H-measure

Quantil Matemáticas Aplicadas
Miguel Bernal

30 de octubre de 2014

Contenido

Introducción

- ¿Cómo comparar objetivamente el rendimiento de varios modelos?

Medidas utilizadas

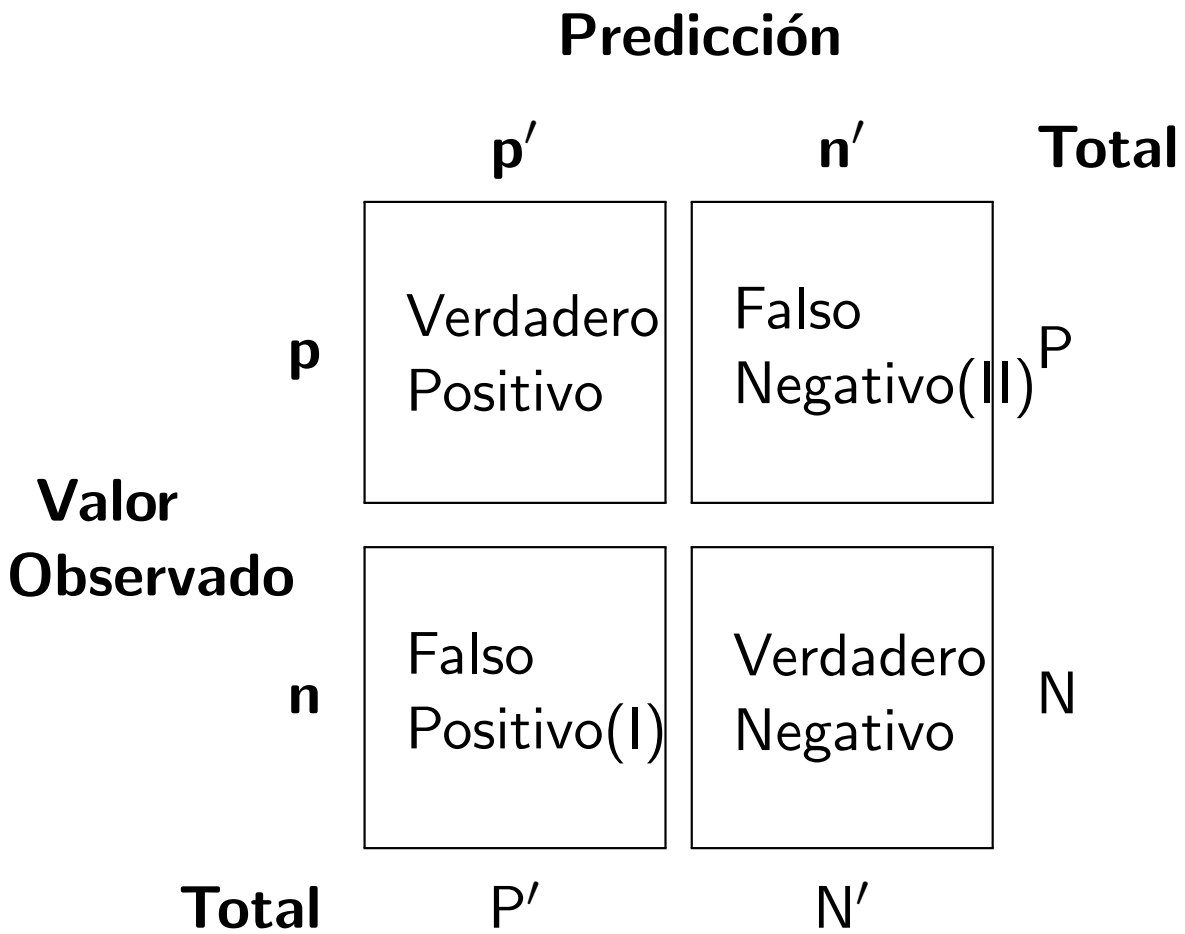
- RMSE, R^2, AIC, BIC, etc.
- Nos concentraremos en el área bajo la curva ROC: AUC
- Uno de los más populares.

Contenido

¿Qué es?

