uc3m | Universidad Carlos III de Madrid

OPENCOURSEWARE
APRENDIZAJE AUTOMÁTICO PARA EL ANÁLISIS DE DATOS
GRADO EN ESTADÍSTICA Y EMPRESA
Ricardo Aler

CLASIFICACIÓN CON MUESTRAS DESBALANCEADAS

ORGANIZACIÓN

- **Evaluación** con muestras desbalanceadas
  - Matriz de confusión
  - La métrica AUC: el área bajo la curva ROC
- **Aprendizaje** con muestras desbalanceadas

Usar particiones estratificadas

- Tanto en train/test (holdout) como en validación cruzada.
- Para que el test sea mas representativo
- La distribución de clases que existe en el conjunto de datos disponibles, se intenta mantener en los conjuntos de train y test (o en cada uno de los folds de validación cruzada)
  - Ejemplo: si en el conjunto original un 99% de los datos pertenecen a la clase negativa y 1% a la positiva, la estratificación intentará que esa proporción se mantenga en train y test

EVALUACIÓN CON MUESTRAS DESBALANCEADAS

- Ya sabemos que usar el error de clasificación (o accuracy = 1 – error) es engañoso para muestras desbalanceadas.
- Ejemplo:
  - Tenemos varios clasificadores c_1, c_2, c_3 que predicen si se ha de abrir o cerrar la válvula del módulo de refrigeración de una central nuclear.
  - Para evaluar los clasificadores usamos un conjunto de datos obtenido en el último mes, donde un operario ha decidido en cada momento si se había de abrir o cerrar la fórmula.
    - 100.000 ejemplos, de los cuales 99.500 son de la clase “Cerrar” y 500 son de la clase “Abrir”.
  - Supongamos que el clasificador c_2 obtiene una tasa de aciertos (accuracy) del 99.5%, que es muy alta, pero es lo que puede obtener un clasificador trivial (clase mayoritaria)
  - Además, tenemos un 100% de fallos en la clase “Cerrar” (que en este caso es la más importante).

EVALUACIÓN CON MUESTRAS DESBALANCEADAS

- En el ejemplo anterior vemos que un único número (accuracy) no es suficiente para darnos una información correcta, cuando tenemos distribuciones desbalanceadas

- Matriz de confusión/contingencia (p.ej. para el conjunto de test):

Real

Predicho	abrir (p)	cerrar (n)
ABRIR (P)	TP	FP
CERRAR (N)	FN	TN

Diagonal de los aciertos

Real

Pred	c_1	abrir	cerrar
	ABRIR	300	500
	CERRAR	200	99000

Hay un total de 100000 datos, 500 positivos y 99500 negativos

EVALUACIÓN CON MUESTRAS DESBALANCEADAS

- En el ejemplo anterior vemos que un único número (accuracy) no es suficiente para darnos una información correcta, cuando tenemos distribuciones desbalanceadas

- Matriz de confusión/contingencia (p.ej. para el conjunto de test):

Real

Predicho	abrir (p)	cerrar (n)
ABRIR (P)	TP	FP
CERRAR (N)	FN	TN

Diagonal de los errores

Real

Pred	c_1	abrir	cerrar
	ABRIR	300	500
	CERRAR	200	99000

Hay un total de 100000 datos, 500 positivos y 99500 negativos

EVALUACIÓN CON MUESTRAS DESBALANCEADAS

- Normalmente, a la matriz de confusión se la normaliza por columnas

Real

	abrir (p) tpos=TP+FN	cerrar (n) tneg=FP+TN
Predicho: ABRIR (P)	TPR=TP/tpos	FPR=FP/tneg
Predicho: CERRAR (N)	FNR=FN/tpos	TNR=TN/tneg

Normalizada por columnas

Sin normalizar

Real

Pred c_1	abrir	cerrar
ABRIR	300	500
CERRAR	200	99000

tpos = 300+200 = 500
tneg = 500+99000 = 99500

Normalizada

Real

Pred c_1	abrir	cerrar
ABRIR	0.60	0,005
CERRAR	0.40	0.995

EVALUACIÓN CON MUESTRAS DESBALANCEADAS

- Ejemplo: (conjunto de test de 100.000 instancias)

Real

Pred	c_1	abrir	cerrar
	ABRIR	300	500
	CERRAR	200	99000

tpos = 300+200 = 500, tneg = 500+99000 = 99500

ACC: 0.993 = (300+99000)/100000

TPR= 300 / 500 = 0.60
FNR= 200 / 500 = 0.40
TNR= 99000 / 99500 = 0.995
FPR= 500 / 99500 = 0.005

BAC = macromedia = (0.60 + 0.995) / 2 = 0.80

Real

Pred	c_1	abrir	cerrar
	ABRIR	0.60	0,005
	CERRAR	0.40	0.995

EVALUACIÓN CON MUESTRAS DESBALANCEADAS

pos = 300+200 = 500, neg = 500+99000 = 99500

- Ejemplo: (conjunto de test de 100.000 instancias)

Real

Pred

c_1	abrir	cerrar
ABRIR	300	500
CERRAR	200	99000

ACC: 0.993 = (300+99000)/100000

TPR= 300 / 500 = 0.60
FNR= 200 / 500 = 0.40
TNR= 99000 / 99500 = 0.995
FPR= 500 / 99500 = 0.005

BAC = macromedia = (0.60 + 0.995) / 2 = 0.80

Real

c_2	abrir	cerrar
ABRIR	0	0
CERRAR	500	99500

ACC: 0.995 = (0+99500)/100000

TPR= 0 / 500 = 0.0
FNR= 500 / 500 = 1.0
TNR= 99500 / 99500 = 1.0
FPR= 0 / 99500 = 0.0

BAC = Macromedia= (0 + 1) / 2 = 0.5

Real

Pred

c_1	abrir	cerrar
ABRIR	0.60	0,005
CERRAR	0.40	0.995

Real

c_2	abrir	cerrar
ABRIR	0.0	0.0
CERRAR	1.0	1.0

EVALUACIÓN CON MUESTRAS DESBALANCEADAS

pos = 300+200 = 500, neg = 500+99000 = 99500

- Ejemplo: (conjunto de test de 100.000 instancias)

Real

Pred	c_1	abrir	cerrar
	ABRIR	300	500
	CERRAR	200	99000

ACC: 0.993 = (300+99000)/100000

TPR= 300 / 500 = 0.60
FNR= 200 / 500 = 0.40
TNR= 99000 / 99500 = 0.995
FPR= 500 / 99500 = 0.005

BAC = macromedia = (0.60 + 0.995) / 2 = 0.80
= balanced accuracy

Real

c_2	abrir	cerrar
ABRIR	0	0
CERRAR	500	99500

ACC: 0.995 = (0+99500)/100000

TPR= 0 / 500 = 0.0
FNR= 500 / 500 = 1.0
TNR= 99500 / 99500 = 1.0
FPR= 0 / 99500 = 0.0

BAC = Macromedia= (0 + 1) / 2 = 0.5

Real

c_3	abrir	cerrar
ABRIR	400	5400
CERRAR	100	94100

ACC: 0.945 = (400+94100)/100000

TPR= 400 / 500 = 0.80
FNR= 100 / 500 = 0.20
TNR= 94100 / 99500 = 0.946
FPR= 5400 / 99500 = 0.054

Macromedia= (0.80 + 0.946) / 2 = 0.873

Real

Pred	c_1	abrir	cerrar
	ABRIR	0.60	0,005
	CERRAR	0.40	0.995

Real

c_2	abrir	cerrar
ABRIR	0.0	0.0
CERRAR	1.0	1.0

Real

c_3	abrir	cerrar
ABRIR	0.80	0.054
CERRAR	0.20	0.946

EVALUACIÓN CON MUESTRAS DESBALANCEADAS

pos = 300+200 = 500, neg = 500+99000 = 99500

- Ejemplo: (conjunto de test de 100.000 instancias)

Real

Pred	c_1	abrir	cerrar
	ABRIR	300	500
	CERRAR	200	99000

ACC: 0.993 = (300+99000)/100000
BAC = macromedia = (0.60 + 0.995) / 2 = 0.80

Real

Pred	c_2	abrir	cerrar
	ABRIR	0	0
	CERRAR	500	99500

ACC: 0.995 = (0+99500)/100000
BAC = Macromedia= (0 + 1) / 2 = 0.5

Real

Pred	c_3	abrir	cerrar
	ABRIR	400	5400
	CERRAR	100	94100

ACC: 0.945 = (400+94100)/100000
Macromedia= (0.80 + 0.946) / 2 = 0.873

Real

Pred	c_1	abrir	cerrar
	ABRIR	0.60	0,005
	CERRAR	0.40	0.995

Real

Pred	c_2	abrir	cerrar
	ABRIR	0.0	0.0
	CERRAR	1.0	1.0

Real

Pred	c_3	abrir	cerrar
	ABRIR	0.80	0.054
	CERRAR	0.20	0.946

¿Qué clasificador es mejor?

