



Técnicas escalables de análisis de datos en entornos Big Data: Clasificadores

Curvas ROC y otras métricas de evaluación de
clasificadores





Contenido

1. Motivación: Limitaciones de la tasa de error.
2. Costes de clasificación.
3. Estructura del error: Matriz de confusión.
4. Clasificación con coste.
5. Curvas ROC.
6. Otras métricas.
7. Referencias.

1. Motivación: Limitaciones de la tasa de error

- Hasta ahora, hemos utilizado la tasa de error como criterio para evaluar hipótesis.
- Suposiciones:
 - Distribuciones de clases no muy desequilibradas.
 - Los costes de los errores son los mismos.
- No siempre es así
 - 99,99% de la población no es terrorista.
 - «No terrorista» cierto 99,99%.
 - El 97% de los días del mes las vacas no están en celo
 - «No celo» cierto 97,00%.
- ¿Coste de diagnosticar erróneamente la presencia de una enfermedad frente a no diagnosticar una enfermedad real?



2. Costes de clasificación

- Optimizar la tasa de error sin tener en cuenta el coste de cometer un error de clasificación puede producir resultados no deseados.
- En la práctica, diferentes tipos de error de clasificación suelen tener costes diferentes.
- Interesa conocer la estructura del error para buscar una buena solución de compromiso.
- Otros tipos de costes que no estamos considerando, pero que son importantes en un proyecto de Big Data
 - Coste de obtención datos.
 - Coste aplicación hipótesis.

3. Estructura del error: matriz de confusión

- Problema binario:

		Clase predicha	
		sí	No
Clase real	sí	Ciertos Positivos (<i>TP</i>)	Falsos Negativos (<i>FN</i>)
	no	Falsos Positivos (<i>FP</i>)	Ciertos Negativos (<i>TN</i>)

$$P = TP + FN$$

$$N = FP + TN$$

- Tasa de acierto: $[TP + TN]/[P + N]$
- Tasa de ciertos positivos: $tp = \frac{TP}{P} = \frac{TP}{TP+FN}$
- Tasa de falsos positivos: $fp = \frac{FP}{N} = \frac{FP}{FP+TN}$

Matriz de confusión problema multiclase

		Predicted class			
		a	b	c	total
Actual class	a	88	10	2	100
	b	14	40	6	60
	c	18	10	12	40
total		120	60	20	

- Tasas de acierto = $\frac{88+40+12}{88+10+2+\dots+12} = 0.7$
- Tasa de ciertos positivos por clase
 $TP_a = \frac{88}{88+10+2} = 0,88$
- Tasa de falsos positivos por clase
 $FP_a = \frac{14+18}{14+40+6+18+10+12} = 0,32$

- Tasas de acierto = $\frac{\sum_{i=1}^k x_{i,i}}{\sum_{i=1,j=1}^k x_{i,j}}$
- Tasa de ciertos positivos por clase $TP_i = x_{i,i} / \sum_{j=1}^k x_{i,j}$
- Tasa de falsos positivos por clase $FP_i = \frac{\sum_{j=1, j \neq i}^k x_{j,i}}{\sum_{j,l, j \neq i}^k x_{j,l}}$
- **No utilizar métricas promediadas:** enmascaran el mal comportamiento en las clases de interés (que suelen tener pocos ejemplos).

4. Clasificación con coste

- Si se conoce el coste de los errores, se puede incluir en el proceso de toma de decisión.
- Matriz de coste

		Clase predicha		Clase predicha		
		a	b	a	b	c
Clase actual	a	0	1	Clase actual	a	0
	b	5	0		b	1
				c	8	16
						0

- Clasificación: remplazar la tasa de error por el coste promedio por predicción y asignar la clase con menor coste.
- **PERO** los costes de los errores rara vez se conocen con precisión.
- Alternativa: analizar distintos escenarios (curvas ROC, PR, etc...).

5. Curvas ROC

- Análisis ROC: *Receiver Operation Characteristic.*
- Origen en Teoría de la señal, para mostrar el compromiso entre la tasa de detección (tp) y la tasa de falsas alarmas (fp) sobre canales con ruido.
- Muy utilizados en diagnosis médica.
- Especialmente interesante si costes muy diferentes o distribución de clases muy desigual (*skewed*).
- Especialmente útil en problemas binarios.



