emails importantes entre los emails *que llegaron al Inbox*.

Precision = Positive Predicted Value
PPV = TP / (TP + FP)

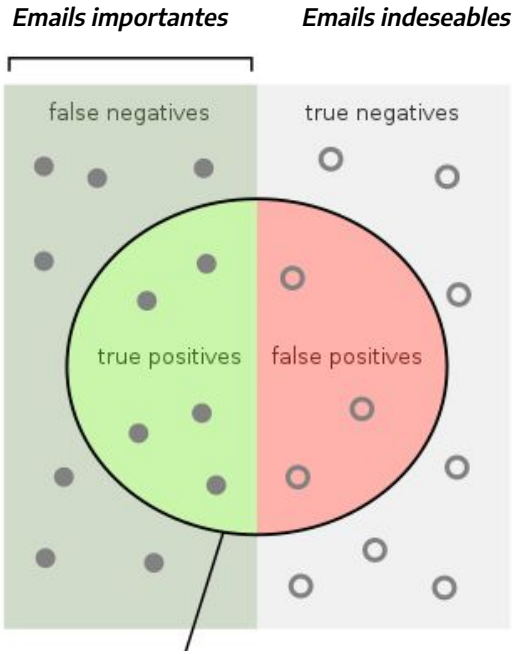
Exhaustividad (Recall): Fracción de *emails importantes que llegaron al inbox*, con respecto a todos los emails importantes.

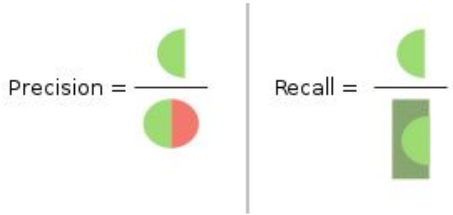
Recall = True Positive Rate
TPR = TP / P = TP / (TP + FN)

Accuracy = Exactitud
ACC = TP + TN / P + N



verde = acierto
rojo = error





https://en.wikipedia.org/wiki/Precision_and_recall

Qué pasa si dejamos entrar todos los emails al Inbox?

Precision: Fracción de emails importantes entre los emails *que llegaron al Inbox*.

Precision = Positive Predicted Value
PPV = TP / (TP + FP) = 12 / 22 = 54% 😐

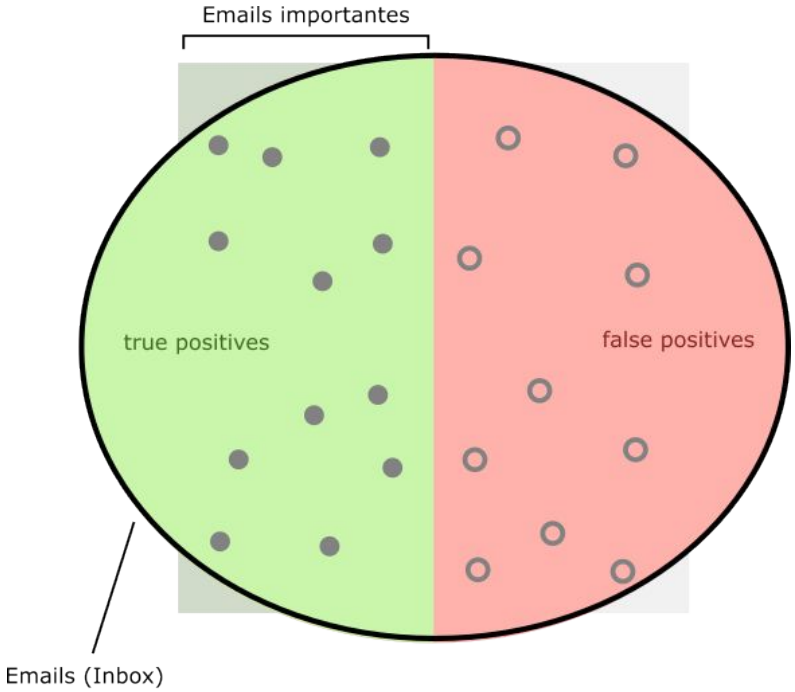
Exhaustividad (Recall): Fracción de *emails importantes que llegaron al inbox*, con respecto a todos los emails importantes.