NOTA

- Medidas que combinan los aciertos en los positivos y en los negativos. Todas están entre cero y uno.
  - Balanced accuracy (BAC) o macro-media: (TPR+TNR)/2
  - Youden's J index: TPR+TNR-1 = BAC*2 - 1
    - (equivalente a BAC)
  - F1-score: 2*TP / (2*TP+FN+FP)

EVALUACIÓN VS. APRENDIZAJE

- Evaluación vs. Aprendizaje
- Ya sabemos cómo evaluar un modelo con muestras desbalanceadas (matriz de confusión)
- Pero el modelo se sigue construyendo según las técnicas vistas hasta ahora: no hay garantía de que un método construya un modelo que optimice la Macromedia o el TPR.
- Vamos por tanto a tratar el caso de aprender buenos modelos cuando hay muestras desbalanceadas.

APRENDIZAJE CON MUESTRAS DESBALANCEADAS

- Los algoritmos de aprendizaje típicos tienden a maximizar la tasa de aciertos (accuracy)
- En problemas de muestra desbalanceada eso suele equivaler a aprender bien la clase mayoritaria a costa de aprender mal la minoritaria
- Es decir, los métodos vistos hasta ahora, tienden a aprender modelos que resultan en matrices de confusión como las siguientes:

		Real	
	c_1	abrir	cerrar
Pred	ABRIR	**0.60**	0,005
	CERRAR	0.40	**0.995**

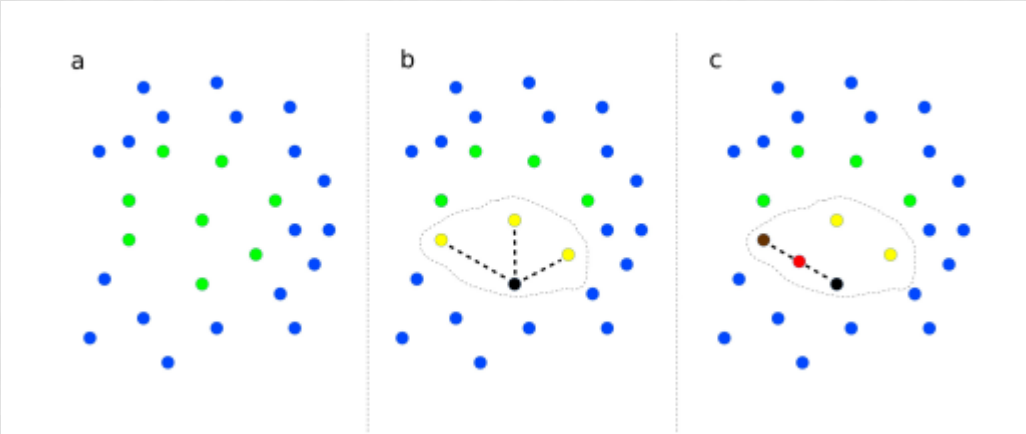
	Real	
c_2	abrir	cerrar
ABRIR	**0.0**	0.0
CERRAR	1.0	**1.0**

APRENDIZAJE CON MUESTRAS DESBALANCEADAS

- Solución 1: entrenar varios modelos (p. ej. con distintos algoritmos) y seleccionar aquel que aprenda razonablemente bien la clase minoritaria (o que maximice el BAC).
  - Es costoso en tiempo y el que aparezca un buen modelo es cuestión de casualidad
- Solución 2: remuestreo:
  - Submuestreo: eliminar datos de la clase mayoritaria para equilibrarla con la minoritaria
  - Sobremuestreo: replicar datos de la clase minoritaria. También se pueden utilizar pesos para las instancias, como se hacía en boosting
  - SMOTE: Synthetic Minority Over-sampling Technique
- Solución 3: thresholding (cambiar el threshold o punto de corte, del modelo)

APRENDIZAJE CON MUESTRAS DESBALANCEADAS

- SMOTE: Synthetic Minority Over-sampling Technique:
  - Se generan instancias situadas entre instancias de la clase minoritaria
  - Hiper-parámetros:
    - ¿Cuántos vecinos?
    - ¿Cuántas muestras de la clase minoritaria hay que generar?



Source: https://www.researchgate.net/publication/317489171_Imbalance-Aware_Machine_Learning_for_Predicting_Rare_and_Common_Disease-Associated_Non-Coding_Variants

APRENDIZAJE CON MUESTRAS DESBALANCEADAS

- La representación ROC de un clasificador discreto es un punto en el espacio ROC