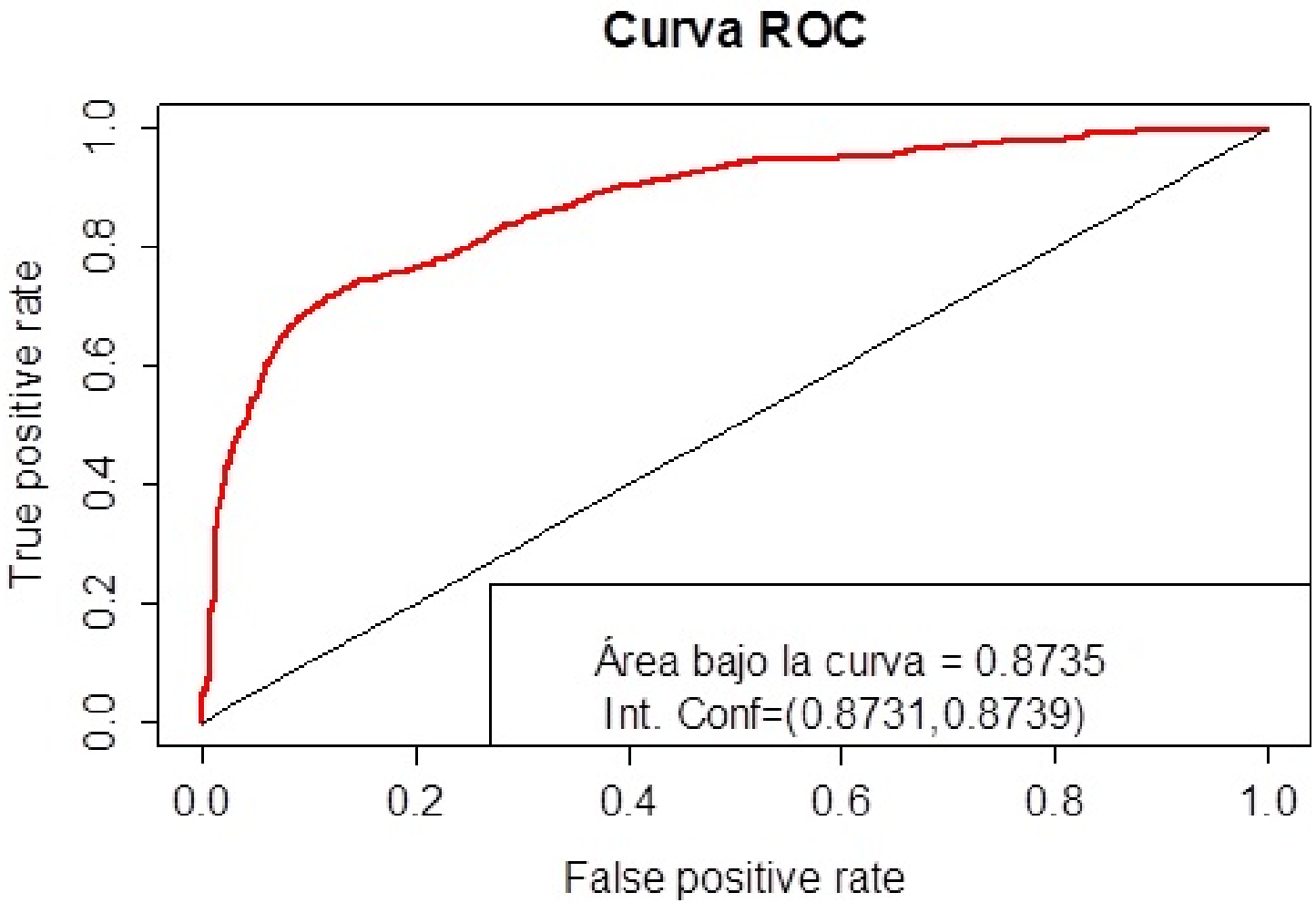
- Medida que cuantifica el rendimiento de clasificadores binarios.
- Teoría de detección de señales.
- Muy utilizada en minería de datos.
- Predicción de un clasificador binario depende del punto de corte(umbral).

Matriz de confusión



R. Verdaderos Positivos $= \frac{VP}{P}$ R. Falsos Positivos $= \frac{FP}{N}$

ROC


Curva ROC
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
Área bajo la curva = 0.8735
Int. Conf=(0.8731,0.8739)

Definición Formal

- Dos clases: 0,1
- $\hat{p}(1|x)$ o en general un puntaje $s(x)$
- Para la clase k
  1. Densidad: $f_k(s)$
  2. Acumulada: $F_k(s)$
  3. Prior: π_k $(\pi_0 + \pi_1 = 1)$
- Clasificación: Para un umbral t si $s > t$ se clasifica 1.
- Entonces: RVP=$F_0(s)$ y RFP=$F_1(s)$

AUC

- Denotando $v = F_1(s)$

$$AUC = \int_0^1 F_0(F_1^{-1}(v))\,\mathrm{d}v \tag{1}$$

$$AUC = \int_{-\infty}^{\infty} F_0(s) f_1(s)\,\mathrm{d}s \tag{2}$$

Interpretaciones

- Probabilidad que un miembro de la clase 0 escogido aleatoriamente tenga un score más bajo que un miembro de la clase 1 escogido aleatoriamente.
- GINI = 2AUC - 1
- Muchos otros indicadores asociados.