Recall = True Positive Rate
TPR = TP / P = TP / (TP + FN) = 12 / 12 💯!!

Pero también:
False Positive Rate (Probabilidad de falsa Alarma)
FPR = FP / N = FP / (FP + TN) = 10/10

💯!!

https://en.wikipedia.org/wiki/Precision_and_recall



Qué pasa si dejamos entrar pocos emails al Inbox?

Precision (Especificidad)

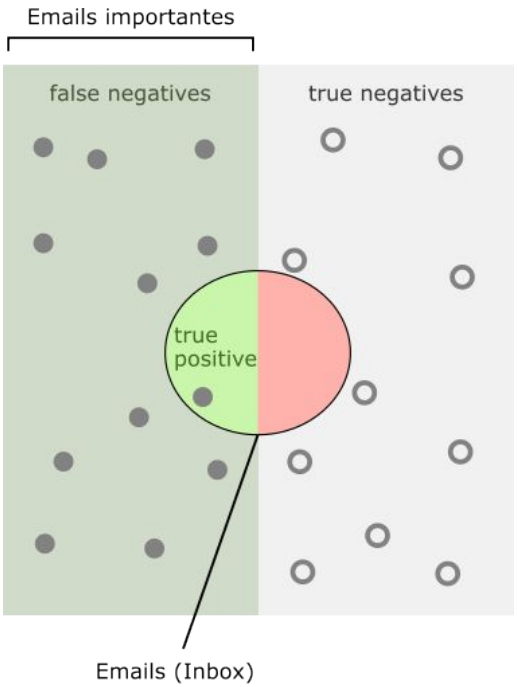
Fracción de emails importantes entre los emails *que llegaron al Inbox*.

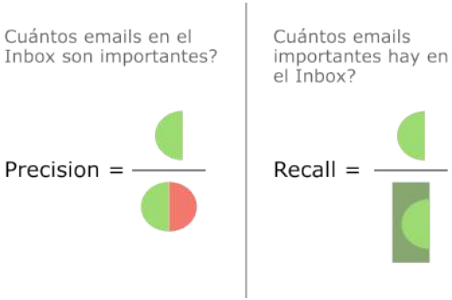
Precision = Positive Predicted Value
PPV = TP / (TP + FP) = 1 / 1 = 100% 💯!!

Exhaustividad (Recall)

Fracción de *emails importantes que llegaron al inbox*, con respecto a todos los emails importantes.

Recall = True Positive Rate
TPR = TP / P = TP / (TP + FN) = 1 / 12 = 8.33% 😞!!





https://en.wikipedia.org/wiki/Precision_and_recall

Hablemos de umbrales

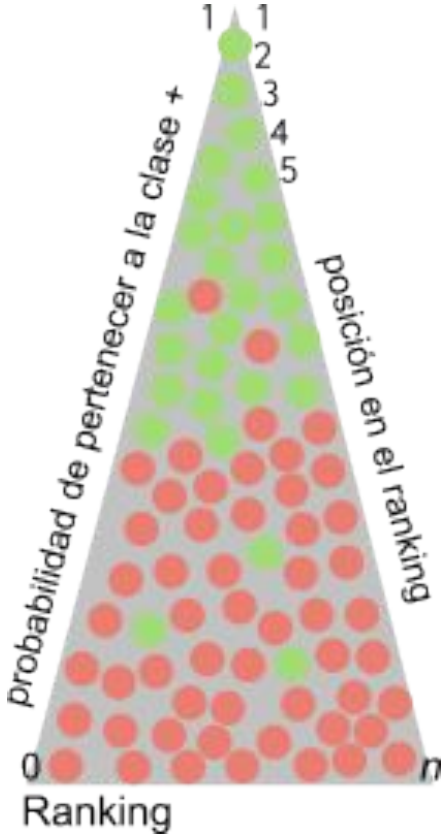
Tenemos datos ***clasificados*** (positivos, negativos)

Cada dato tiene una ***probabilidad de pertenencia*** a su clase

(funciona como un puntaje/score, podemos rankear los datos!)