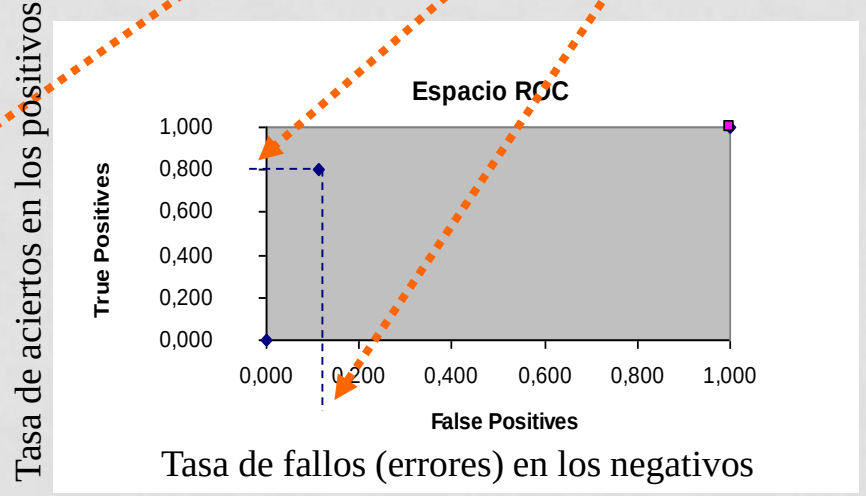
Real

Pred	abrir	cerrar
ABRIR	400	12000
CERRAR	100	87500

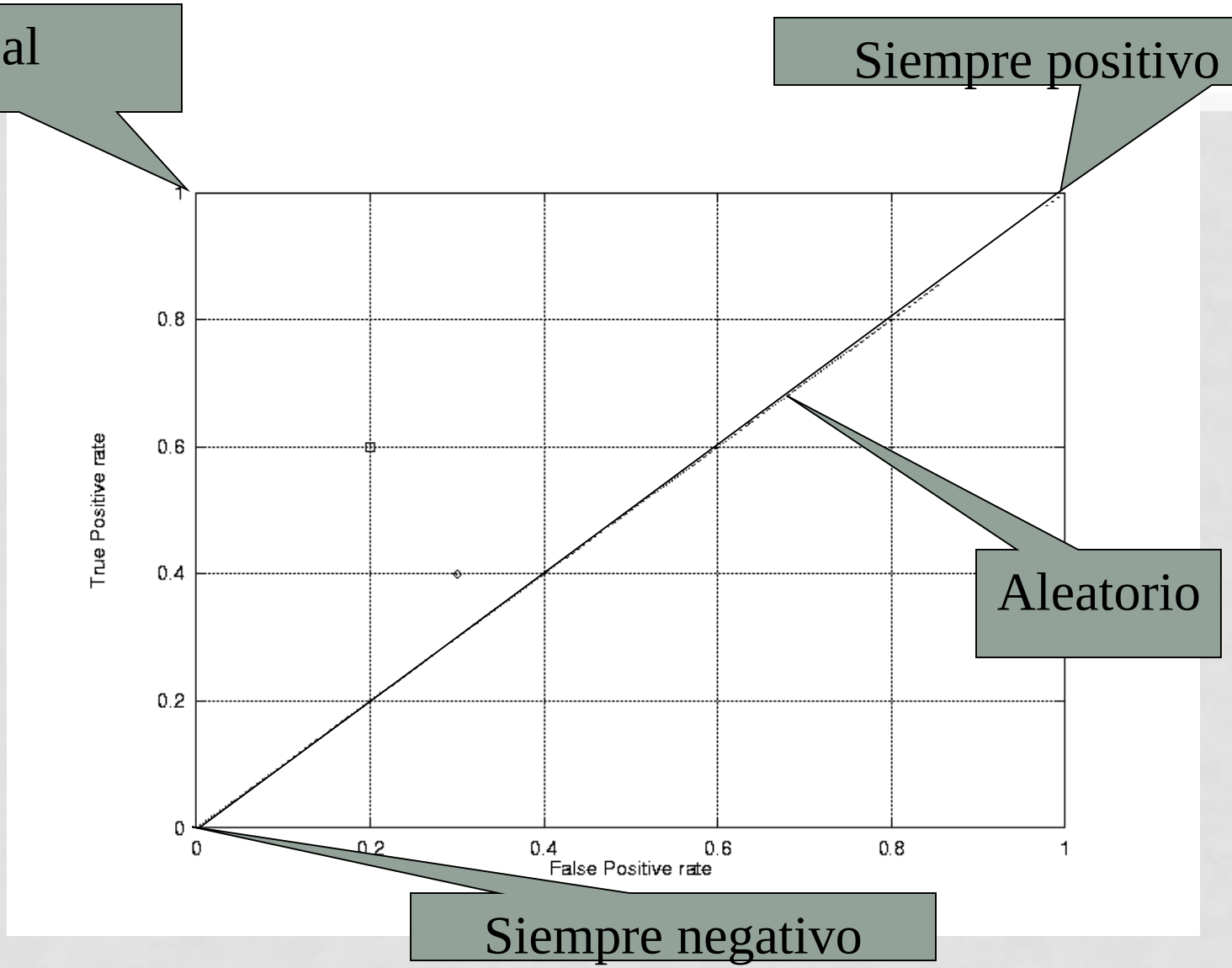
Real

Pred	abrir	cerrar
ABRIR	0,8	0,121
CERRAR	0,2	0,879

TPR= 400 / 500 = 0.80
FNR= 100 / 500 = 0.20
TNR= 87500 / 99500 = 0.879
FPR= 12000 / 99500 = 0.121



ESPACIO ROC


Ideal
Siempre positivo
Aleatorio
Siempre negativo
True Positive rate
False Positive rate
0
0.2
0.4
0.6
0.8
1

¿PORQUÉ LA DIAGONAL REPRESENTA UN CLASIFICADOR ALEATORIO?

- TPR = FPR
- Supongamos que un clasificador clasifica **aleatoriamente** los datos como:
  - positivo el 50% de las veces
  - negativo el 50% de las veces
  - Por pura casualidad, acertará con el 50% de los positivos y fallará con el 50% de los negativos
    - TPR = Pr(P | p) = 0.5; FPR = Pr(P | n) = 0.5

¿PORQUÉ LA DIAGONAL REPRESENTA UN CLASIFICADOR ALEATORIO?

- TPR = FPR
- Supongamos que un clasificador clasifica **aleatoriamente** los datos como:
  - positivo el 90% de las veces
  - negativo el 10% de las veces
  - Por pura casualidad, acertará con el 90% de los positivos y fallará con el 90% de los negativos
    - TPR = Pr(P | p)= 0.9; FPR = Pr(P | n)= 1-Pr(N | n) =0.9

CLASIFICADOR DISCRETO VS. SCORING CLASSIFIER (SC)

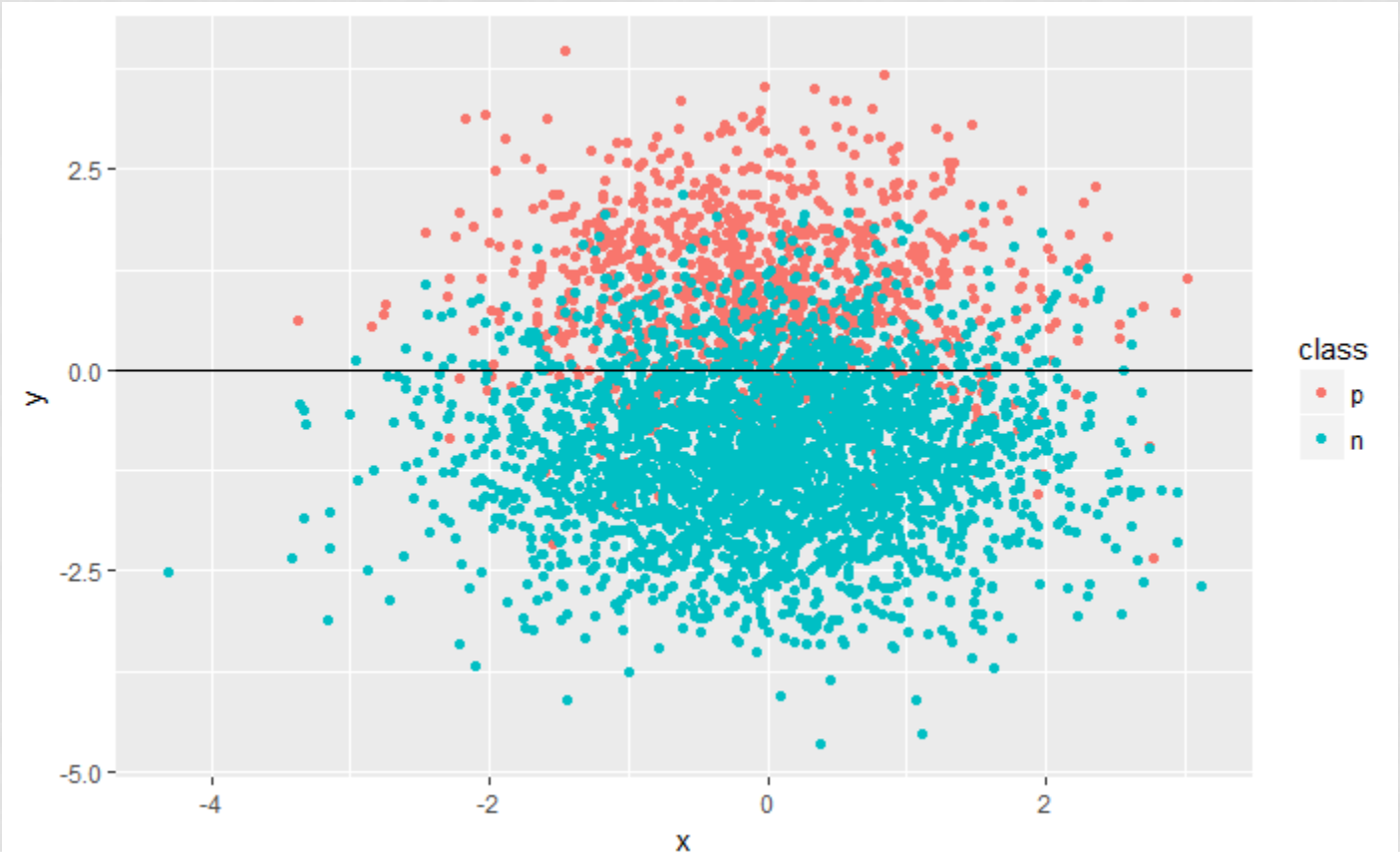
- Un clasificador discreto predice una clase entre las posibles.
- Un scoring classifier (*sc*) predice una clase, pero acompaña un valor de fiabilidad a cada predicción.
- Un ejemplo de *sc* es Random Forest, que puede estimar probabilidades
- Pero existen *sc* que no son capaces de devolver probabilidades, pero si valores (scores g(x)) que indican la certidumbre de que un dato pertenezca a una clase
- Por ejemplo, un clasificador lineal, o una máquina de vectores de soporte, que devuelven una distancia del dato a la frontera:
  - Si el valor es muy negativo, cercano a la clase 0
  - Si el valor es muy positivo, cercano a la clase 1

CLASIFICADOR DISCRETO VS. SCORING CLASSIFIER (SC)