Espacio ROC

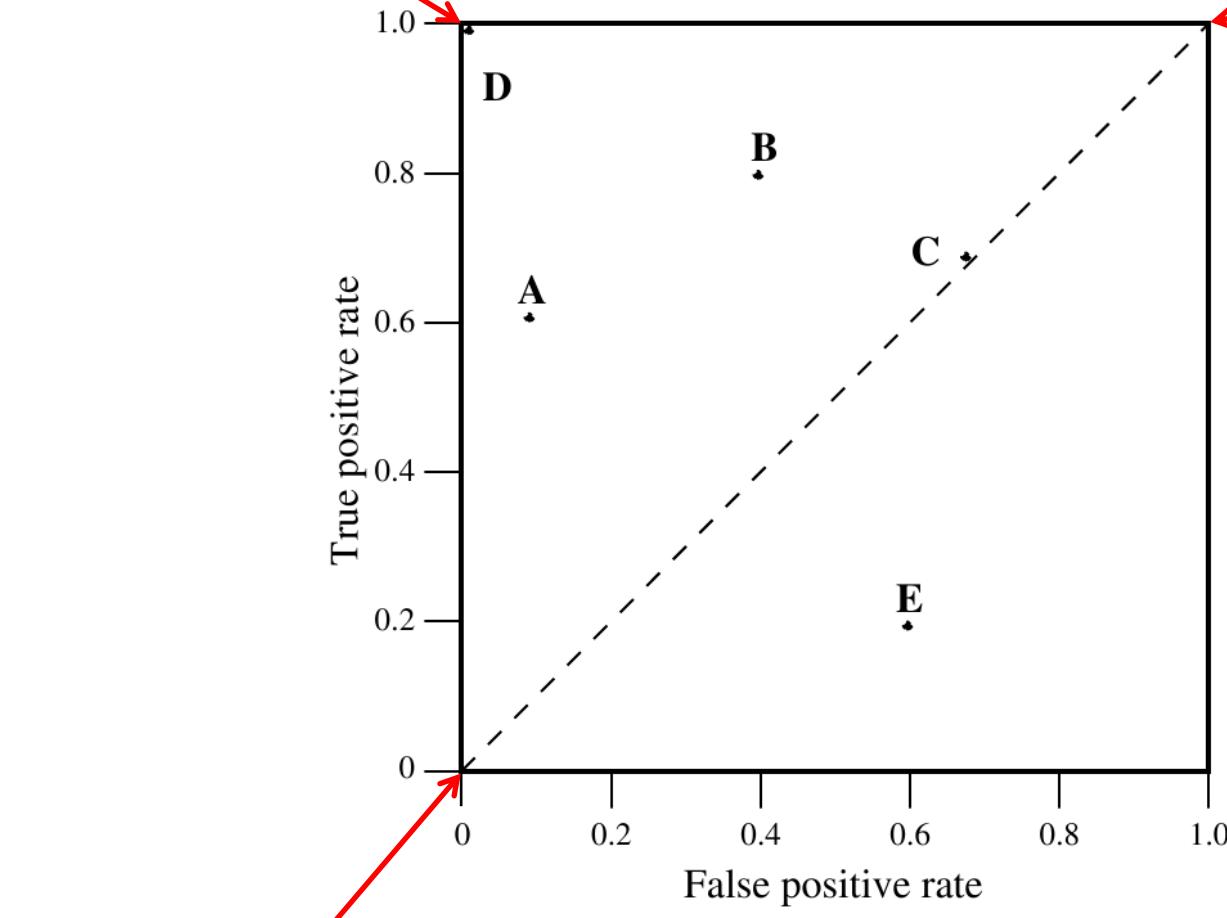
- Gráficos bidimensionales.
- Eje Y: tasa de ciertos positivos.
- Eje X: tasa de falsos positivos.
- Punto en el espacio: pares (fp, tp) .
 - Tasa de ciertos positivos: $tp = TP/P$.
 - Tasa de falsos positivos: $fp = FP/N$.
- Relación entre los beneficios (ciertos positivos) y los costes (falsos positivos).

- Cada clasificador binario: un punto en el espacio ROC.
- $(0,0)$ clasificador que siempre predice la clase negativa.
- $(1, 1)$ clasificador que siempre predice la clase positiva.
- Un clasificador es mejor que otro si está encima y a la izquierda.
- Predicción aleatoria: diagonal.

