PROBLEMAS DE CLASIFICACIÓN

CLASIFICACIÓN: Cuando apàrece una nueva observación, identificar a qué categoría pertenece.

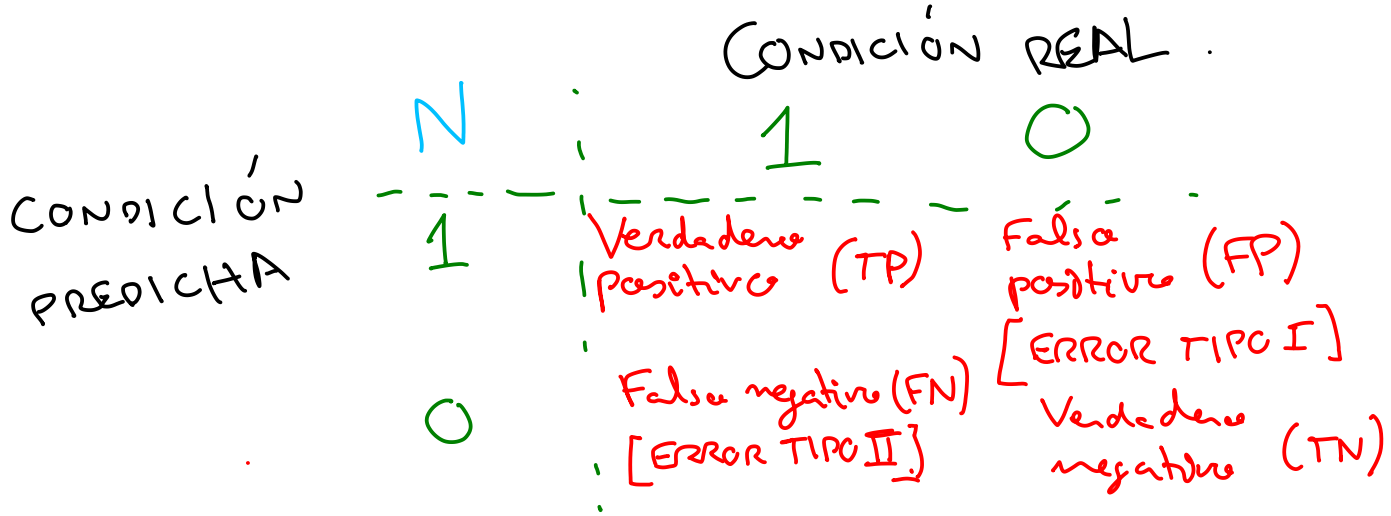
Ejemplo: Nos llega un email. Identificar si es spam o no es spam.

Dada una nueva observación, nos preguntamos si pertenece (1) o no pertenece (0) a una categoría dada.

La predicción de si pertenece o no se da con una probabilidad

Posibles errores:

Matriz de confusión:



Si hacemos predicciones sobre una muestra de N elementos podemos obtener varios parámetros:

Precisión (Accuracy, ACC)

$$ACC = \frac{TP + TN}{N}$$

Tasa de detección (TPR)

$$TPR = \frac{TP}{TP + FN}$$

Tasa de verdaderos negativos (TNR)

$$TNR = \frac{TN}{TN + FP}$$

Tasa de falsa alarma (FPR)