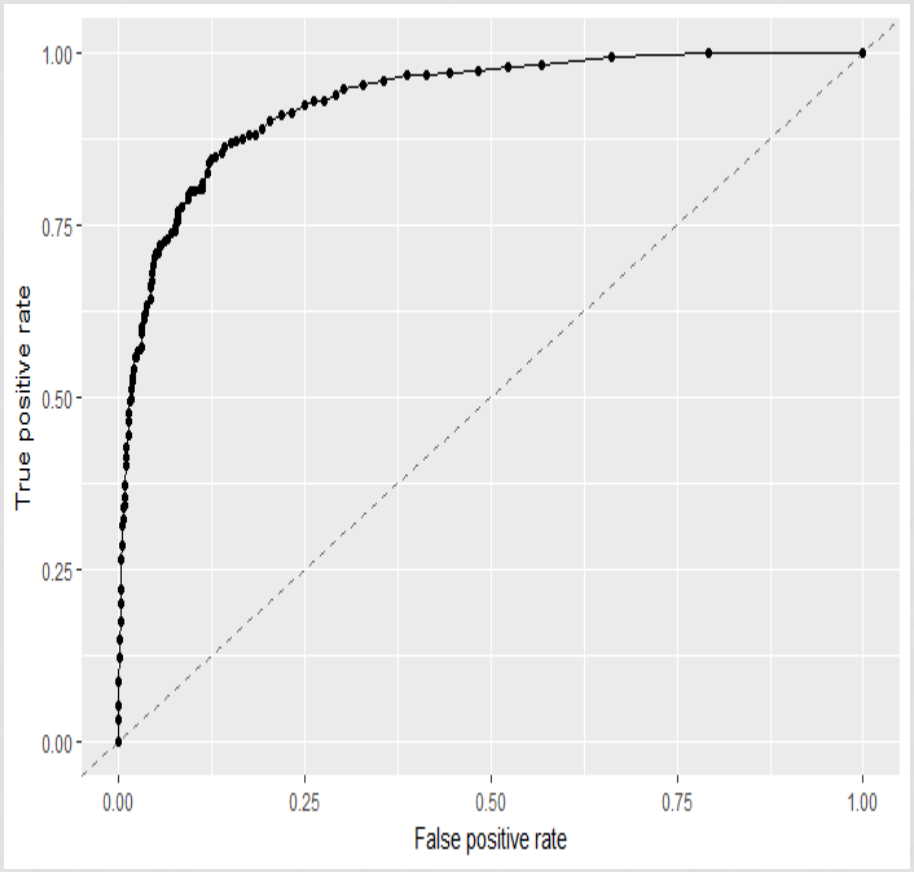
- Por ejemplo, un clasificador lineal, o una máquina de vectores de soporte, que devuelven una distancia del dato a la frontera:
  - Si el valor g(x) es muy negativo, cercano a la clase 0
  - Si el valor g(x) es muy positivo, cercano a la clase 1
- Los scores también pueden estar entre 0 y 1 (como con RF)
- Es fácil transformar un *sc* en un clasificador binario discreto, sin mas que fijar un threshold *t* (valor de corte):
  - Si g(x) <= t entonces clase 0
  - Si g(x) > t entonces clase 1
- Eso quiere decir que para distintos t, tenemos distintos clasificadores discretos, es decir, distintos puntos en el espacio ROC

EJEMPLO DE *SC*

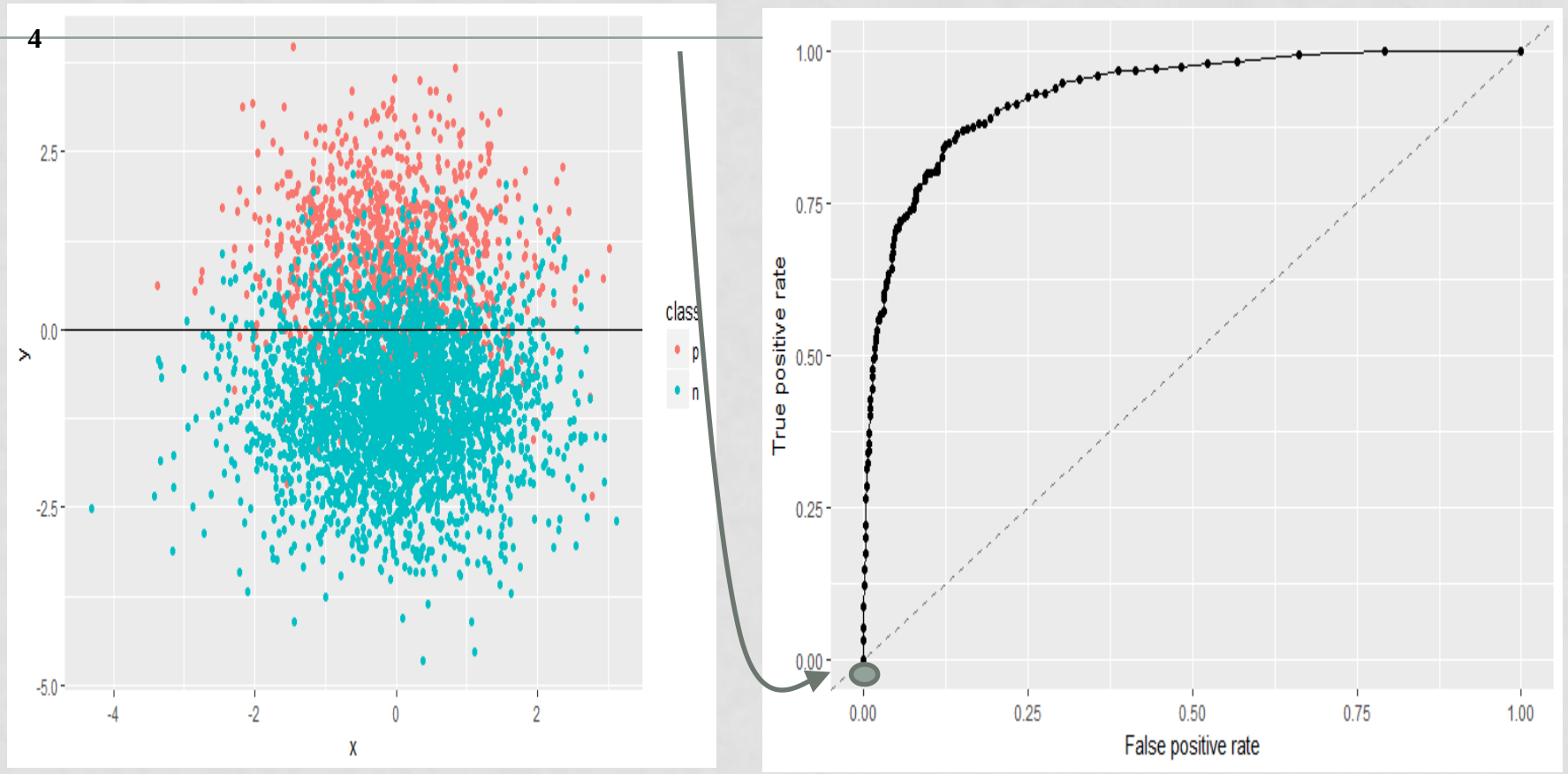
- Modelo lineal para este problema
- El "score" es la distancia a la frontera (valor positivo a un lado y negativo al otro)
- Clasificaremos un dato como positivo si la distancia (score) a la frontera >= threshold
  - (en este caso, las distancias (scores) en la zona roja serán positivas y las de la zona azul, negativas)