Puntos característicos

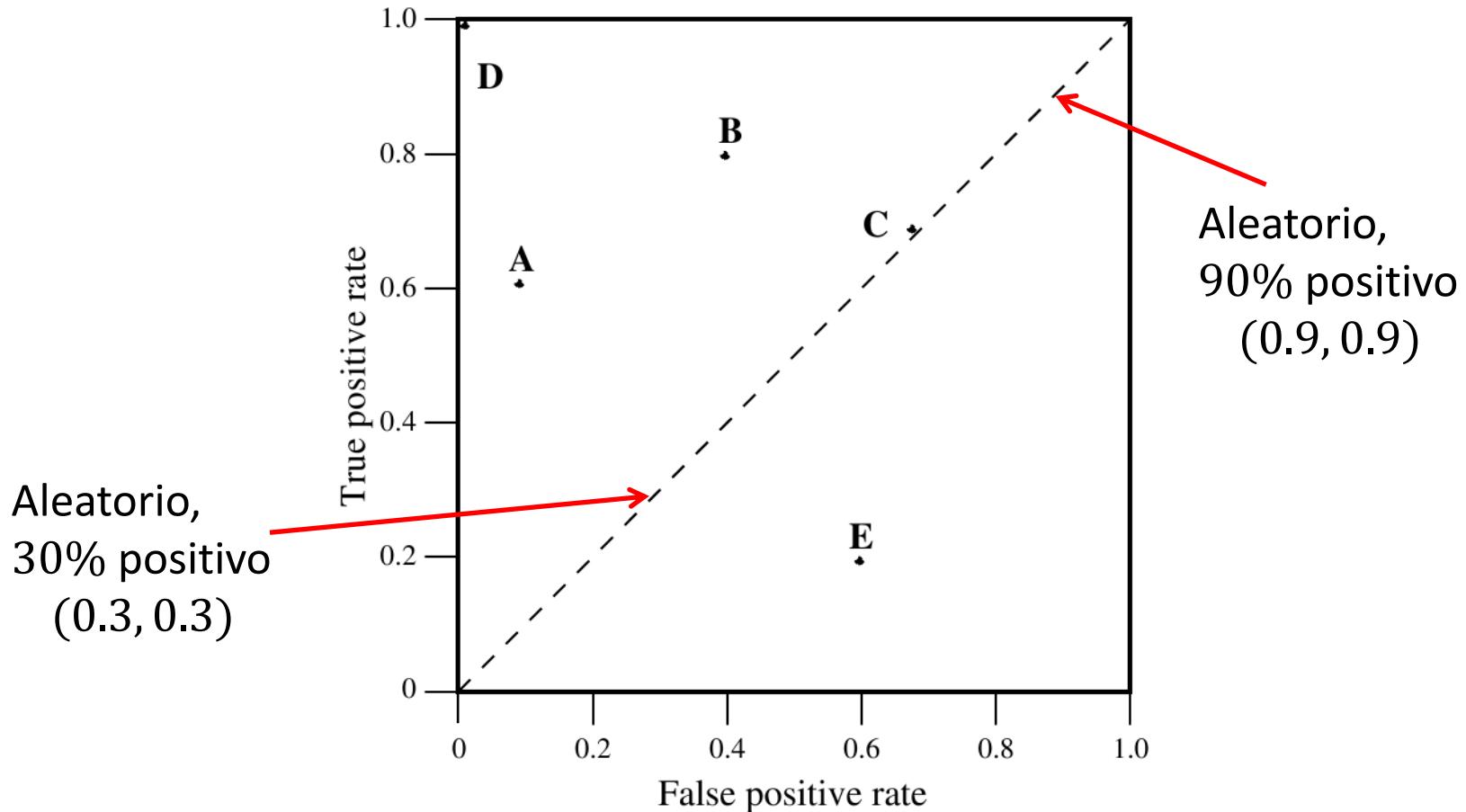
Ideal: $(0, 1)$

Siempre positivo: $(1, 1)$

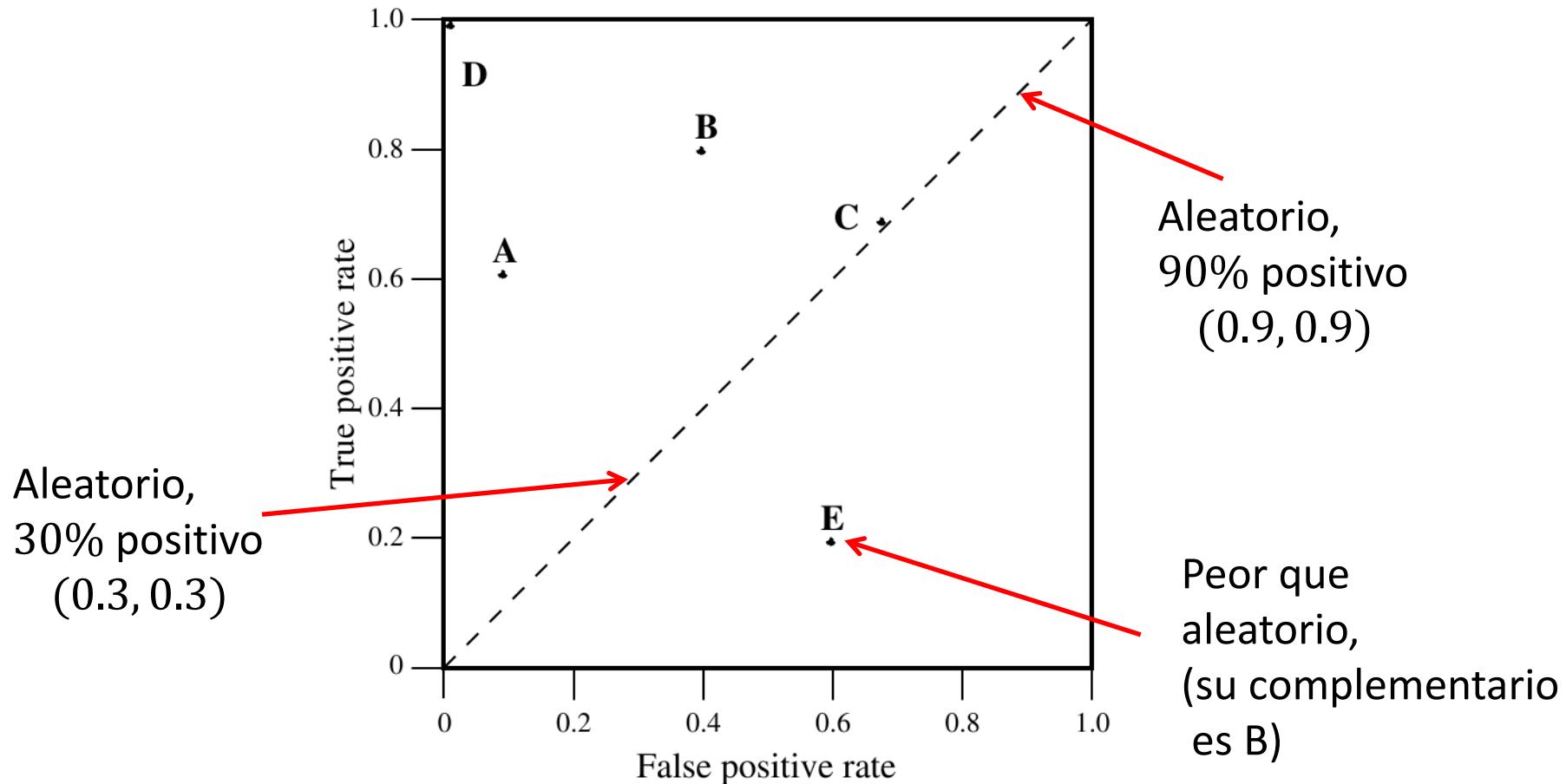


Siempre negativo: $(0, 0)$

Clasificador aleatorio: diagonal

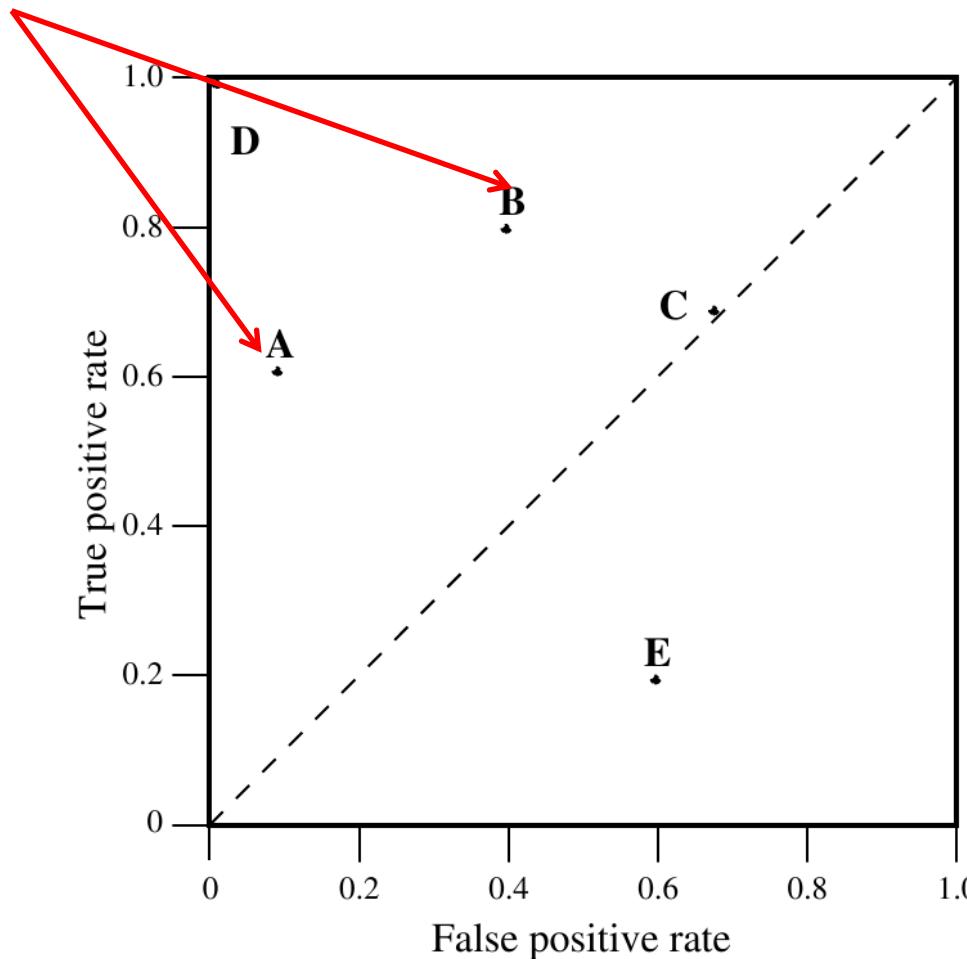


Peor que aleatorio: bajo la diagonal



Comparación de clasificadores

A es más conservador que B



Si el coste de los falsos positivos es elevado, preferiríamos A

Si el coste de los falsos positivos es asumible, preferiríamos B

Generación de curvas ROC