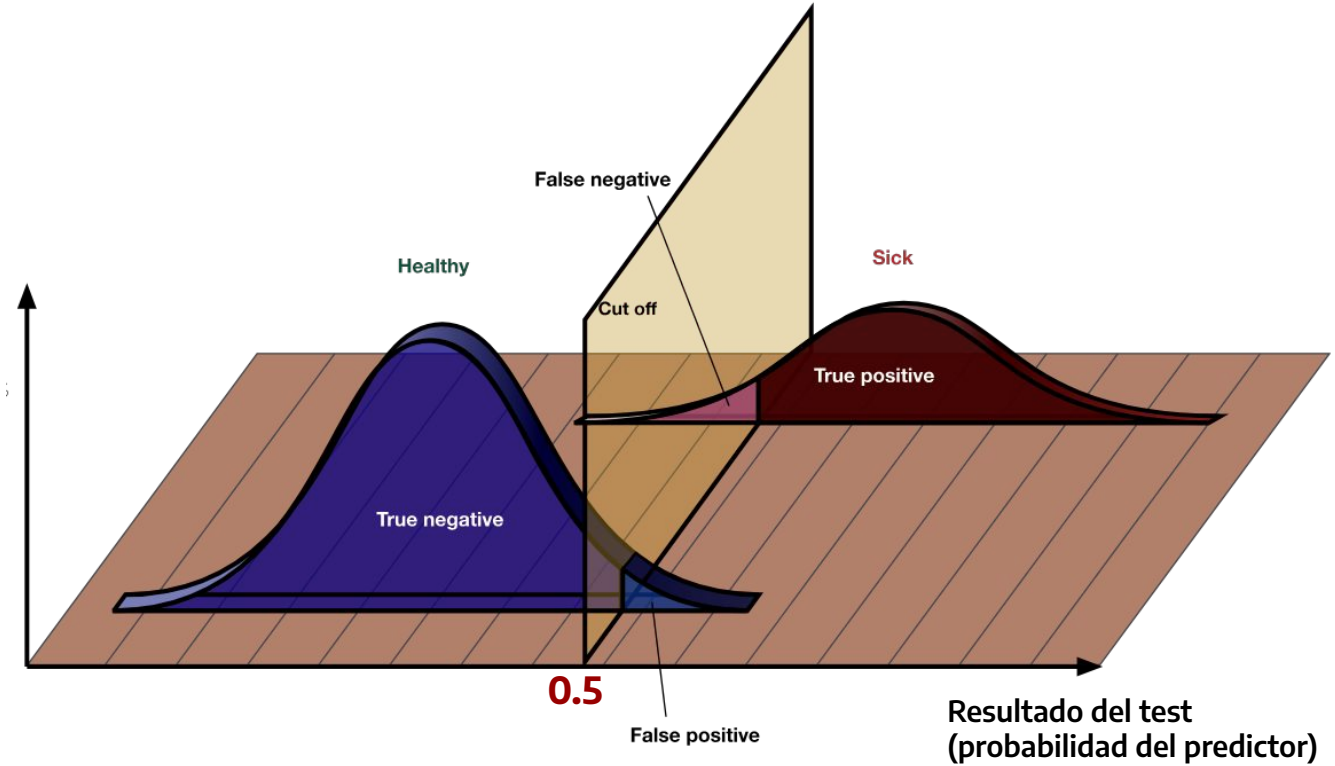
Aunque no lo veamos, hay un umbral!

a. Es lo que define Verdaderos Positivos, Verdaderos Negativos, Falsos Positivos, Falsos Negativos
b. El umbral por defecto es p = 0.5