- Para cada threshold, habrá un punto en la curva ROC

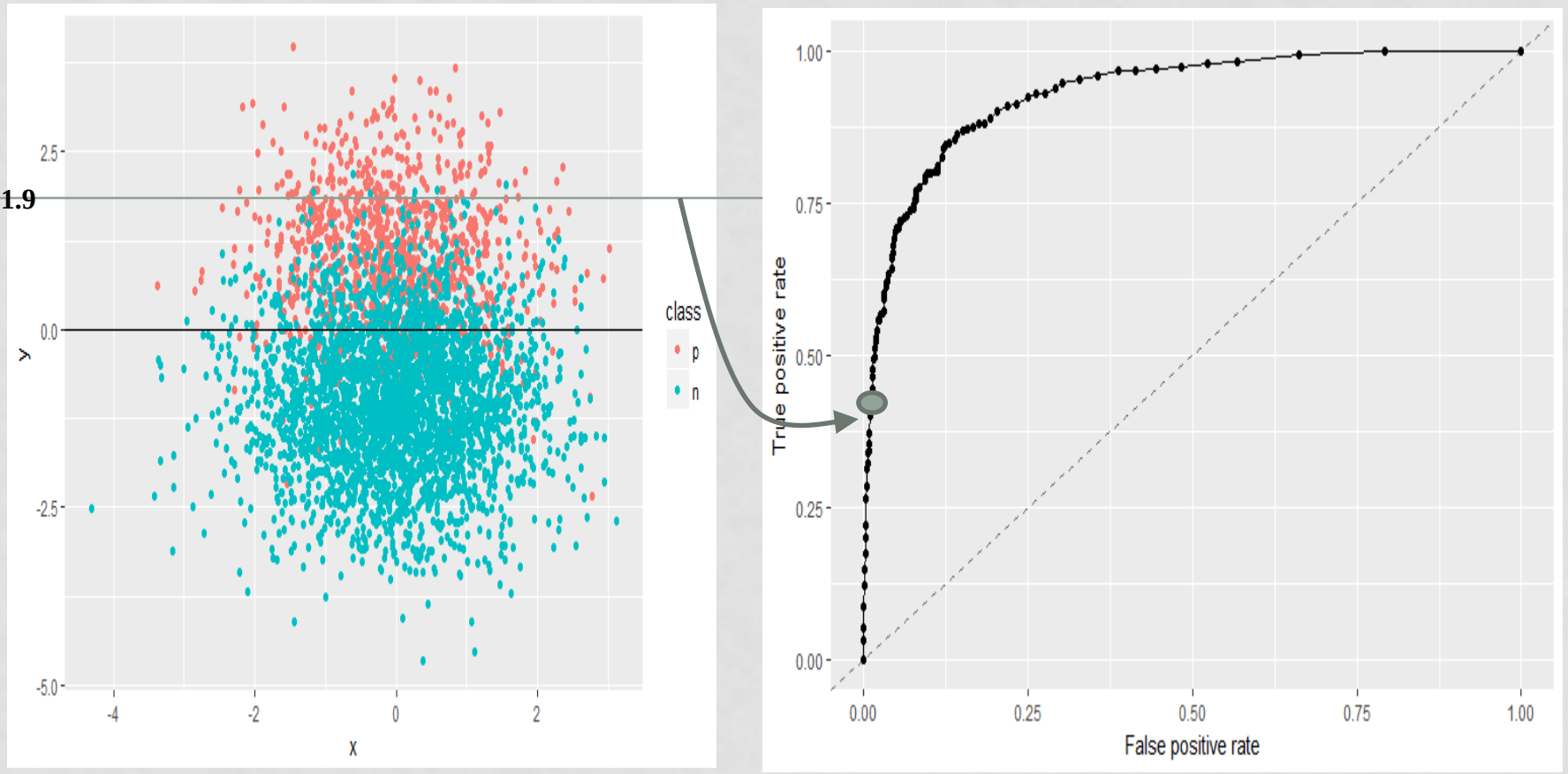


- Para cada threshold, habrá un punto en la curva ROC
- Si score >=4 entonces “positivo”, en caso contrario “negativo”



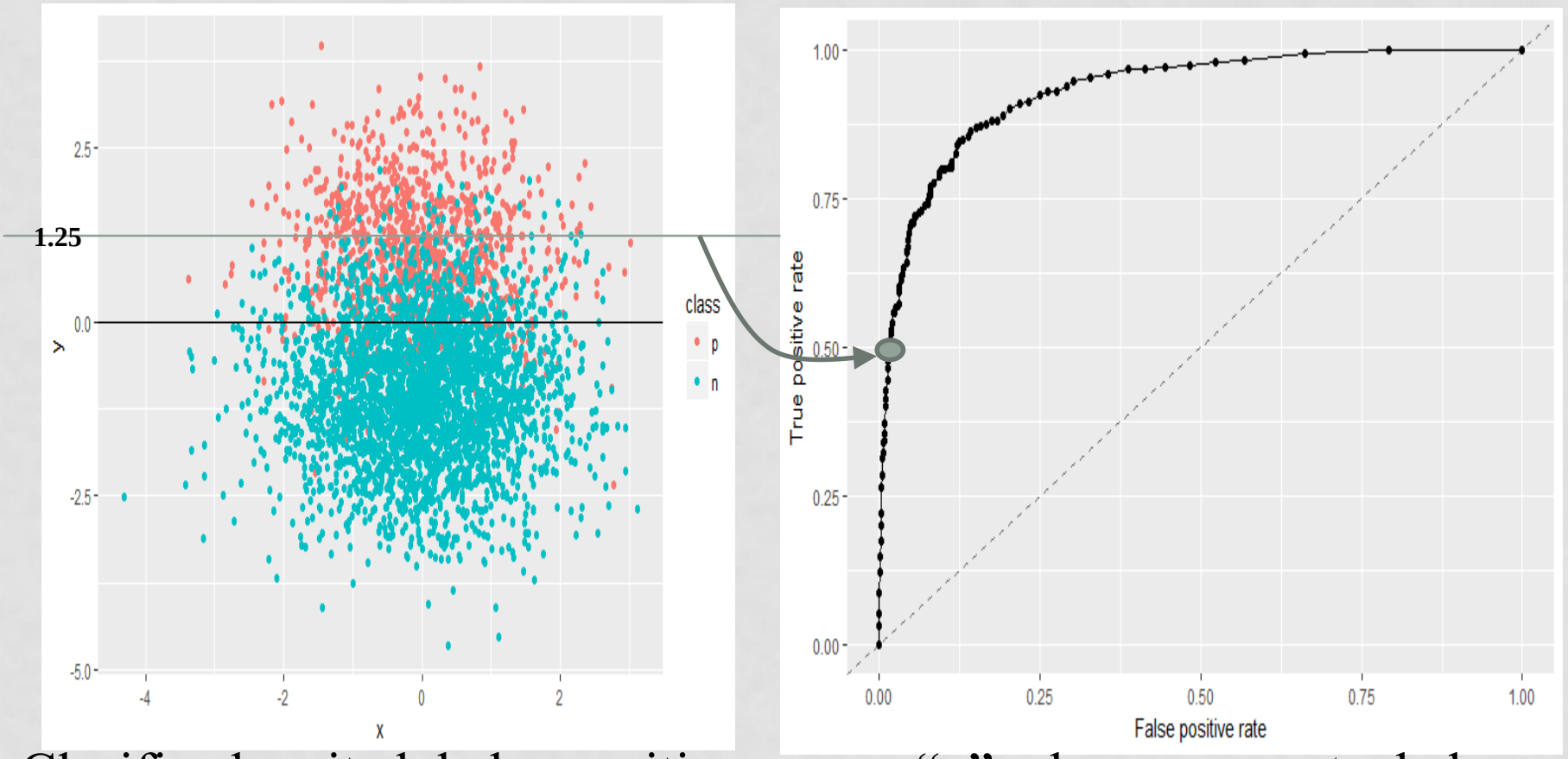
Todo lo clasifica como “n”: TPR=0, FPR=0

- Para cada threshold, habrá un punto en la curva ROC
- Si distancia >=1.9 entonces "positivo", en caso contrario "negativo"