- Un clasificador binario produce un punto en el espacio ROC.
- Hay clasificadores que proporcionan algún tipo de “puntuación continua” (*score*): Naïve Bayes, Regresión logística, Redes de Neuronas.
- Utilizando un umbral para decidir la clase, tenemos una familia de clasificadores
 - Clase positiva si $score > umbral$.
- Variando el umbral, generamos una curva.

- Si el clasificador solo devuelve la clase, adaptarlo para que devuelva un *score*.

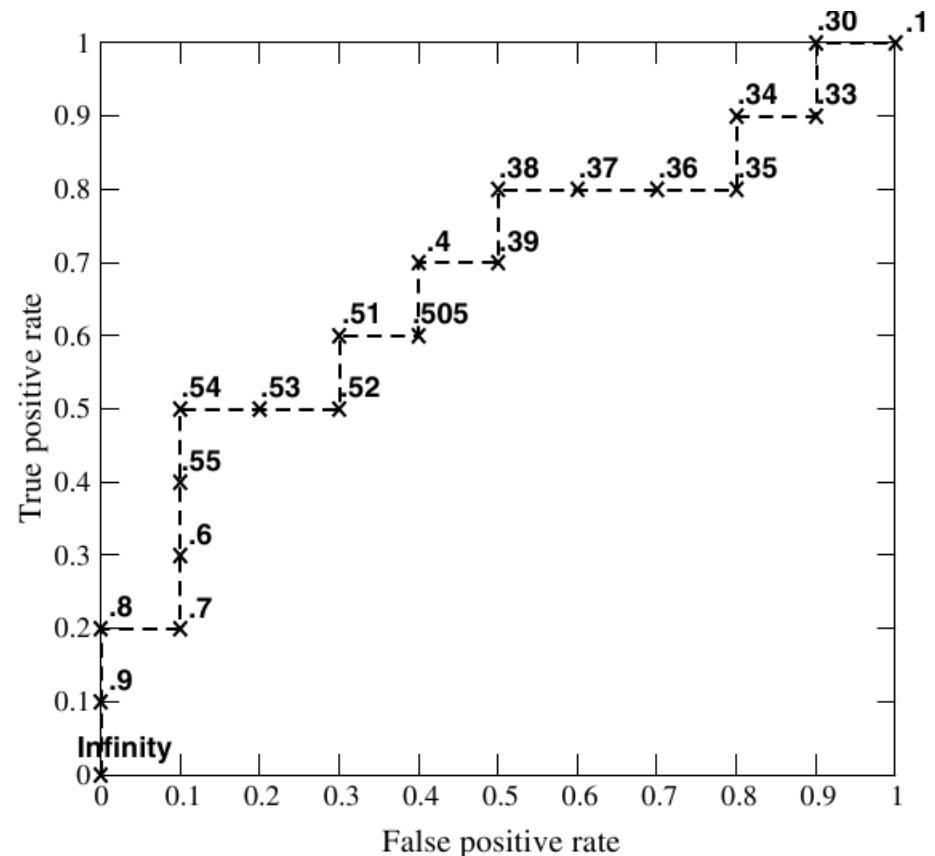
Ejemplo: curva ROC sobre conjunto de Prueba con 20 instancias

