Tema 3: TÉCNICAS DE EVALUACIÓN EN CLASIFICACIÓN SUPERVISADA

MINERÍA DE DATOS

Alicia Pérez
alicia.perez@ehu.es

Lengoaia eta Sistema Informatikoak Saila
Bilboko Ingeniaritza Eskola

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Índice

EVALUACIÓN DE UN CLASIFICADOR
Clasificación Supervisada

Introducción: Clasificación Supervisada

Clasificación supervisada:

- Disponemos de un conjunto de N muestras etiquetadas (con su clase asociada)
- Cada muestra se caracteriza por un conjunto de n atributos $(X_1, \cdots, X_n)$
- Objetivo: inferir un modelo $(\mathcal{M})$ que dé cuenta de las muestras y que sea capaz de clasificar (asignar una etiqueta) a una nueva muestra (descrita por sus n características o atributos)

		X_1	$\cdots$	X_n	C
Muestras	$(\mathbf{x}^{(1)}, c^{(1)}) \Leftrightarrow$	$(x_1^{(1)},$	$\cdots,$	$x_n^{(1)},$	$c^{(1)})$
	$(\mathbf{x}^{(2)}, c^{(2)}) \Leftrightarrow$	$(x_1^{(2)},$	$\cdots,$	$x_n^{(2)},$	$c^{(2)})$
	$\cdots$		$\cdots$		$\cdots$
	$(\mathbf{x}^{(N)}, c^{(N)}) \Leftrightarrow$	$(x_1^{(N)},$	$\cdots,$	$x_n^{(N)},$	$c^{(N)})$
Test	$(\mathbf{x}^{(N+1)}, ???) \Rightarrow$	$(x_1^{(N+1)},$	$\cdots,$	$x_n^{(N+1)},$	$c^{(N+1)})$

Matriz de confusión

- Evaluación: medida de la calidad de un clasificador
- Objetivo: problema de 2 clases: $\mathcal{C} = \{+, -\}$

Matriz de Confusión o Tabla de Contingencia

Real \ Estimado	+	-
+	TP	FN
-	FP	TN

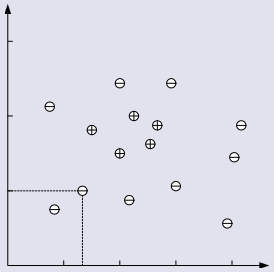


Figura: Problema de 2 clases

Figuras de mérito

Valores normalizados en la matriz de confusión:

TPR True Positive Rate $$TPR = \frac{TP}{TP + FN} \quad (1)$$

FNR False Negative Rate $$FNR = \frac{FN}{TP + FN} \quad (2)$$

FPR False Positive Rate $$FPR = \frac{FP}{FP + TN} \quad (3)$$

TNR True Negative Rate $$TNR = \frac{TN}{FP + TN} \quad (4)$$

Figuras de mérito típicas:

$$Precision = 100 \times \frac{TP}{TP + FP} \quad (5)$$

$$Recall = TPR = \frac{TP}{TP + FN} \quad (6)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

$$F - measure = \left[\frac{1}{2}\left(\frac{1}{Precision} + \frac{1}{Recall}\right)\right]^{-1} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (8)$$

Figuras de mérito

Ejercicio: discusión figuras de evaluación

Evaluar la calidad del clasificador en los siguientes casos:

1. 10.000 muestras (9.900 C-, 100 C+). El clasificador predice siempre C-
2. 10.000 muestras (9.900 C-, 100 C+). El clasificador predice siempre C+
3. 10.000 muestras (100 C-, 9.900 C+). El clasificador predice siempre C+
4. 10.000 muestras (100 C-, 9.900 C+). El clasificador predice siempre C-
5. Clasificador con la siguiente matriz de confusión asociada:

Estimado \ Real	+	-
+	500	500
-	100	10.000

Figuras de mérito

Ejercicio: Compara tu clasificador con predicciones aleatorias

La tabla 1 muestra la matriz de confusión de un clasificador. Supongamos que tenemos otro clasificador, uno que hace predicciones aleatorias pero que produce el mismo número de instancias de cada clase que el clasificador original.

Real \ Predicted	a	b	c	Total
a	88	10	2	
b	14	40	6	
c	18	10	12	
Total				

Tabla: MC de clasificador propuesto

Real \ Predicted	a	b	c	Total
a				
b				
c				
Total				

Tabla: MC de predicciones aleatorioas

1. ¿Cuál sería la matriz de confusión del clasificador aleatorio?
2. ¿Son mejores nuestras predicciones que hacerlas de forma aleatoria pero manteniendo las proporciones?