Bondades

- No hay que especificar un valor de t.
- Con los mismos datos todo el mundo obtiene el mismo valor.
- Medida objetiva que permite comparar diferentes clasificadores.

Problemas

- Si se cruzan las curvas?
- Modelos parecidos al ajustar.
- Otros conocidos:
- Lobo, J. M., Jiménez-Valverde, A. and Real, R. (2008), AUC: a misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography, 17. 145?151
- Problemota:

Problemas

- Si se cruzan las curvas?
- Modelos parecidos al ajustar.
- Otros conocidos.
- Problemota:

Incoherente!!!

¿Por qué?

- Escoger un t implica incurrir en un costo de clasificar mal.
- Para cada pareja de costos de clasificar mal se puede escoger un t óptimo que minimce el costo total.
- Generalmente no se conocen los costos exactos de clasificar mal, sólo cuál es mayor.
- Uno puede intentar construir una distribución de la razón de los costos de clasificar mal.
- Esto se traduce en una distribución sobre t.
- Para tener una medida de rendimiento se puede integrar la pérdida ponderada por la distribución de t.
- De hecho esto es lo que el AUC hace!!!

¿Por qué?

- Como la distribución de t corresponde a una distribución de la razón de costos, entonces el AUC es equivalente a promediar la pérdida de mal-clasificación sobre la distribución de la razón de costos que depende la distribución del score.
- La distribución del score depende del clasificador.
- **Entonces el AUC evalúa el rendimiento de un clasificador utilizando una métrica que depende del mismo clasificador!!!**
- Esto quiere decir que evalúa diferentes clasificadores con diferentes métricas.
- Como si midieramos la longitud de un objeto con una regla en centímetros y otra en pulgadas y las comparáramos. Peras con Manzanas.
- Problema muy profundo. (Prueba Formal en el artículo)
- Entonces?

Contenido

Preliminares

- Sea c_k el costo de clasificar mal un elemento de la clase k.
- $c_1 = c_0$ daría el error de clasificación.
- Para un t el costo total es:
- $c_0\pi_0(1 - F_0(t)) + c_1\pi_1(F_1(t))$
- Se puede escoger el t óptimo, independientemente de la escala de c_k.

Preliminares

- Lo importante es la razón de costos.
- Se puede transformar la pareja (c_0, c_1), en la pareja(b, c)
- $b = c_0 + c_1, c = \frac{c_o}{c_0+c_1}$ y así c sólo depende de la razón.
- Entonces la pérdida se puede escribir de manera general:

$$Q(t; b, c) = (c\pi_0(1 - F_0(t)) + c\pi_1(F_1(t)))b \tag{3}$$

- Entonces la pérdida general sería:

$$L = \int Q(T(c); b, c) w(c) \, \mathrm{d}c \tag{4}$$

- Donde $w(c)$ es la distribución que pondera las pérdidas asociadas a los diferentes valores de c.

H-Measure

- Función de peso ideal que no dependa del score
  $w(c) = \int bv(b, c)db$
- Pero desconocemos b y c.
- Si b y c son independientes y cambiando las unidades de b:
- $w(c) = u(c)$
- Cuál $u(c)$ usar? Uniforme?

H-Measure

- Propuesta:
- $u_{\alpha,\beta}(c) = beta(c; \alpha, \beta) = \frac{1}{B(\alpha,\beta)} c^{\alpha-1}(1-c)^{\beta-1}$
- $\alpha, \beta > 0, \ c \in (0,1)$
- $\beta(\alpha, \beta) = \int_0^1 c^{\alpha-1}(1-c)^{\beta-1}\mathrm{d}c$
- Entonces la pérdida general sería:

$$L_{\alpha,\beta} = \int Q(T(c); b, c) u_{\alpha,\beta} \, \mathrm{d}c \tag{5}$$

H-Measure

- α y β?
- Si $c_1 > c_0$ entonces $\alpha = 2$, $\beta = 4$.
- Para ser consistente $\alpha = 2 = \beta$. (Arbitario)
- En clases desbalanceadas?