- Crear el clasificador.
- Evaluarlo sobre el conjunto de prueba.
- Ordenar los ejemplos de prueba por orden decreciente de *score*.
- Mover el umbral entre *scores* consecutivos.
- Cada elección de umbral define un clasificador y un punto en el espacio ROC.

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

Curva ROC sobre un conjunto de prueba

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

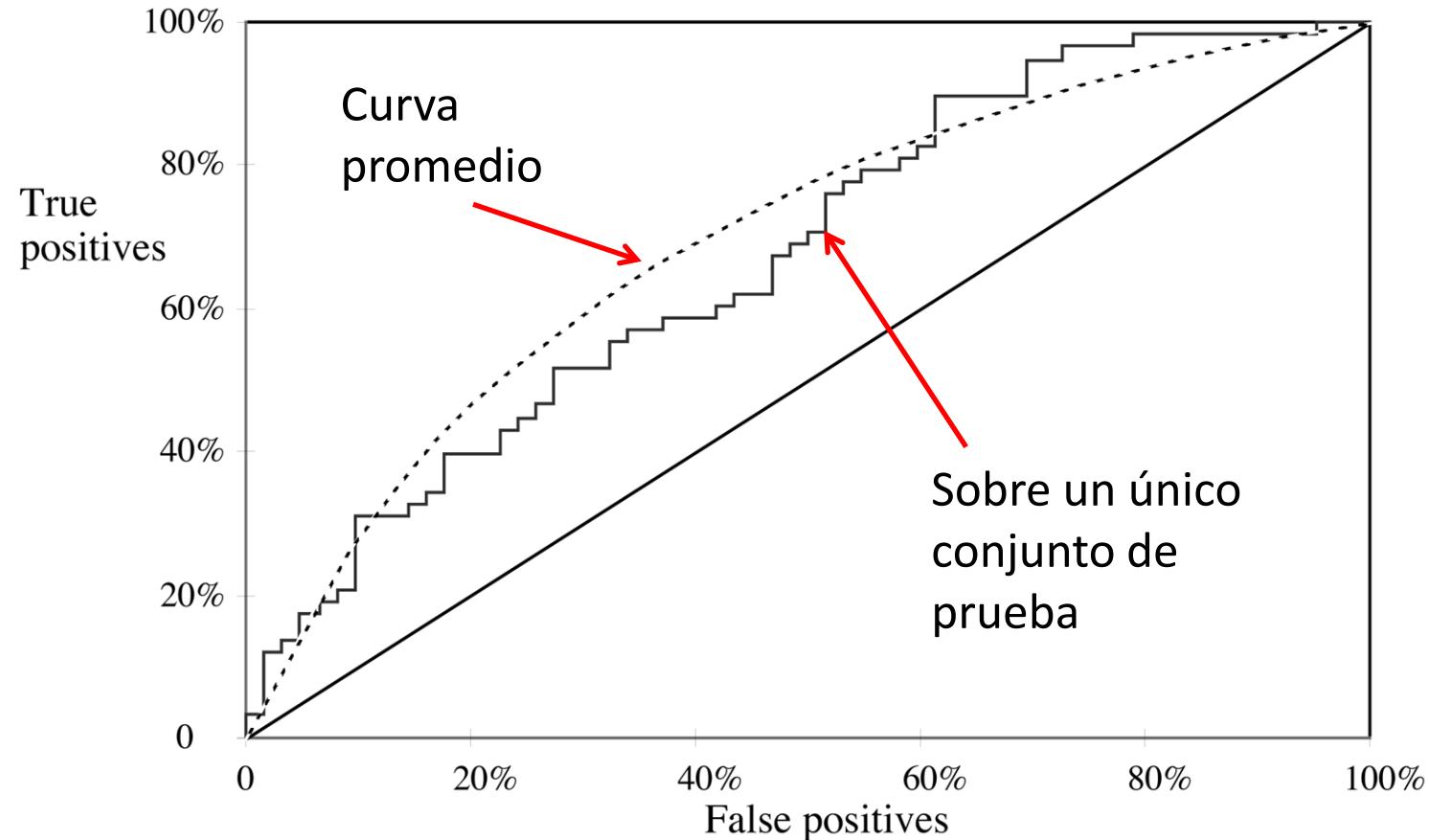




Generación de curvas ROC

- La forma de la curva ROC depende del conjunto de prueba utilizado.
 - Se puede obtener una curva más suave promediando valores de varios experimentos.
-
- Promediando sobre tasa tp (promedio vertical).
 - Mediante validación cruzada.
 - En cada partición
 - Obtener tasa tp para cada valor fp .
 - Obtener tasa tp media para cada valor fp .