Figuras de mérito

Ejercicio: *Kappa statistics* [Witten et al., 2011, Sec. 5.7]

La tabla 3 muestra la matriz de confusión de un clasificador. Supongamos que tenemos otro clasificador, uno que hace predicciones aleatorias pero que produce el mismo número de instancias de cada clase que el clasificador original.

		Predicted a	b	c	Total
Real	a	88	10	2	
	b	14	40	6	
	c	18	10	12	
Total					

Tabla: MC de clasificador propuesto

		Predicted a	b	c	Total
Real	a				
	b				
	c				
Total					

Tabla: MC de predicciones aleatorioas

1. Accuracy del clasificador original y de las predicciones aleatorias
2. Calcula la estadística Kappa

Esquemas de validación

Evaluación por holdout

Evaluación basada en precisión: estimación de la probabilidad de clasificación correcta

Método no honesto: N-Train y N-test

$$\hat{p}_M = \frac{1}{N}\sum_{i=1}^{N}\delta(c^{(i)}, c_M^{(i)}) \quad (9)$$

Resubstitution error: utilizar la muestra completa (S) para inferir el modelo ($\mathcal{M}$) y evaluarlo con la misma muestra (S). ¡No! porque premia el sobreajuste (*over-fiting*) [Witten et al., 2011, Sec. 5.1]

Método *holdout*: N_1-Train y N_2-test siendo $N = N_1 + N_2$

$$\hat{p}_M = \frac{1}{N_2}\sum_{i=N_1+1}^{N_1+N_2}\delta(c^{(i)}, c_M^{(i)}) \quad (10)$$

- Se barajan las muestras y después se hacen dos particiones disjuntas: $\mathcal{S} = \mathcal{S}_{Train} \bigcup \mathcal{S}_{test} \quad : \quad \mathcal{S}_{Training} \bigcap \mathcal{S}_{test} = \emptyset$
- Se emplea la partición $\mathcal{S}_{Training}$ para estimar el modelo $\mathcal{M}$ y la partición $\mathcal{S}_{test}$ para evaluarlo
- Sólo se utilizan parte de los datos para estimar el modelo: $N_2 < N_1 < N$
- $N_1 \downarrow \implies$ sesgo (*bias*) en el estimador $\uparrow\uparrow$