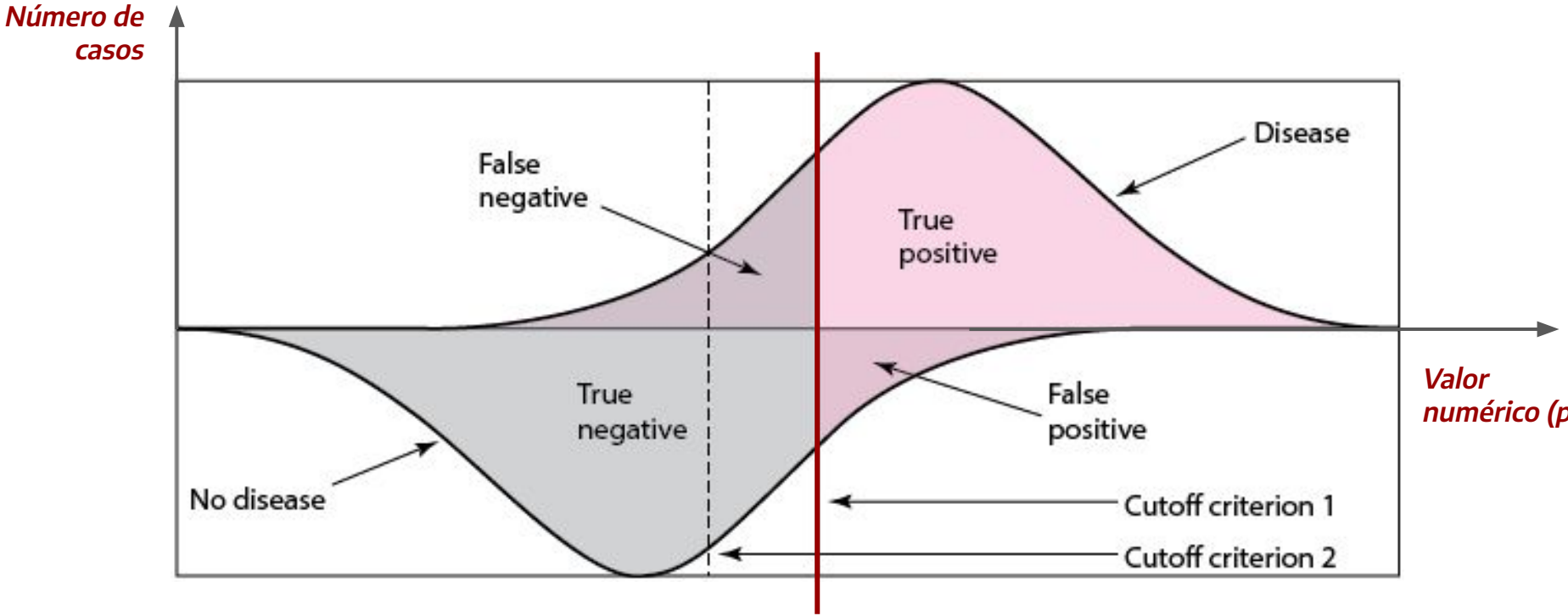
El problema de la predicción

Numero de casos
(pacientes)