Ejemplo curva promedio

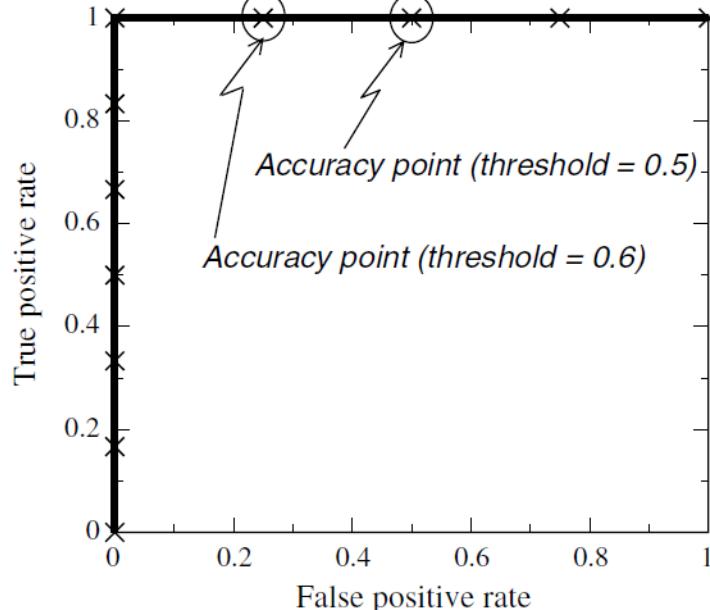




Propiedades de las curvas ROC

- Las curvas ROC miden la capacidad del clasificador para producir buenas puntuaciones relativas.
- Las curvas ROC son invariantes ante cambios en la distribución de clases (en el conjunto de prueba).

Scores relativos



Inst no.	Class		Score
	True	Hyp	
1	p	Y	0.99999
2	p	Y	0.99999
3	p	Y	0.99993
4	p	Y	0.99986
5	p	Y	0.99964
6	p	Y	0.99955
7	n	Y	0.68139
8	n	Y	0.50961
9	n	N	0.48880
10	n	N	0.44951

- Tasa de error 0.2 (umbral 0.5, por defecto) pero curva ROC ideal
- Las curvas ROC mide la capacidad del clasificador para ordenar las instancias positivas frente a las negativas.

Propiedades de las curvas ROC

- Las curvas ROC son invariantes ante cambios en la distribución de clases (en el conjunto de prueba)
 - Si la proporción de ejemplos positivos/negativos cambia en el conjunto de prueba, la curva ROC no cambia.
 - Motivo: las tasas tp (y fp) son ratios respecto a los positivos (negativos) totales, P (N)
 - cada ratio se calcula con elementos de la misma fila.

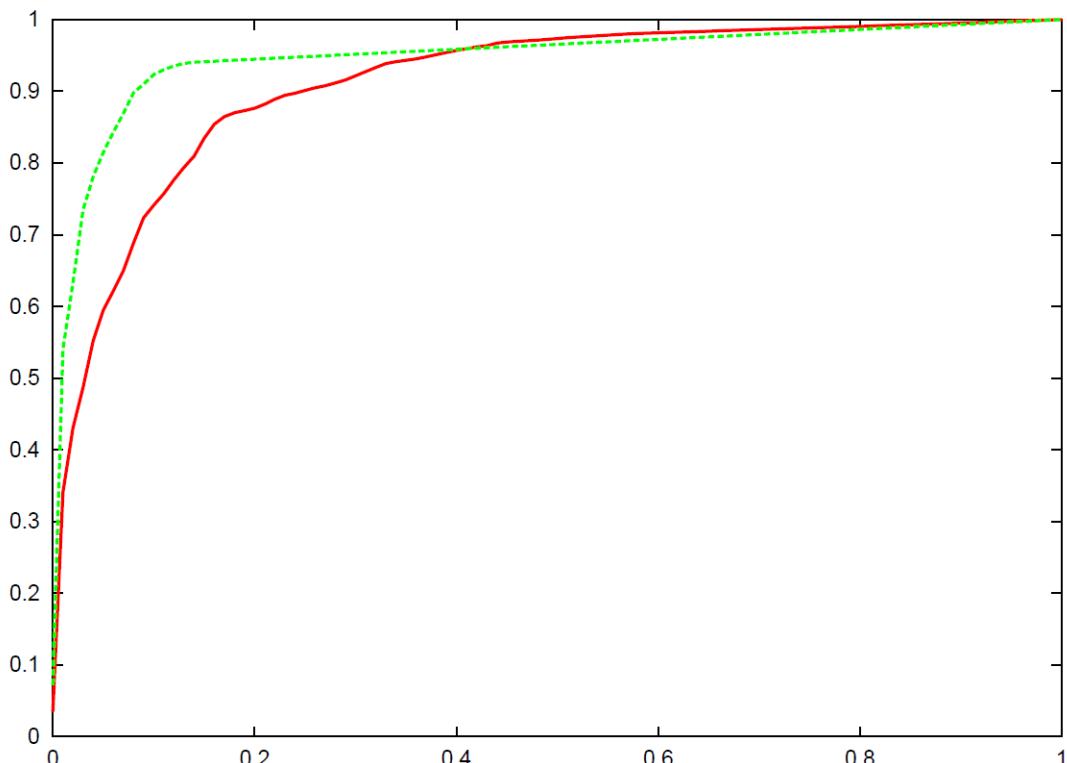
		Clase predicha	
		sí	no
Clase real	sí	TP	FN
	no	FP	TN

$$tp = \frac{TP}{TP + FN}$$
$$fp = \frac{FP}{FP + TN}$$

- Frente a otras métricas, que mezclan elementos de ambas filas.