Clases desbalanceadas

- Moda de la distribución $Beta(\alpha, \beta) = \frac{\alpha-1}{\alpha+\beta-2}$
- Queremos que sea igual a π_1.
- Varias formas de hacerlo.
- Se propone $\alpha = \pi_1 + 1, \beta = \pi_0 + 1$. (Arbitario pero tiene sentido)

- Queremos que más sea mejor y que esté entre 0 y 1.
- Para normalizar, el peor caso:

$$L_{Max} = \pi_0 \int_0^{\pi_1} cu(c)\,\mathrm{d}c + \pi_1 \int_{\pi_1}^{1} (1-c)u(c)\,\mathrm{d}c \qquad (6)$$

$$H = 1 - \frac{L_{\alpha,\beta}}{L_{Max}} = 1 - \frac{\int Q(T(c); b, c) u_{\alpha,\beta} \, \mathrm{d}c}{\pi_0 \int_0^{\pi_1} c u(c) \, \mathrm{d}c + \pi_1 \int_{\pi_1}^{1} (1 - c) u(c) \, \mathrm{d}c} \quad (7)$$

Estimación

- Similar al AUC, ver en detalle sección 7 del artículo.
- Implementado recientemente en R. Paquete "hmeasure".

Ejemplo

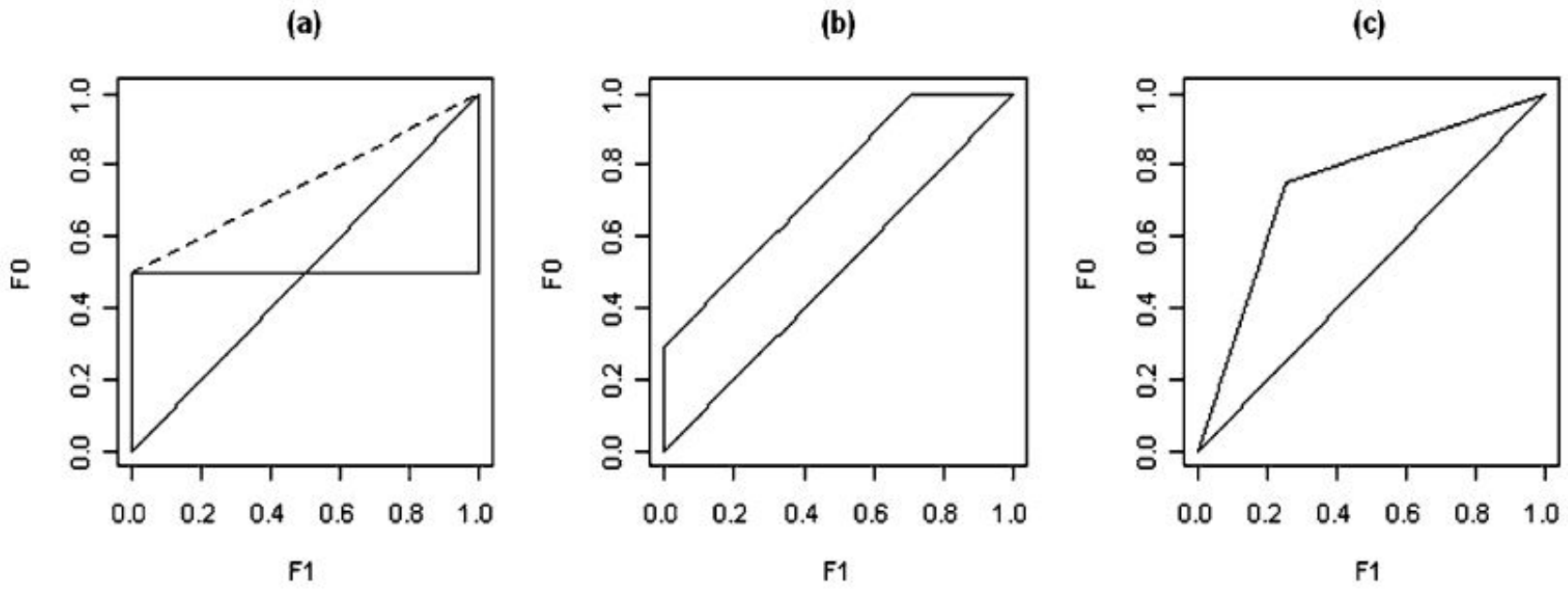


Fig. 3 Examples of three ROC curves

Ejemplo

	(a)	(b)	(c)
AUC	0.5	0.75	0.75
G	0	0.5	0.5
AUCH	0.75	0.75	0.75
H	0.348	0.293	0.288