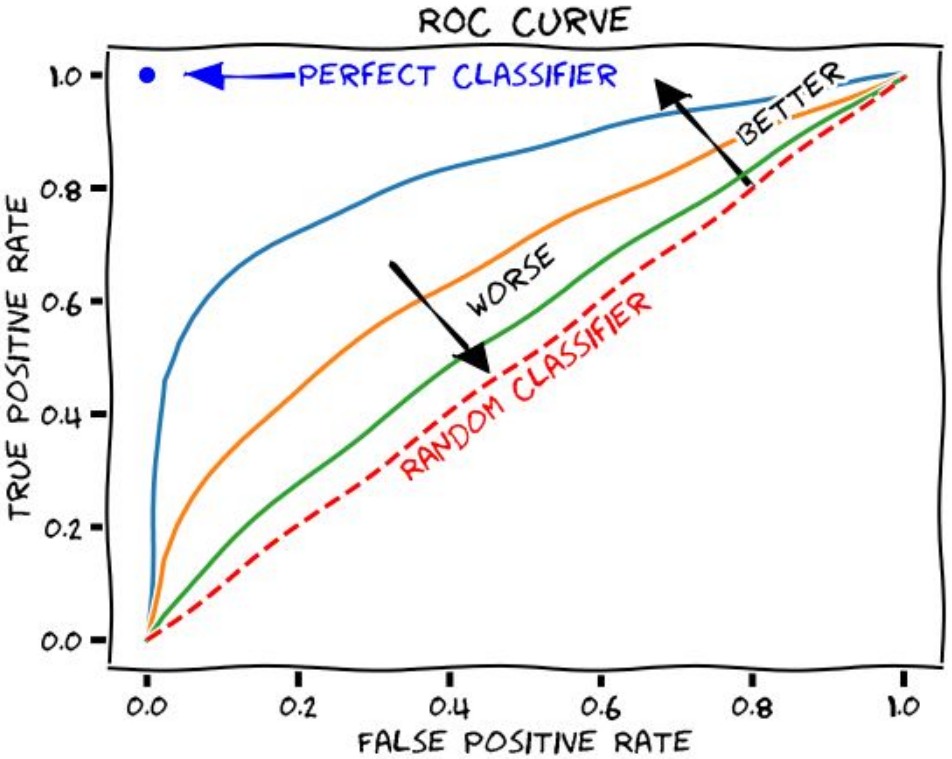


Resultado del test
(probabilidad del predictor)

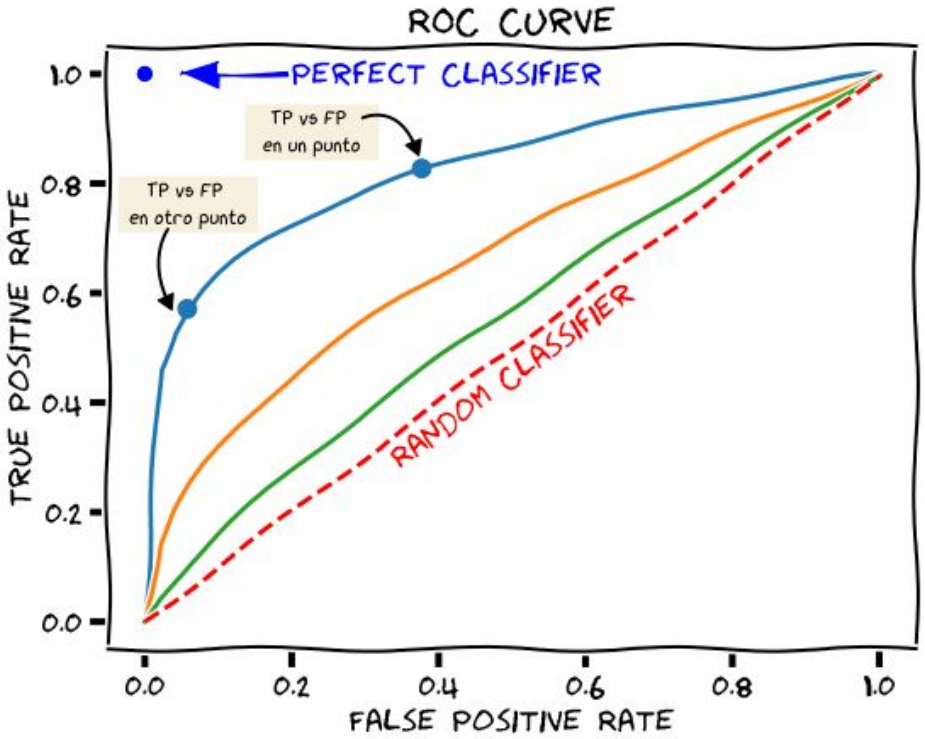
El problema de la predicción (visto de otra manera)