Curvas ROC para dos clasificadores

- Punto de trabajo:
numero de fp .
- En cada punto de
trabajo, es preferible el
clasificador con tasa tp
más alta.
- Rara vez el mismo
clasificador se comporta
mejor en todos los
puntos de trabajo.

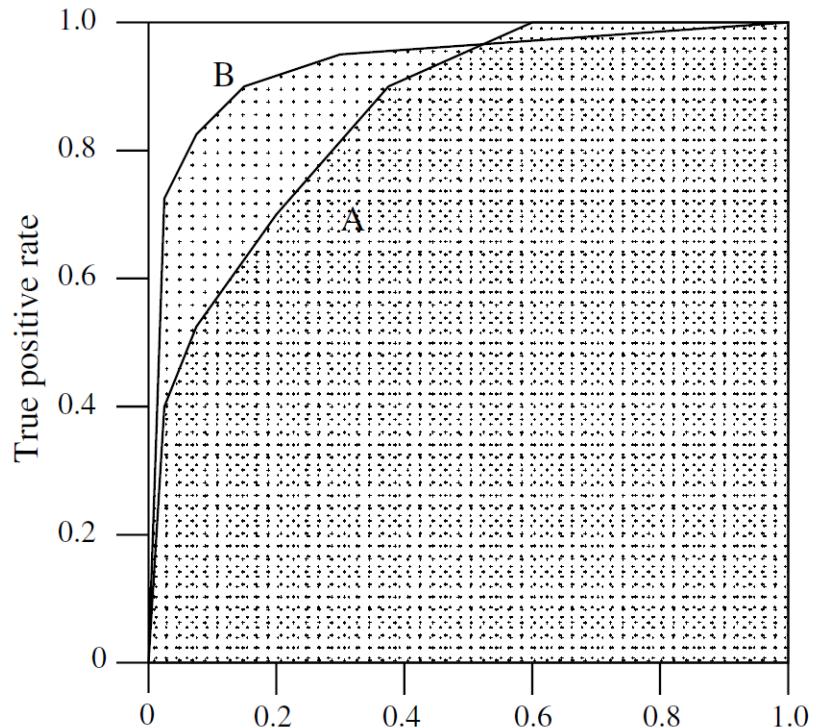


Área bajo la curva ROC: AUC

- Único valor numérico que facilita la comparación de dos clasificadores.
- AUC: porción del área bajo la curva en un cuadrado unitario
 - Entre 0 y 1.
 - Debería ser mayor que 0.5
 - Correspondiente a un clasificador aleatorio.
- Preferibles clasificadores con mayor AUC
 - Pero un clasificador con menor AUC puede comportarse mejor en alguna región.

Ejemplo AUC

- $AUC_B > AUC_A$
- Pero si estamos dispuestos a aceptar el coste de más fp con tal de identificar más positivos, el clasificador A es preferible.





Propiedad AUC

- El área bajo la curva es equivalente a la probabilidad de que el clasificador asigne un score superior a una instancia positiva que a una negativa.
- AUC es un buen indicador de la capacidad de predicción de un clasificador.

Resumen curvas ROC

- Herramienta para visualizar el comportamiento de un clasificador.
- Interesante porque no es sensible a la distribución de clases (en el conjunto de prueba) ni al coste.
- Conocido el coste de los errores, permite seleccionar el punto de trabajo
 - En combinación con un umbral.
- Utilizadas tradicionalmente en comunicaciones (radar) y diagnosis médica.

7. Otras métricas

- Marketing
 - Clientes que responden: TP .
 - Tamaño de la muestra: $(TP + FP)/(TP + FP + TN + FN)$ % clientes a los que se le envía.
 - Gráfico de elevación (*lift charts*).
- Recuperación de información
 - **Precision** ($TP/(TP + FP)$): % de documentos relevantes entre los que son recuperados.
 - **Recall** ($tp = TP/(TP + FN)$): % de documentos relevantes que son recuperados.
 - Curvas *precision/recall*.
- Diagnóstico médico: resultados de test
 - Sensibilidad (tp): porcentaje con enfermedad y test positivo.
 - Especificidad ($1 - fp$): porcentaje sin enfermedad y test negativo.