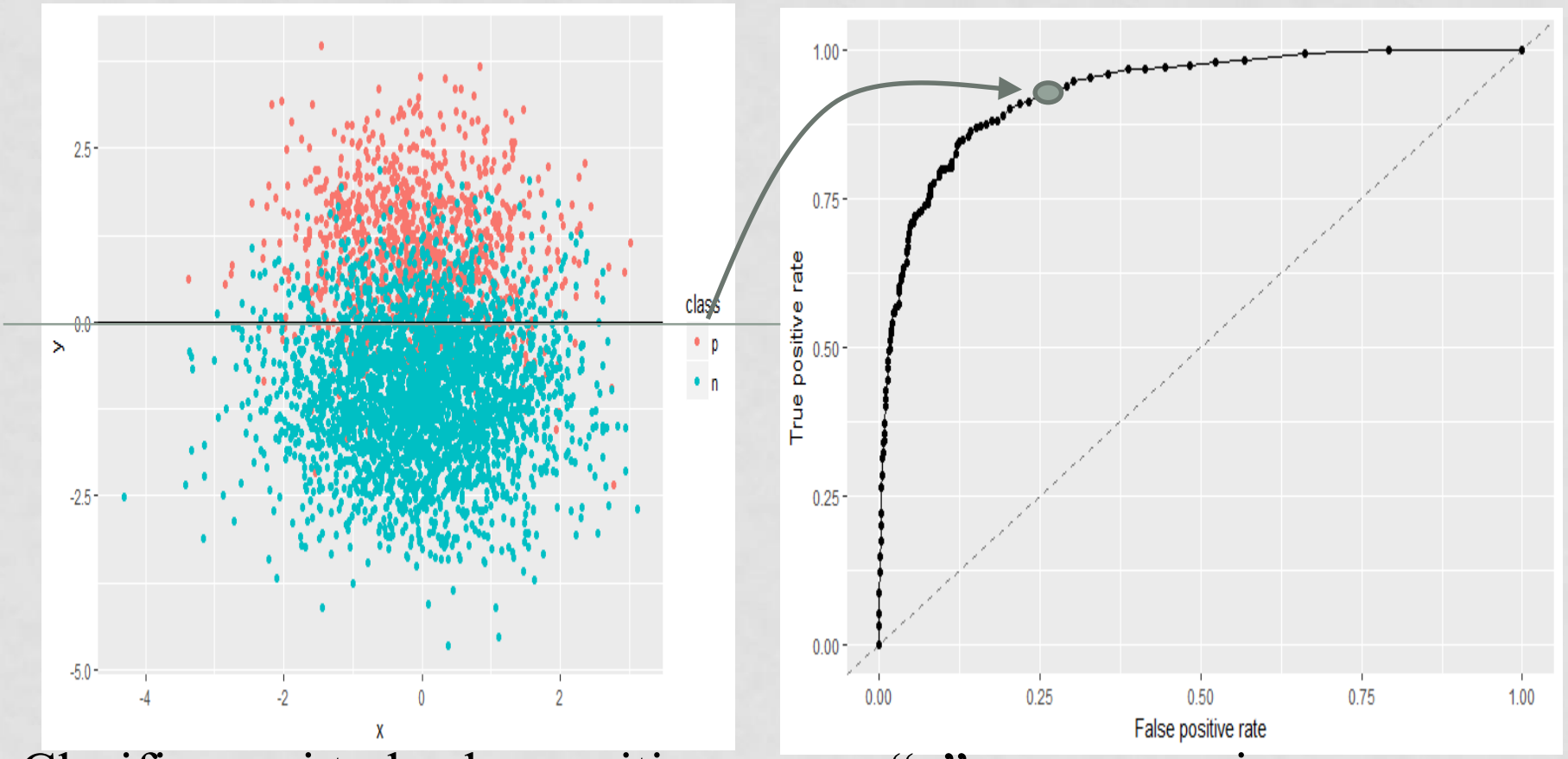
Clasifica los de encima como "p" y los de debajo como "n".

- Para cada threshold, habrá un punto en la curva ROC
- Si distancia >=1.25 entonces "positivo", en caso contrario "negativo"



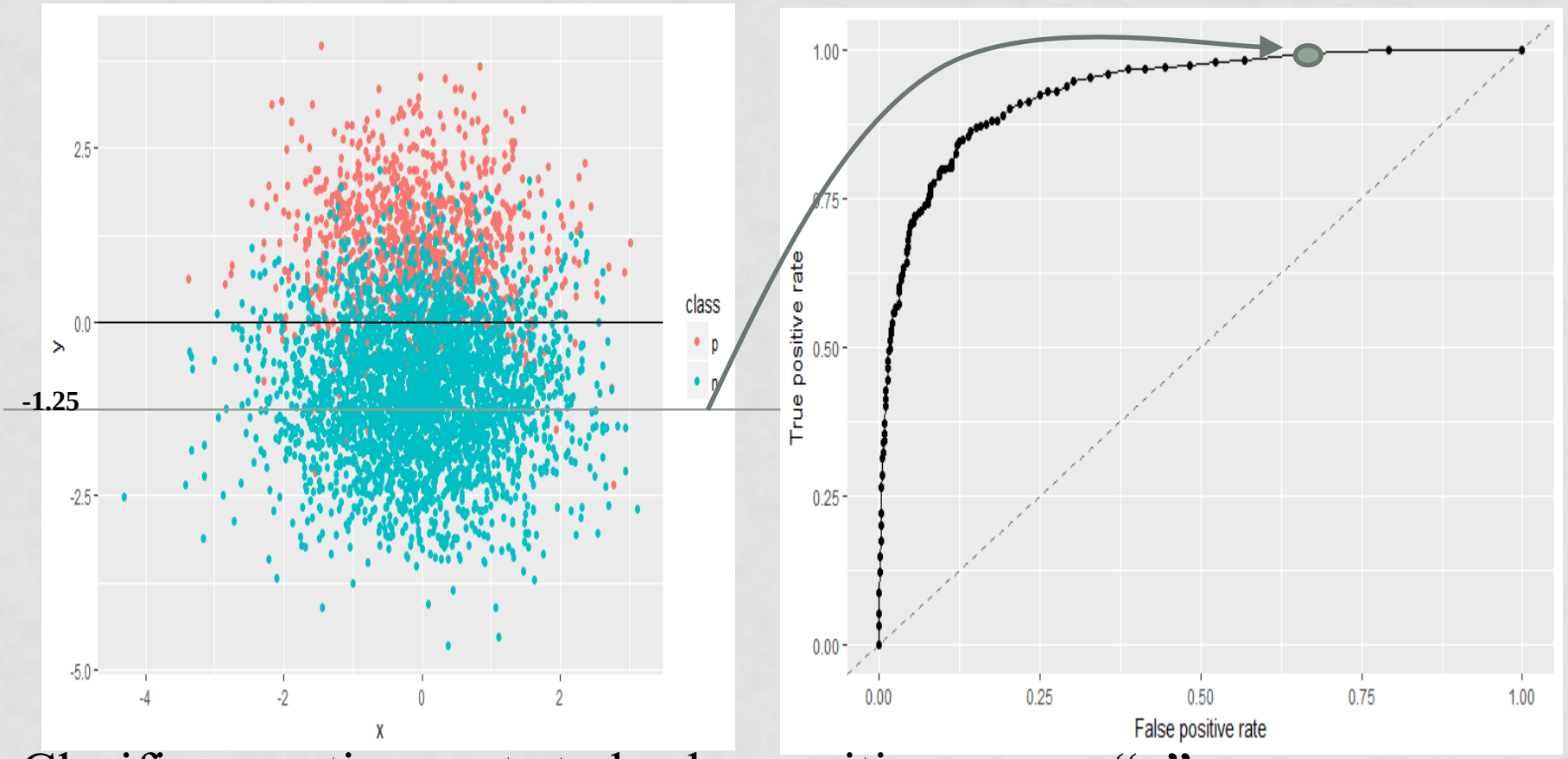
Clasifica la mitad de los positivos como "p" y la mayor parte de los negativos como "n".

- Para cada threshold, habrá un punto en la curva ROC
- Si distancia >=0 entonces "positivo", en caso contrario "negativo"



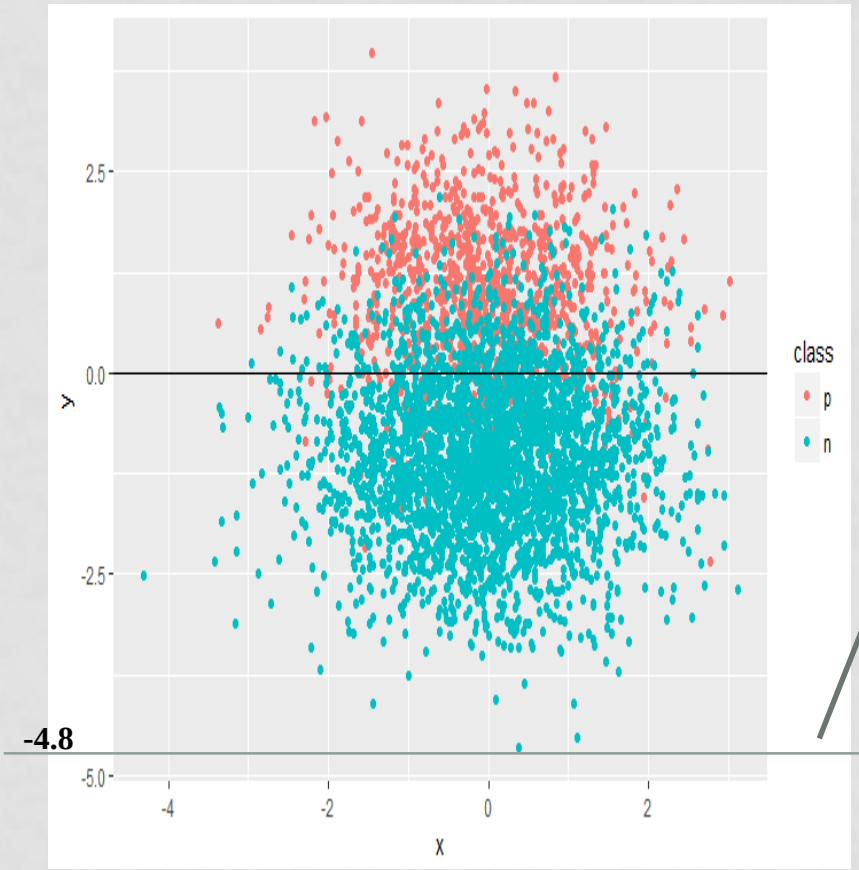
Clasifica casi todos los positivos como "p" pero se equivoca en unos cuantos negativos (25%).