Esquemas de validación

Evaluación por holdout

Método no honesto: N-Train y N-test

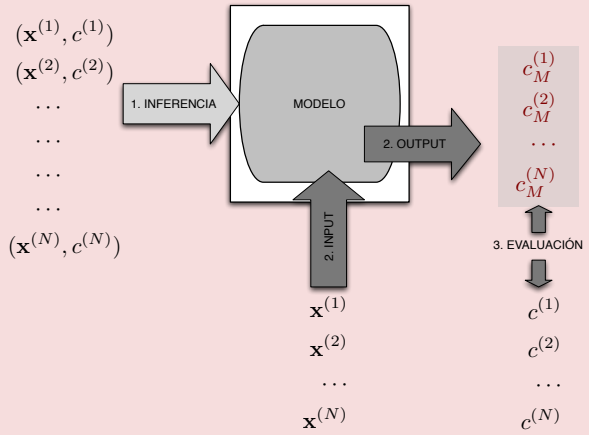


Figura: Premia sobre-ajuste

Esquemas de validación

Evaluación por holdout

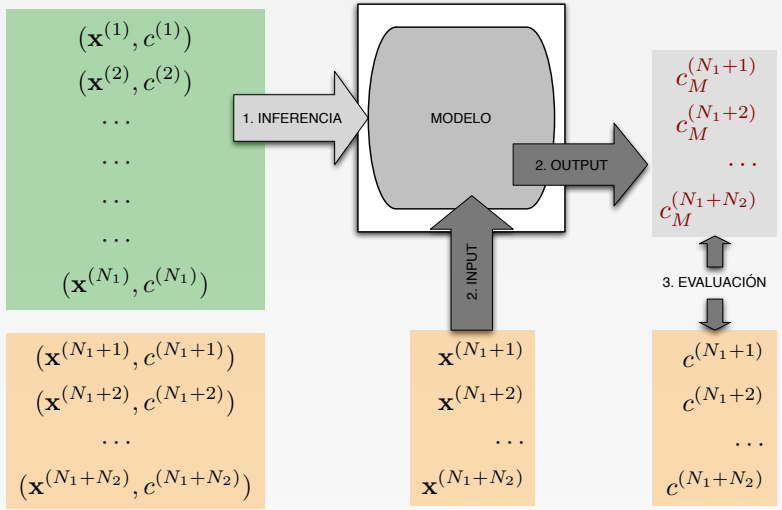


Figura: Método *holdout* de estimación: N_1-Train y N_2-test

Esquemas de validación
Evaluación por holdout

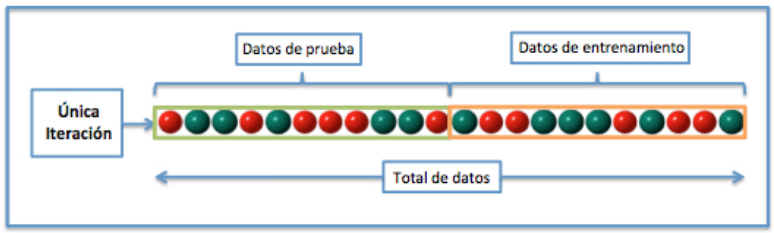


Figura: Método *holdout*[1]

[1] Fuente de la figura: Wikimedia-Commons.

Esquemas de validación
Evaluación mediante validación cruzada

Método *k-fold cross validation*

[Witten et al., 2011, Chap. 5, Cross validation]

- Se barajan las muestras y después se hacen k particiones disjuntas de la misma talla (aprox.): $\mathcal{S} = \mathcal{S}_1 \bigcup \mathcal{S}_2 \bigcup \cdots \mathcal{S}_k : \quad |\mathcal{S}_i| = \frac{N}{k}$
- Se repiten k procesos de entrenamiento (M_i) y test empleando (k-1) particiones para entrenar y la otra para test: $\forall i : 1 \leq i \leq k \; \mathcal{S}^i_{Train} = \mathcal{S} \bigcap (\mathcal{S}_i)^c$ y $\mathcal{S}^i_{test} = \mathcal{S}_i$

$$\hat{p}_M = \frac{1}{k} \sum_{j=1}^{k} \hat{p}_{M_j} \tag{11}$$

- Resultados similares $\Leftrightarrow$ Modelo estable $\Leftrightarrow$ No hay sesgo en el estimador
- En el límite cuando $k = N$: N-fold cross validation $\equiv$ *leaving one out*

Esquemas de validación
Evaluación mediante validación cruzada

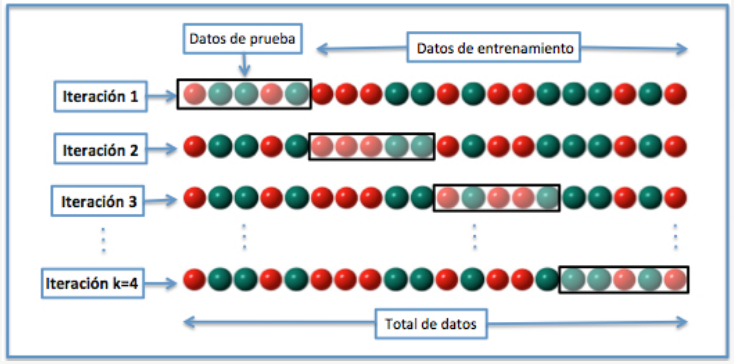


Figura: *k-fold cross validation*[2]

[2] Fuente de la figura: Wikimedia-Commons

Esquemas de validación
Evaluación mediante validación cruzada

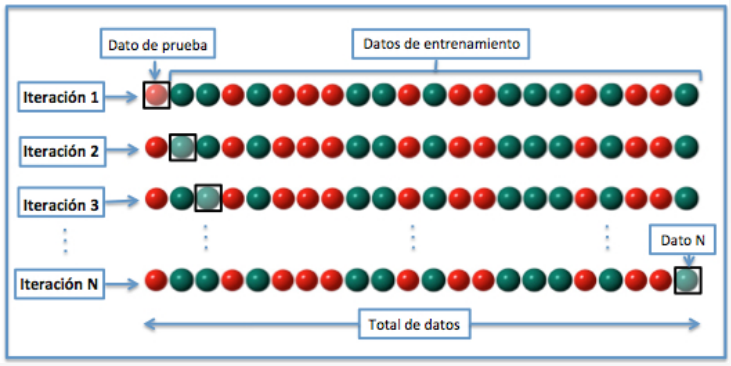


Figura: *Leave one out*[3]