precision/recall

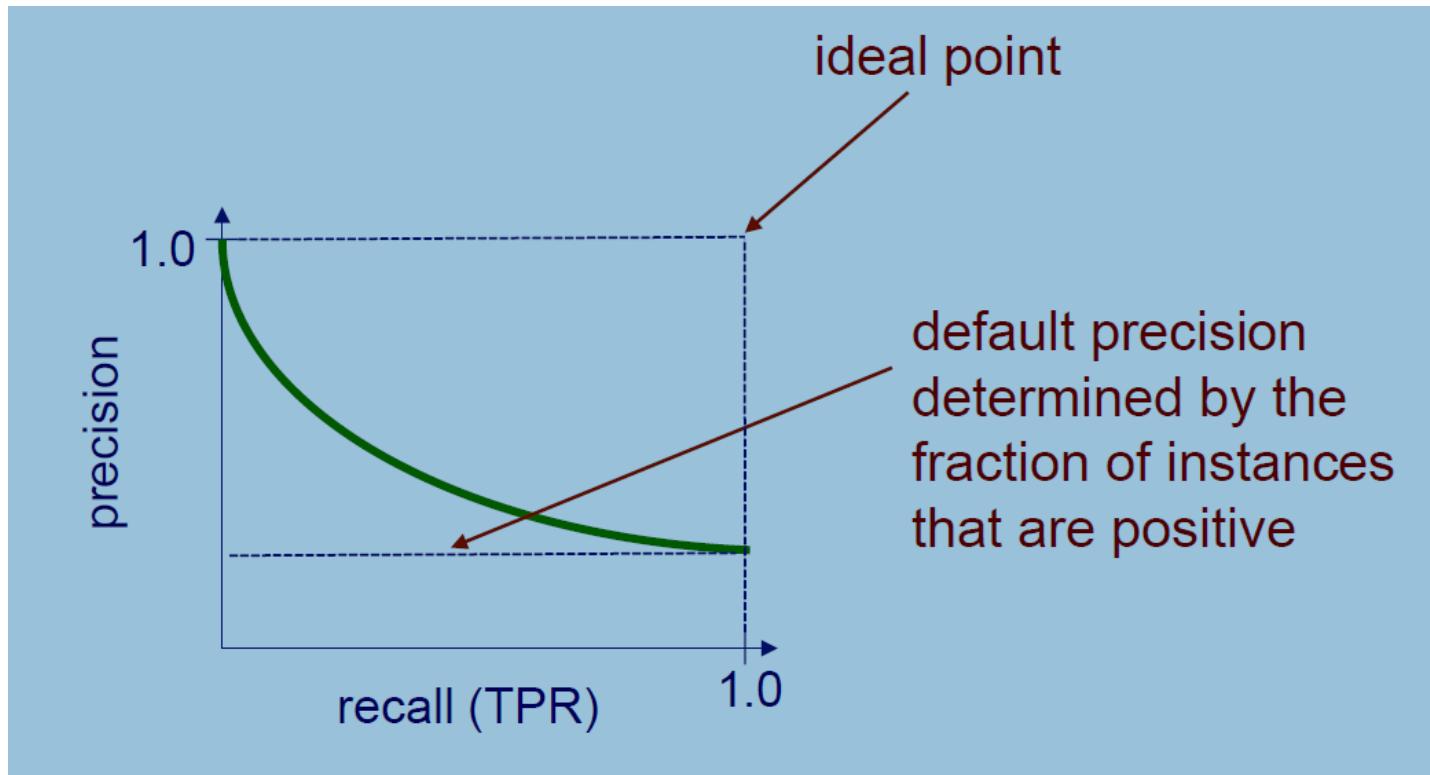
- Utilizadas en el ámbito de la recuperación de información.
 - También de utilidad en clasificación si la distribución de valores de la clase es muy desequilibrada.
- Documentos recuperados: equivalente a predicción positiva de un clasificador.
- *Precision* ($TP/(TP + FP)$): % de documentos relevantes entre los que son recuperados
 - *Precision* perfecto: 100, todos los documentos recuperados son relevantes (aunque no estén todos, FN; no hay FP).
- *Recall* ($tp = TP/(TP + FN)$): % de documentos relevantes que son recuperados
 - *Recall* perfecto: 100, se recuperaron todos los documentos relevantes (y quizás muchos no interesantes también, FP; no hay FN).

Espacio PR: curvas *precision/recall*

- Gráficos bidimensionales.
- Eje Y: *precision*.
- Eje X: *recall*.
- Punto en el espacio: pares (*recall, precision*)
 - *Precision*: $TP / (TP + FP)$: en %.
 - *Recall*: $tp = TP / (TP + FN)$: en %.
- Relación entre la calidad (Ciertos Positivos entre todos los recuperados) y la exhaustividad (*recall*: tasa de recuperados entre todos los relevantes).

- Un clasificador binario produce un punto en el espacio PR.
- Variando el umbral se obtiene una curva.
- La curva PR muestra la fracción de las predicciones que son ciertos positivos.
- (1, 1): clasificador ideal.

Ejemplo curva PR



Área bajo la curva PR y *F-measure*

- Ambas permiten resumir el comportamiento con una única medida.
- AUC mide la *precision* media de un clasificador.
- AUC=1: clasificador perfecto, con 100% *precision* y *recall*.
- Pero *precision* y *recall* suelen tener una relación inversa.
- *F-measure* es la media harmónica de *precision* y *recall*.
- $$F\text{-measure} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} = \frac{2 \times TP}{2 \times TP + FP + FN}$$
- Varía entre 0 y 1.
- Utilizada en recuperación de información.
- Discutible en clasificación, pues no considera TN.
 - Pero de utilizad en localización en imágenes (hay muchos TN).

8. Referencias

- Ian H. Witten, Eibe Frank, Mark A. Hall, Christopher Pal. Data Mining: practical machine learning tools and techniques (4th Edition). Morgan Kaufmann, 2016. ISBN: 9780128042915
- Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters* 27, 861-874, 2006.
- Davis, J. and Goadrich, M. The relationship between Precision-Recall and ROC curves. ICML '06 Proceedings of the 23rd international conference on Machine learning, pp 233-240, Pittsburgh, PA, 2006.
- Thomas Kautz, Bjoern M. Esko er, Cristian F. Pasluosta. Generic Performance Measure for Multiclass-Classifiers. *Pattern Recognition* , 8, 111-125, 2017. doi: 10.1016/j.patcog.2017.03.008
- David Page. Evaluating Machine Learning Methods.
<http://pages.cs.wisc.edu/~dpage/cs760/evaluating.pdf>. Último acceso: octubre 2025.