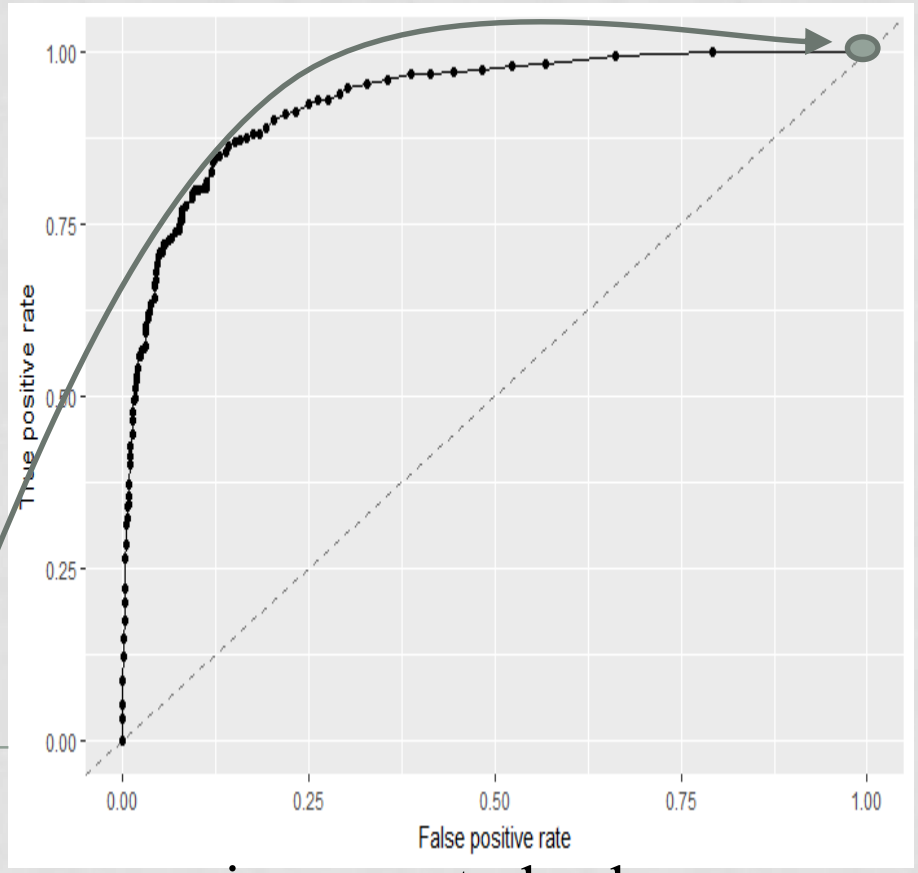
- Para cada threshold, habrá un punto en la curva ROC
- Si distancia >= -1.25 entonces "positivo", en caso contrario "negativo"



Clasifica practicamente todos los positivos como "p" pero se equivoca ya en bastantes negativos (25%).

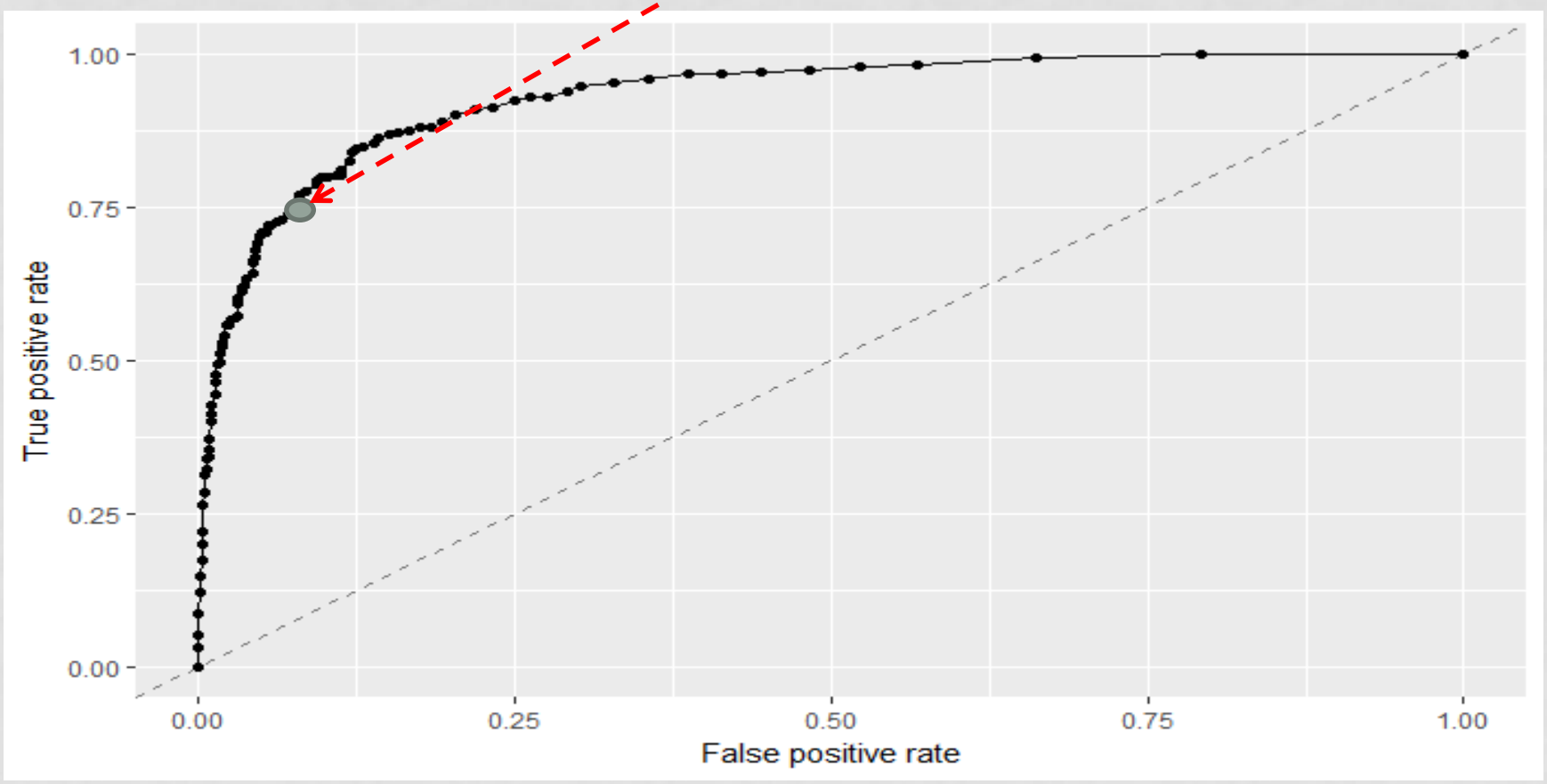
- Para cada threshold, habrá un punto en la curva ROC
- Si distancia >= -4.8 entonces "positivo", en caso contrario "negativo"