[3] Fuente de la figura: Wikimedia-Commons

Esquemas de validación
Evaluación por bootstrap

Método 0.632 *bootstrapping*

[Witten et al., 2011, Chap. 5, The 0.632 bootstrap]

- En los casos anteriores: muestreo aleatorio sin reemplazo
- Se genera una muestra de *bootstrap* ($\mathcal{S}_{bootstrap}$) extrayendo N datos de forma aleatoria y con reemplazo del conjunto de muestra $\mathcal{S}$; $|\mathcal{S}| = N \Longrightarrow |\mathcal{S}_{bootstrap}| = N$
  - $p(s \in \mathcal{S}, s \notin \mathcal{S}_{bootstrap}) = \lim_{N\to+\infty} \left(1 - \frac{1}{N}\right)^N \approx e^{-1} \approx 0{,}368$
  - El número esperado de datos no repetidos: $0{,}632 \cdot N$
- Para estimar la probabilidad de clasificación correcta: $\mathcal{S}_{Train} = \mathcal{S}_{bootstrap}$ y $\mathcal{S}_{test} = \mathcal{S} \bigcap \mathcal{S}^c_{bootstrap}$
  - $|\mathcal{S}_{Train}| = N$
  - $|\mathcal{S}_{test}| \approx 0{,}368 \cdot N$
- Se repite el proceso anterior k veces ($M_1, M_2, \ldots, M_k$):

$$\hat{p}_M = \frac{1}{k}\sum_{j=1}^{k}\left(0{,}632 \cdot \hat{p}_{M_j}^{Train} + 0{,}368 \cdot \hat{p}_{M_j}^{test}\right) = 0{,}632 \cdot \hat{p}_M^{Train} + 0{,}368 \cdot \hat{p}_M^{test} \quad (12)$$

- Resultados similares $\Leftrightarrow$ Modelo estable $\Leftrightarrow$ No hay sesgo en el estimador

Esquemas de validación
Intervalos de confianza

- ¿Es fiable la estimación de $\hat{p}_M$?
- ¿Coincide p con $\hat{p}_M$? ¿con un margen de error?

$$p = \hat{p}_M \pm \Delta e_M \quad (13)$$

- El intervalo del error-verdadero a un nivel de confianza del c % siendo $N = |S|$:

$$e_M(c) = Z_c\sqrt{\frac{\hat{p}_M\,(1-\hat{p}_M)}{N}} \quad (14)$$

Donde Z_c es una constante relacionada con la distribución normal para el nivel de confianza:

Nivel de confianza (%c)	50 %	80 %	90 %	95 %	99 %
Z_c	0,67	1,28	1,64	1,6	2,58

Esquemas de validación
Observaciones sobre cada técnica en la práctica

Observaciones sobre cada técnica en la práctica:

- Holdout: se utiliza cuando N es grande para tratar de que el sesgo no sea grande
- k-fold cross validation:
  - ofrece $\hat{p}_M$ no sesgada :-)
  - sin embargo, la varianza es alta :-(
- Bootstrap:
  - ofrece $\hat{p}_M$ no sesgada en el límite :-)
  - la varianza es baja :-)
- Ver: [Witten et al., 2011, Sec. 5.5]

Esquemas de validación
Observaciones sobre cada técnica en la práctica

Ejercicios:

Dado el conjunto de datos `iris.arff` calcular media y varianza de la figura de mérito F1-measure para las técnicas de evaluación siguientes aplicadas en 5 experimentos con conjuntos de instancias ligeramente distintas (es decir, repitiendo el experimento 5 veces con los conjuntos de instancias generados mediante 5 *random-seeds* distintas). Se pide hacer este ejercicio con dos métodos de clasificación distintos (eg. `ZeroR`, `J48`, `BayesNet`, `MultilayerPerceptron`,...).