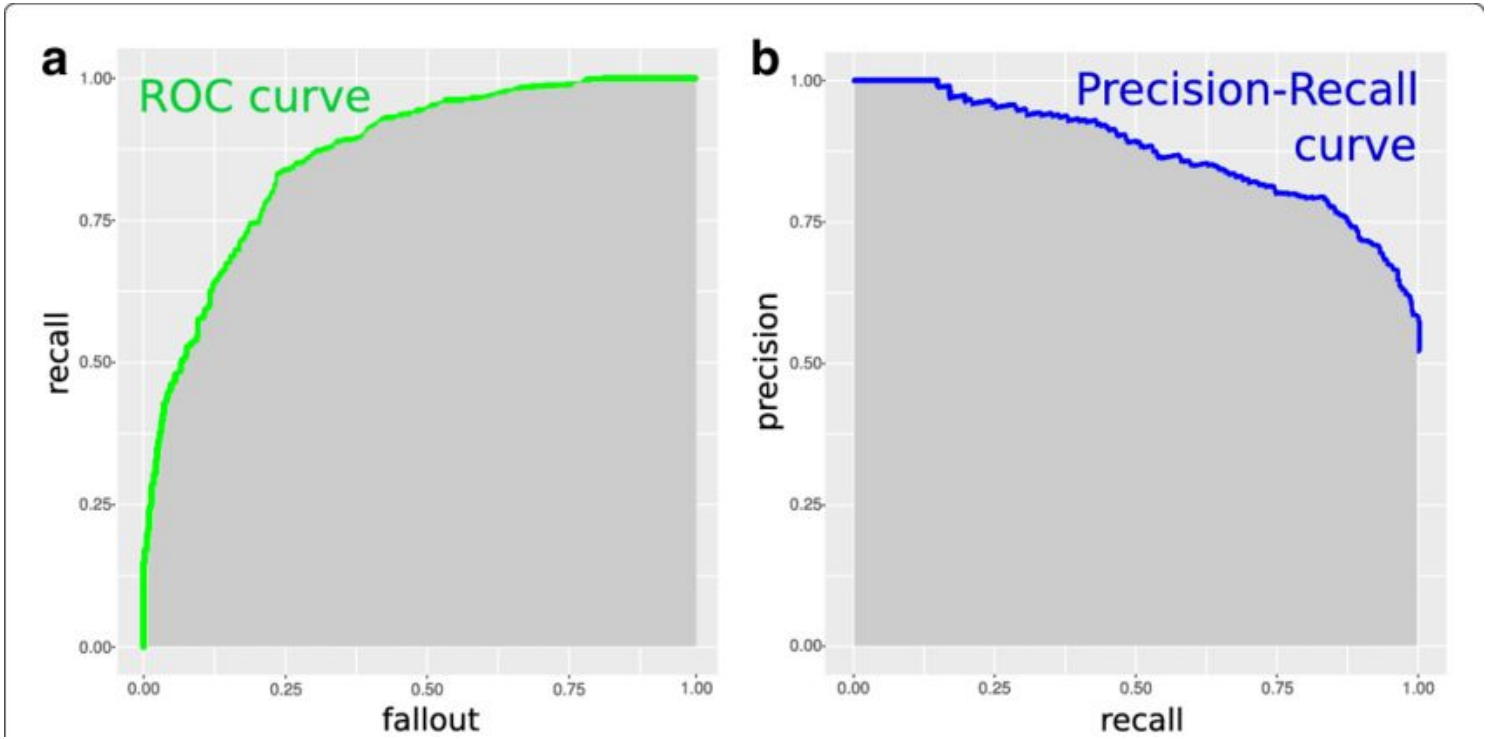


Curvas ROC


ROC CURVE
PERFECT CLASSIFIER
BETTER
WORSE
RANDOM CLASSIFIER
TRUE POSITIVE RATE
FALSE POSITIVE RATE
0.0
0.2
0.4
0.6
0.8
1.0

Como leer una curva ROC




a
ROC curve
recall
fallout
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
b
Precision-Recall curve
precision
recall

		Predicted condition			
	Total population $= P + N$	**Positive (PP)**	**Negative (PN)**	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP),** hit	**False negative (FN),** type II error, miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	Negative (N)	**False positive (FP),** type I error, false alarm, overestimation	**True negative (TN),** correct rejection	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN}$ $= 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$

Sources: [4][5][6][7][8][9][10][11][12] view · talk · edit

Caso difícil: cancer, pacientes (vida, muerte)

Predicción de Pronóstico de progresión de Cancer.