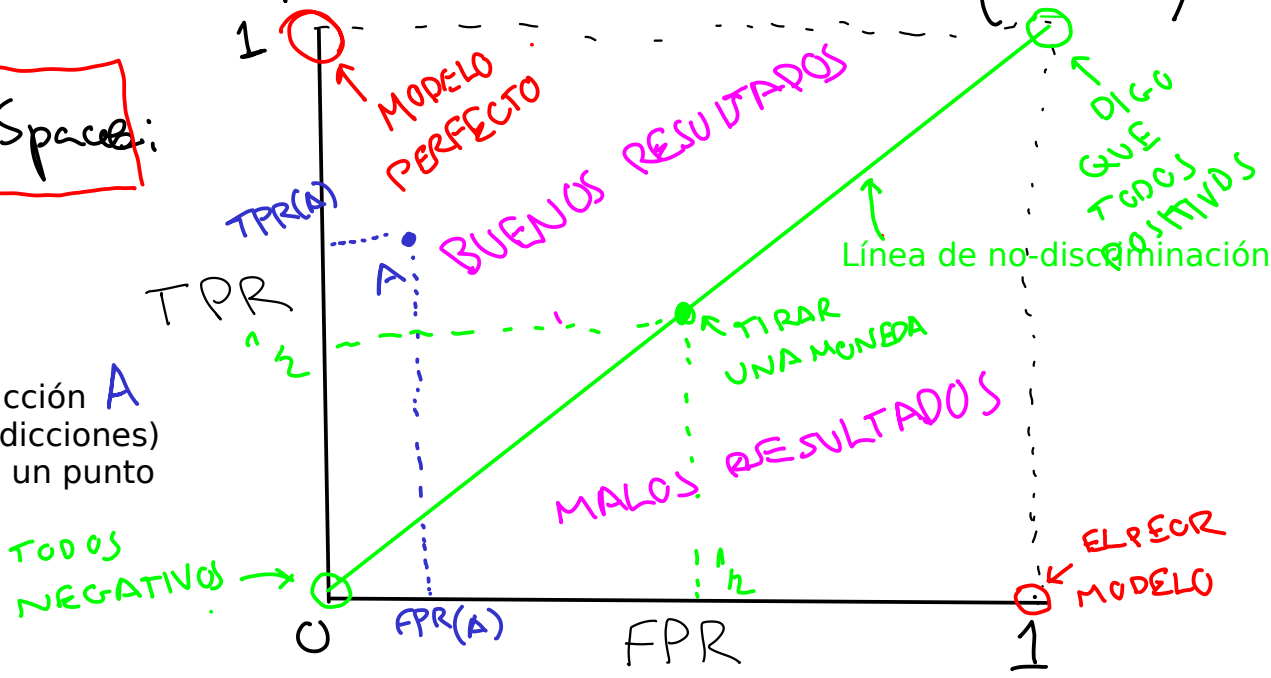
$$FPR = \frac{FP}{FP + TN}$$

		True condition			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

ROC space [edit]

Receiver operator characteristic (ROC)

ROC Space:



Cada posible predicción A (o conjunto de predicciones) puede verse como un punto en el ROC Space.

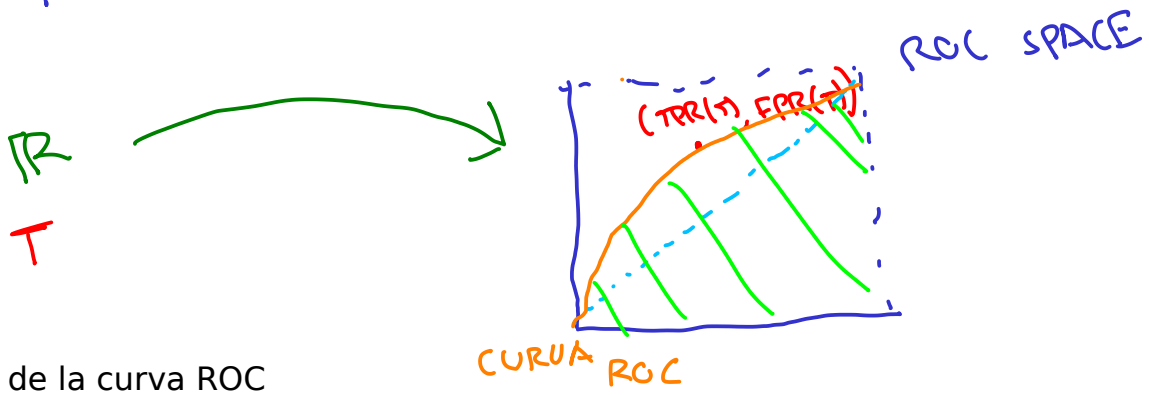
Curva ROC:

En general las predicciones se dan como una variable aleatoria continua X de forma que

$$TPR(T) = \int_T^{\infty} f_1(x)\,dx$$

$$FPR(T) = \int_T^{\infty} f_0(x)\,dx$$

$$\begin{cases} X > T \rightarrow 1 \leftarrow f_1 \\ X \leq T \rightarrow 0 \leftarrow f_0 \end{cases}$$



El área debajo de la curva ROC informa de lo bueno que es el modelo predictivo:

A>0.5 BIEN
A=0.5 RANDOM
A<0.5 MAL

REGRESIÓN LOGISTICA

Odds ratio

Ejemplo:

Participación en una huelga por sexo:

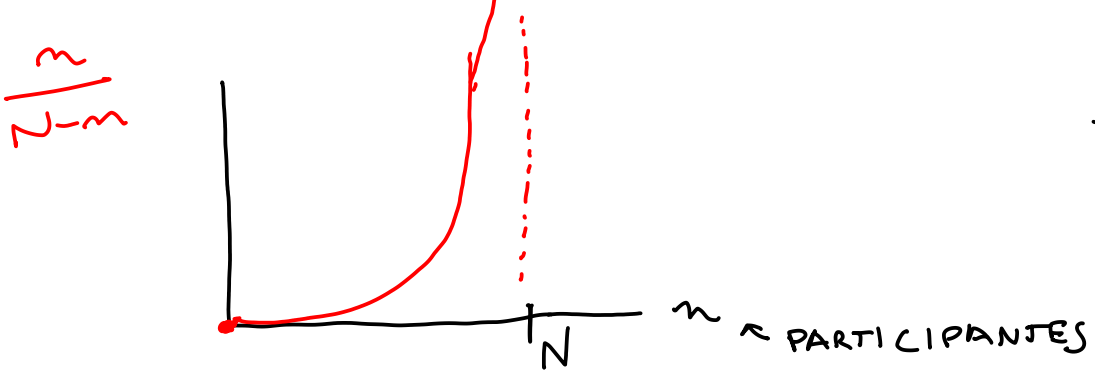
	HOMBRE	MUJER	TOTAL
NO	960	1057	2017
SÍ	261	206	467
	1221	1263	2484

Porcentaje de participación:

$467/2484 = 18,8\,\%$

Odd (los que sí, frente a los que no):

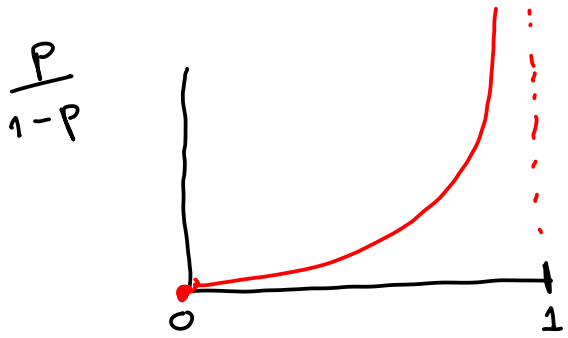
$467/2017 = 0,232$



$\frac{n}{N-n}$ PROB. PARTICIPA

N, $p = \frac{n}{N}$

$\frac{p}{1-p}$



Podemos sacar los ODDs por separado:

	HOMBRE	MUJER
ODD(participa/no participa)	0,272	0,195

ODD RATIO: $OR = \frac{ODD_A}{ODD_B} = \frac{\frac{p_A}{1-p_A}}{\frac{p_B}{1-p_B}}$

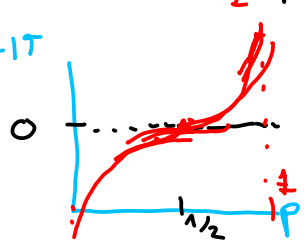
Para estudiar los odds, nos interesa considerar la función LOGIT

$(\ln)$

$ODD(p) = \frac{p}{1-p}$ $\quad Logit(p) = \log(ODD) = \log\left(\frac{p}{1-p}\right)$

$ODD = p/q \longrightarrow logit = \ln(p/q)$

$ODD = q/p \longrightarrow logit = \ln(q/p) = -\ln(p/q)$



Volviendo al caso anterior:

	HOMBRE	MUJER	TOTAL
NO	0,7862	0,8368	0,8119
SÍ	0,2137	0,1631	0,1884
	1	1	1
ODD	0,272	0,195	
LOGIT	-1,30	-1,63	

Regresión logística

Training set

Test

↓

Certeza

p ← PROB. DE QUE PERTENEZCA

Datos, que pertenecen o no pertenecen a una variable categórica

$$y = \begin{cases} 1 & \text{si pertenece} \\ 0 & \text{si no pertenece} \end{cases}$$

X = Variables que usamos para predecir

Idea: Hacer un ajuste lineal a Logit(p)

$$\text{logit}(p) = \alpha + \beta x \rightsquigarrow p = \frac{1}{1 + e^{-(\alpha+\beta x)}}$$

En reg. lineal	En reg. logística
$y = \alpha + \beta x$	$p = \frac{1}{1+e^{-(\alpha+\beta x)}}$

Función logística: $\sigma(t) = \frac{1}{1+e^{-t}}$

En el ejemplo anterior:

Mi target es estimar p, la probabilidad de hacer huelga.