1. Hold-out con (Train, test) en proporción (90 %,10 %)
2. 10-fold cross validation (¿en qué proporción está (Train,test)?)
3. 0.632 bootstrap

Evaluación basada en coste

- Evaluación basada en precisión: todos los errores tienen la misma relevancia.
  - Directriz: modelo mejor cuanto menos errores asociados tenga
- En la práctica: cada error tiene asociado un coste distinto
- Ejemplos:
  - ¿Es *SPAM* este e-mail? Clasificar como *SPAM* un e-mail relevante suele ser más desfavorable que clasificar como no-*SPAM* un e-mail no deseado.
  - ¿Es arriesgado conceder un crédito a este cliente? Denegar un crédito a un cliente sin tiene distintas consecuencias que concedérselo a un cliente de riesgo, es decir, son errores con distinto coste.
  - Medicina: que el clasificador etiquete una prueba como posible enfermedad y que el médico tras haber estudiado la prueba diagnostique que no hay enfermedad es un error de distinto riesgo a que el clasificador etiquete una prueba como ausente de enfermedad porque en este caso posiblemente no se le daría al médico para corroborarlo y podría ser un error.
  - Detección de alarmas
  - Campañas de marketing
- Alternativa: Evaluación Basada en Coste (*CBE*)
  - Directriz: modelo mejor cuanto menor sea el coste asociado a los errores
  - Ver: [Witten et al., 2011, Sec. 5.7], [Orallo et al., 2004]

Evaluación basada en coste

Evaluación basada en coste

- Generalización más realista del aprendizaje predictivo
- Directriz: modelo mejor cuanto menor sea el coste asociado a los errores
  Objetivo: cometer pocos errores caros
- Matriz de coste entre la clase estimada y la clase real: $C_{i,j}$ coste de predecir clase i cuando la clase real es j
- Matriz de confusión: $m_{i,j}$ número de veces que se ha estimado la clase i cuando la clase real era j en un conjunto de N muestras ($N = |S|$).

$$Coste = \sum_{i=1}^{N} \sum_{j=1}^{N} C(i,j) \cdot N(i,j) \tag{15}$$

- Decisión óptima: asignar a cada ejemplo (x) la clase (i) con el menor coste asociado. La clase esperada para x:
  - $\hat{c}(x) = \arg \min_{i:1 \le i \le N_c} \left[\sum_{j=1}^{N_c} C(i,j) \cdot P(j|x) \right]$ siendo N_c el número de clases distintas.

Evaluación basada en coste

Ejemplo: dada la matriz de confusión asociada a distintos clasificadores y la matriz de coste asociada al problema, evaluar cada clasificador en términos de riesgo (coste).

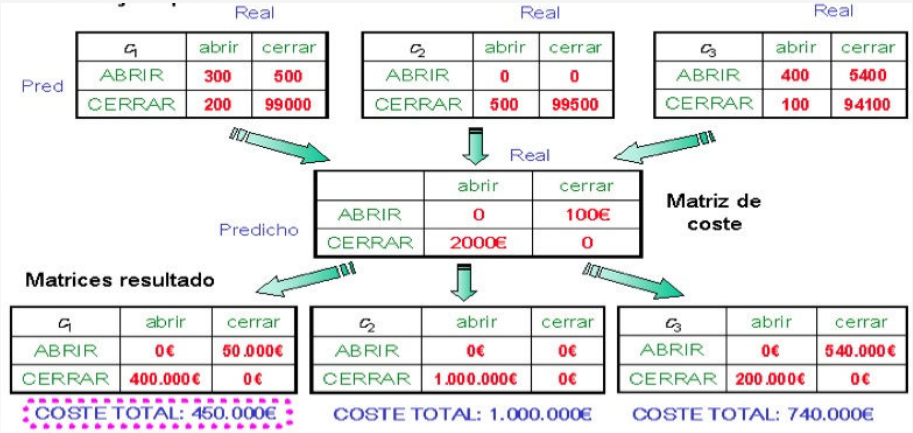


Figura: Matriz de confusión asociada a distintos clasificadores y matriz de coste del problema[4]

[4] Fuente de la figura: [Bielza-Larrañaga,Cap2.1]

Evaluación basada en coste

Ejercicios:

1. Evaluar cada clasificador en términos de precisión y comparar las conclusiones con respecto al caso anterior (evaluación basada en coste).
2. Proponer un clasificador mejor

Evaluación basada en coste

COMPARACIÓN DE CLASIFICADORES
Clasificación Supervisada

Análisis ROC

- Calidad del clasificador determinada en base a dos factores:
  1. Distribución de la clase
  2. Coste de los errores
- En ocasiones estos dos factores se desconocen!
- Alternativa: análisis ROC