Ejemplo

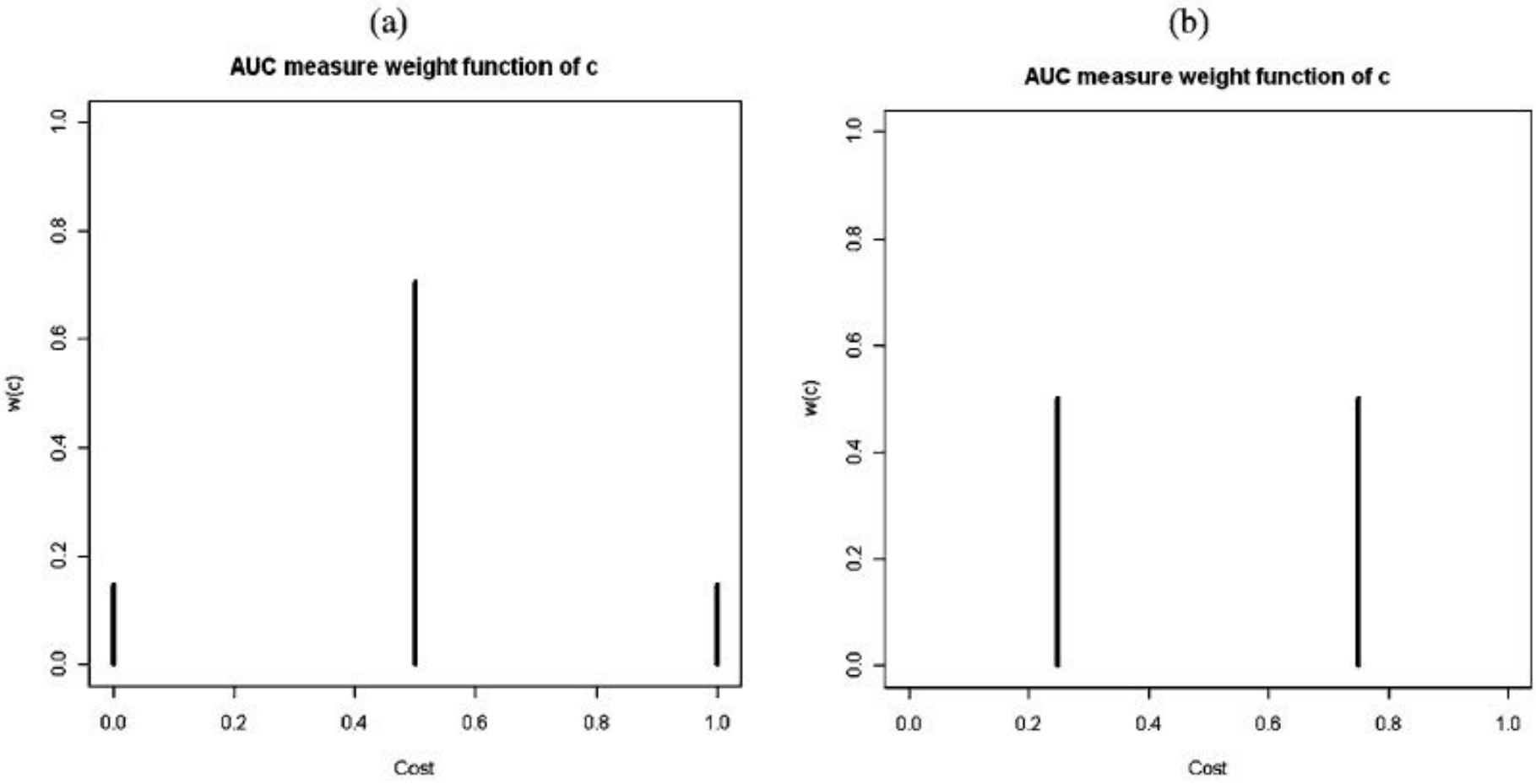


Fig. 4 The $w(c)$ functions corresponding to the ROC curves in Figs. 3(b) and (c)

Contenido

Conclusión

- AUC tiene sus bondades, pero tiene un grave problema descubierto recientemente.
- Mide los modelos con varas diferentes.
- Alternativa: Fijar una distribución de los costos, propuesta que sea *Beta*.
- Problema: A menos que se sepan exactamente los costos del problema (raro), la distribución es arbitraria pero al menos permite hacer una comparación con sentido.