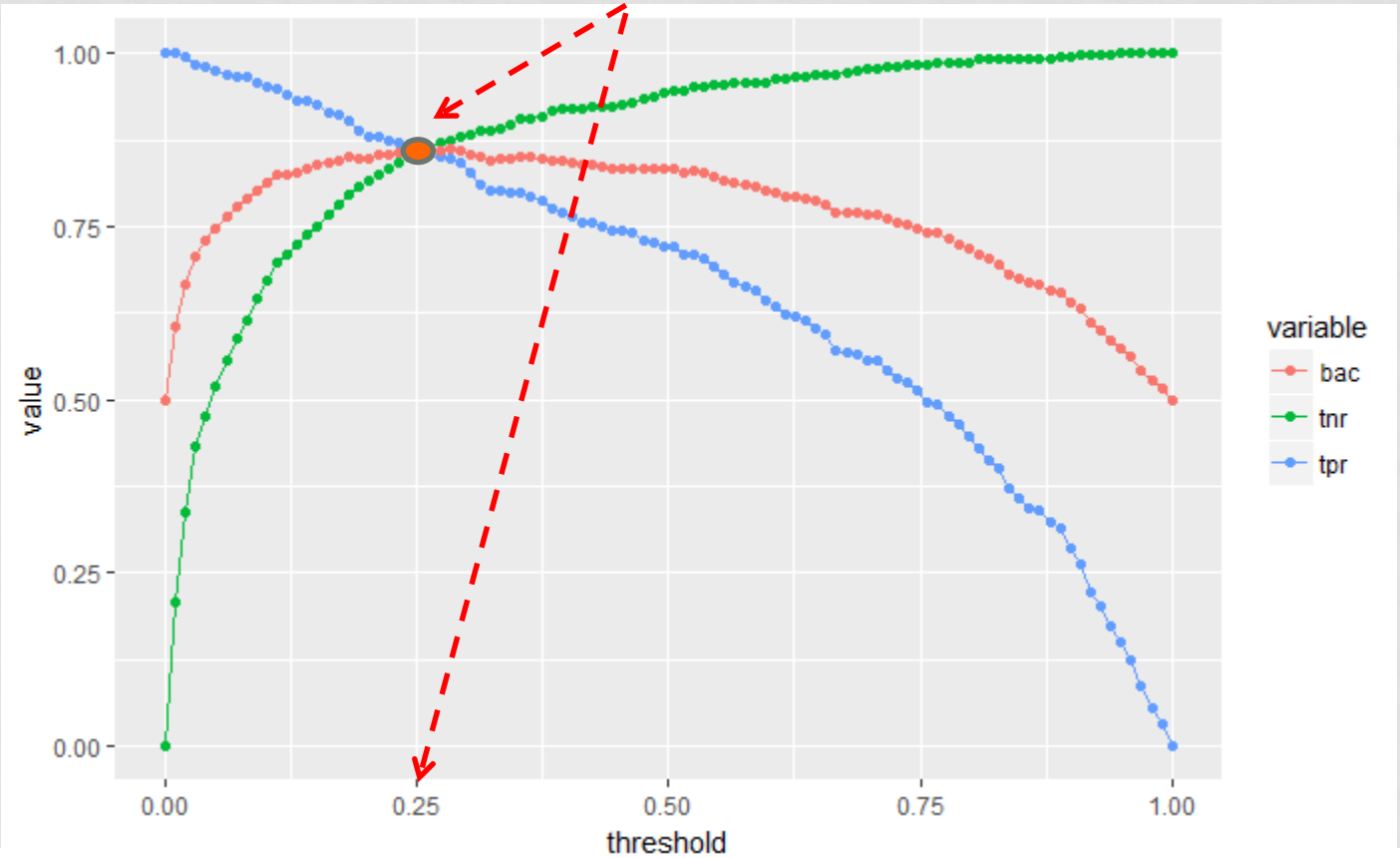
Clasifica todo como "p", con lo que se equivoca en todos los negativos.

- Podemos usar la curva ROC para elegir un threshold adecuado.
- Por ejemplo, si es suficiente con acertar la clase positiva en un 75%:



THRESHOLDING

- En general, podemos usar “thresholding”: ajustar el threshold para optimizar alguna medida. Por ejemplo, el “balanced accuracy” (macromedia)



ORGANIZACIÓN

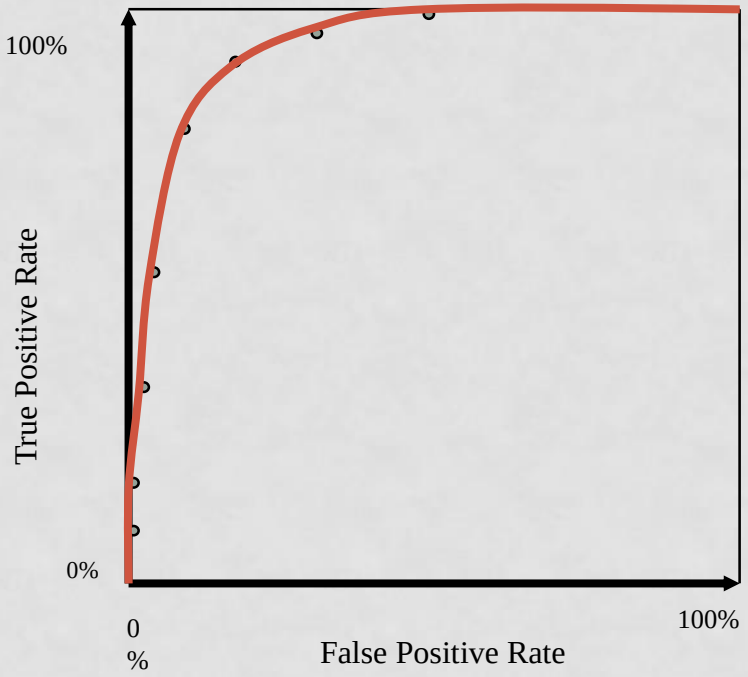
- **Evaluación** teniendo en cuenta distribución y coste
- **Aprendizaje** teniendo en cuenta distribución y coste
- La Métrica AUC: el área bajo la curva ROC

MÉTRICA AUC

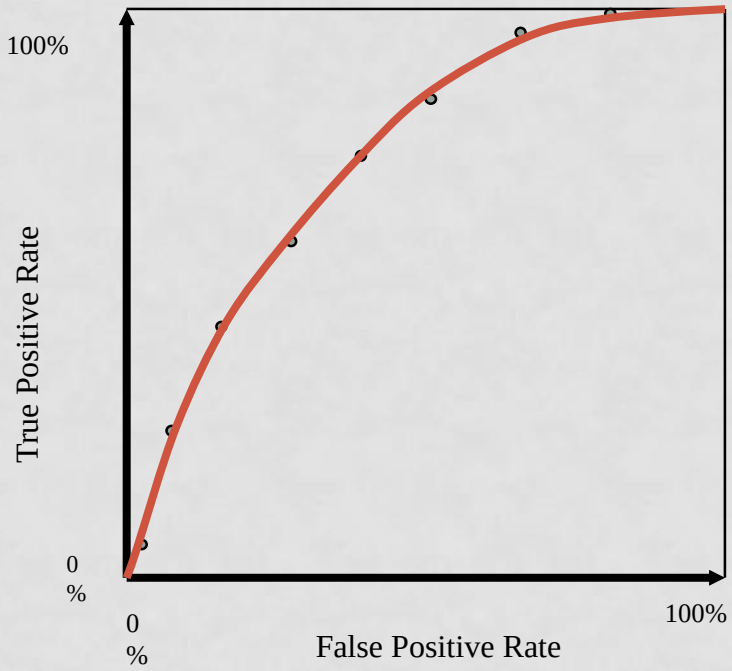
- Aparte de permitirnos elegir un *threshold* adecuado, las curvas ROC tienen otra utilidad en evaluación de modelos.
- Podemos usarlas para evaluar y comparar modelos.

COMPARACIÓN DE CURVAS ROC

Buena



Mala



LA MÉTRICA AUC

- Se puede demostrar que AUC = probabilidad de que el score de una instancia positiva sea mayor que el de una instancia negativa = P(X>Y)
- Nótese que a diferencia del error, no se deja engañar por modelos que clasifican todos los datos como pertenecientes a la clase mayoritaria.
- Podemos ver esto con un clasificador discreto.

AUC DE UN CLASIFICADOR DISCRETO

- La curva ROC de un clasificador discreto está constituida por tres puntos:
  - El punto ROC del clasificador
  - El del clasificador trivial que clasifica todo como negativo
  - El del clasificador trivial que clasifica todo como positivo
- Para clasificadores discretos AUC = (TN + TP)/2 (la macromedia o BAC)
- Podemos ver que el AUC de un clasificador casi trivial será próxima a 0.5 = (1+0)/2

EXTENSIÓN A MAS DE DOS CLASES

- La medida AUC se puede extender a más de dos clases.
  - Extensión de "todos los pares"

$$AUC_{HT} = \frac{1}{c(c-1)} \sum_{i=1}^{c} \sum_{j=1, j<i}^{c} AUC(i, j)$$

  - Extensión "uno contra todos" (Fawcett)