ROC: Receiver Operating Characteristics

- Análisis ROC
- Caracterizar la calidad de los clasificadores en base a su rendimiento (e.g. diagnóstico clínico)
  1. TPR (Sensitivity, Hit Rate) eje y
  2. FPR (= 1-Specificity) eje x [Duda et al., 2000, Alpaydin, 2010]
- Criterios de selección:
  - Cuanto más arriba a la izquierda mejor
  - Cuanto mayor sea el área bajo la curva (*AUC*) tanto mejor
- Ver: [Orallo et al., 2004], [Witten et al., 2011, Chap. 5, ROC curves]

Análisis ROC

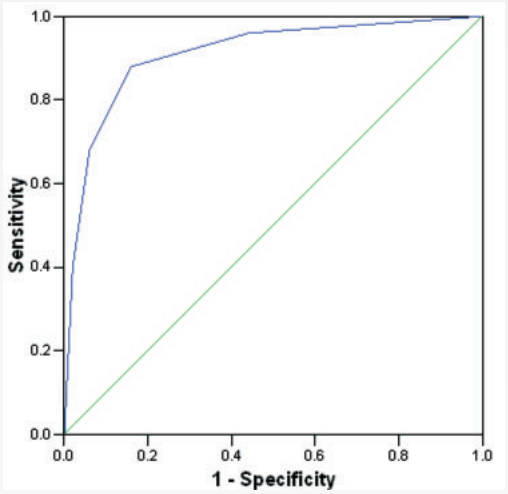


Figura: Curva ROC típica: TPR vs. FPR

Espacio ROC:
- TPR = Sensitivity
- FPR = 1-Specificity

Puntos límite:
- Clasificador en (0,0): predice todo como clase negativa
- Clasificador en (1,1): predice todo como clase positiva
- Puntos de la recta TPR=FPR: clasificador aleatorio (no-discriminante!)

Fuente de la figura: [Bielza and Larrañaga, 2012]

Análisis ROC

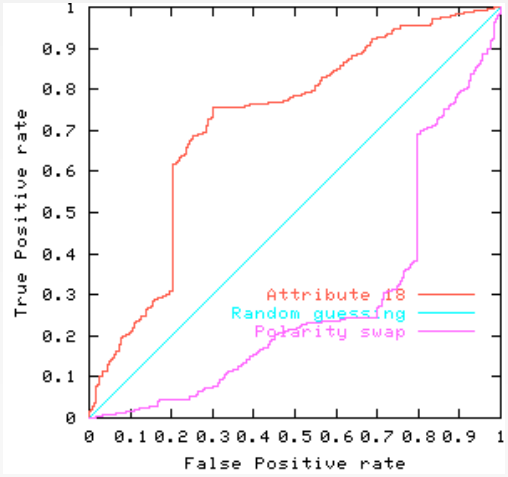
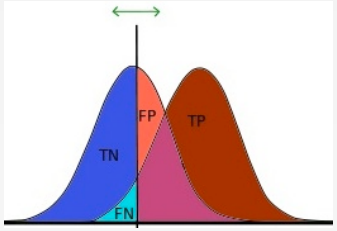


Figura: Worse than random guesser? Never!

Es posible que un clasificador ofrezca resultados peores que el clasificador aleatorio, sin embargo ¡es inadmisible!
- ¿Un resultado peor que el del clasificador aleatorio? Interpretación: la predicción del clasificador guarda cierta correlación con la realidad, pero la correlación es negativa (bajo la diagonal del espacio ROC)
- Solución: tomar la decisión contraria obteniendo así correlación positiva (resultado simétrico sobre la diagonal del espacio ROC)

Fuente de la figura: W. B. Langdon at cs.ucl.ac.uk

Análisis ROC



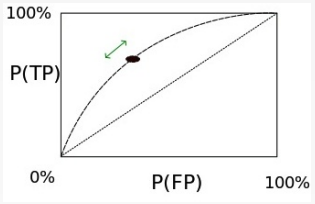


Figura: Modificar el punto de decisión en un clasificador.