Mi feature es $X = \begin{cases} 0 & \text{si hombre} \\ 1 & \text{si mujer} \end{cases}$

$$\text{logit}(p) = \alpha + \beta x$$

$x = 0 \rightarrow \text{logit}(\text{HOMBRE}) = \alpha \Rightarrow \alpha = -1{,}30$

$\text{logit}(\text{HOMBRE}) = -1{,}30$

$x = 1 \rightarrow \text{logit}(\text{MUJER}) = \alpha + \beta \Rightarrow \beta = -0{,}33$

$\text{logit}(\text{MUJER}) = -1{,}63$

$$p = \frac{1}{1 + e^{1{,}30 + 0{,}33x}}$$

$p(0) = 0{,}214$

$p(1) = 0{,}163$

OBSERVACIÓN:

$e^{\alpha} = 0{,}272 = \text{ODD}_{\text{HOMBRES}}$

$e^{\alpha+\beta} = \text{ODD}_{\text{MUJERES}}$

$e^{\beta} = \text{ODD}_M / \text{ODD}_H = \text{OR}$

EJERCICIO:

Participación en la huelga por nivel de estudios

	Básicos	Medios	Universitarios
NO	1116	554	344
SI	138	182	145
	1254	736	489

Se pide: Hallar los ODDS, los logit y los parámetros de la regresión logística.

	x_1	x_2
BÁSICOS	0	0
MEDIOS	1	0
UNIVERSITARIOS	0	1

El ajuste es a:

$$p = \frac{1}{1 + e^{(-\alpha + \beta_1 x_1 + \beta_2 x_2)}}$$

- Función de coste:

La regresión logística tiene la forma:

$$p = \frac{1}{1 + e^{-\vec{\theta}\cdot\vec{x} - \vec{\eta}}}$$

$\vec{\theta} = (\theta_1, \dots, \theta_n)$

$\vec{x} = (x_1, \dots, x_n)$

$\vec{\theta}\cdot\vec{x} = \theta_1 x_1 + \dots + \theta_n x_n$

$\vec{\eta} = (\eta_1, \dots, \eta_n)$

En reg. lineal eran los mínimos cuadrados

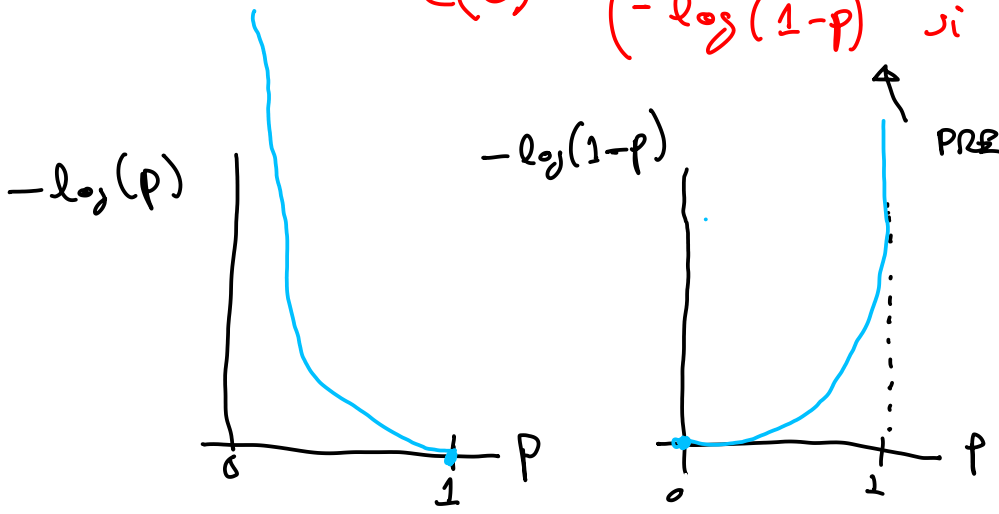
Los a y b del ajuste son los que minimizan c(a,b)

• Para una: θ, x

PREMIO LAS PROBABILIDADES ALTAS

$$c(\theta) = \begin{cases} -\log p & \text{si } y = 1 \\ -\log(1-p) & \text{si } y = 0 \end{cases}$$

PREMIO LAS PROBABILIDADES BAJAS



Para $\vec{x}, \vec{\theta}$, promedio:

$\vec{y} = (y_1, \dots, y_m)$

$$J(\vec{\theta}) = -\frac{1}{m} \sum_{i=1}^{m} \left[y_i \log(p_i) + (1-y_i) \log(1-p_i) \right]$$

ES CONVEXA → Los algoritmos de optimización (p. ej. el "gradient descent") hallan el MÍNIMO GLOBAL.