TP	FP
FN	TN
1	1

Leer: [Duda et al., 2000, Sec. 2.8.3]

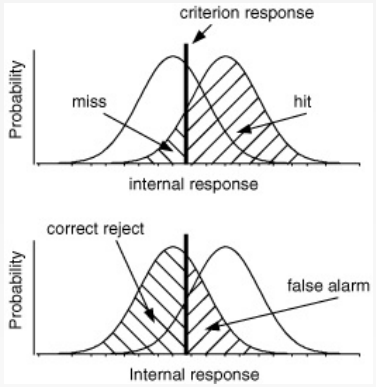
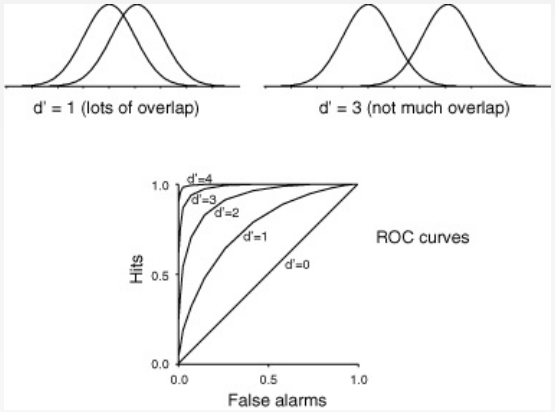


Figura: ROC: Signal to Noise Ratio

Fuente de la figura:D. Heeger

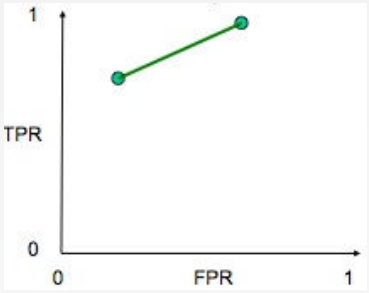
Análisis ROC



Curva ROC para distintos clasificadores (analogía Signal-to-Noise Ratio[5]): la forma de la curva ROC viene determinada por la intersección de las curvas-respuesta.

[5]Fuente de las figuras: D. Heeger

Análisis ROC



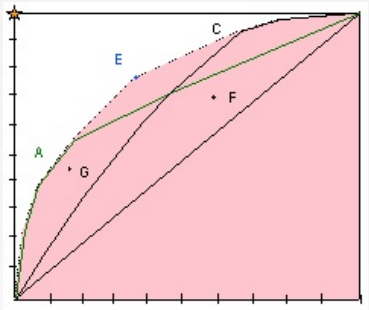


Figura: Construir un clasificador combinando dos clasificadores[6]

Conclusiones:

- El punto de operación en el espacio ROC, cuanto más *arriba a la izquierda*, mejor.
- Mejor clasificador cuanto mayor sea el área bajo la curva ROC (*Area Under the Curve*)

[6]Fuente de las figuras: [Bielza and Larrañaga, 2012] y Tom Fawcett respectivamente

Bibliografia I

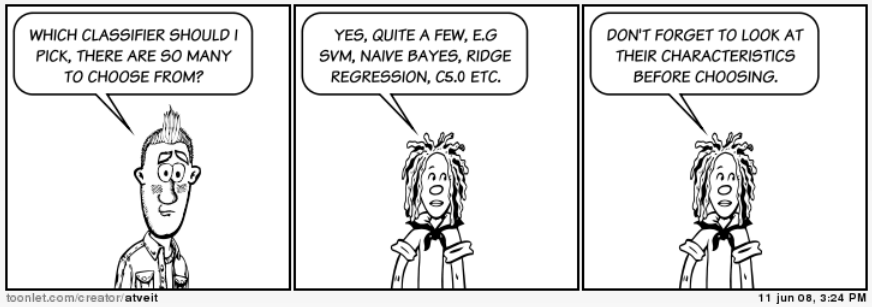
- Alpaydin, E. (2010).
  Introduction to Machine Learning.
  MIT Press.
- Bielza, C. and Larrañaga, P. (2012).
  Minería de datos: Métodos y técnicas.
  Master Universitario en Ingeniería Informática.
- Duda, R. O., Hart, P. E., and Stork, D. G. (2000).
  Pattern Classification.
  Wiley-Interscience.
- Orallo, J. H., Ramírez, M. J., and Ferri., C. (2004).
  Introducción a la Minería de Datos.
  Pearson Educación.
- Witten, I. H., Frank, E., and Hall, M. A. (2011).
  Data Mining: Practical Machine Learning Tools and Techniques.
  The Morgan Kaufmann Series in Data Management Systems, 3rd edition.

Parte II

Apéndice

Índice

Apéndice: ¿Qué clasificador seleccionar?



Fuente de la figura: `http://amundblog.blogspot.com.es/2008/06/pragmatic-classification-of-classifiers.html`