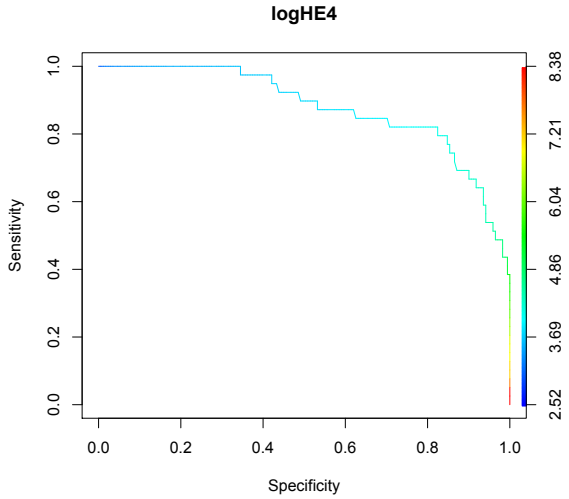
Università degli Studi di Trieste

603SM – Biostatistica

Corso di Laurea Magistrale in Biotecnologie Mediche

La curva ROC con R



Autore
Massimo Borelli, Ph.D.

12 dicembre 2016

Indice

1 La curva ROC in poche parole

In molte condizioni cliniche o sperimentali abbiamo la necessità di trarre una decisione di tipo sì/no sulla base di una misura continua; in altri termini, vogliamo fissare un cut-off nella maniera più opportuna in modo da evitare errori del tipo 'falso positivo' e 'falso negativo'. La curva ROC è uno strumento grafico che consente di scegliere tale cut-off; al contempo, essa consente di valutare anche le attendibilità di due distinti metodi predittivi. Prendiamo a riferimento il dataset `ovarian` nel quale 210 donne di differenti età hanno avuto una neoplasia ovarica con due possibili outcome, fausto ed infausto:

	HE4	CA125	CA199	CEA	ETA	MENOPAUSA	OUTCOME
1	3540	6978	2749	88	34	PRE	BENIGNO
2	3010	23180	12620	91	21	PRE	BENIGNO
3	29340	11190	2426	214	64	POST	MALIGNO
4	6260	5188	3414	547	58	POST	MALIGNO
5	3510	2033	60	244	74	POST	BENIGNO
6	4020	6048	3087	142	40	PRE	BENIGNO
7	130420	196300	1120	119	51	PRE	MALIGNO
8	3510	1160	396	74	21	PRE	BENIGNO
..	..	..	..	..	..	..	..
209	5780	867	62	110	55	POST	BENIGNO
210	5190	5580	494	167	63	POST	MALIGNO

Si sa in letteratura, ad esempio, che il marker HE4 è un valido predittore della malignità delle neoplasie ovariche. La questione potrebbe essere quella di scegliere un valore soglia del marker al di sotto del quale o al di sopra del quale predire la benignità o la malignità del tumore. Importiamo innanzitutto il dataset in R:

```
www = "http://www.dmi.units.it/~borelli/dataset/ovarian.csv"
ovarian  = read.csv(www, header = TRUE)
attach(ovarian)
logHE4 = log(HE4/100)
```

e definiamo la misura `logHE4`.

2 Cenni di probabilità

Facciamo un breve riepilogo su alcuni concetti base di probabilità, creando una **tavola di contingenza**:

```
table(MENOPAUSA, OUTCOME)
```

	BENIGNO	MALIGNO	somma
POST	65	27	92
PRE	106	12	118
somma	171	39	210

Osservando che a 39 donne rispetto al totale di 210 è stato diagnosticato un tumore maligno, possiamo stimare che la prevalenza della patalogia nella popolazione da cui è stato tratto questo campione si attesta attorno al 19 per cento (i.e. una probabilità pari a $39/210 = 0.186$):

$$Pr(maligno) = \frac{39}{210} = 0.186...$$

Questa stima ovviamente non distingue il fatto che le donne siano o meno nel periodo fertile della loro vita, e viene detta anche **probabilità marginale**. Considerando invece il sottoinsieme delle donne che non sono ancora in menopausa, osserviamo che la probabilità di veder diagnosticato un tumore maligno scende attorno al 10 per cento (i.e. una probabilità pari a $12/118 = 0.102$). È questa una **probabilità condizionata**, o come diciamo in gergo, una probabilità *within* e scriviamo:

$$Pr(maligno|premenopausa) = \frac{12}{118} = 0.102...$$

Il fatto che queste due probabilità appena calcolate ci appaiano diverse e non uguali tra loro (l'una è praticamente la metà dell'altra) ci fa intuire che il sapere se una donna sia in premenopausa o in postmenopausa è una informazione rilevante, e non inutile, sulla previsione della malignità del tumore. Detto in altri termini, abbiamo motivo di ritenere che la menopausa e l'outcome non siano **eventi indipendenti** tra loro (in tal caso, in linea di principio, le probabilità dovrebbero essere uguali tra loro) ma siano due **eventi associati**: conoscere l'uno modifica il grado di fiducia che abbiamo sul verificarsi dell'altro.

Per inciso, riguardando la tavola di contingenza, spieghiamo cosa significa misurare la associazione di due variabili per mezzo dell'**odds ratio**, definito in questo modo:

$$O.R. = \frac{65 \cdot 12}{27 \cdot 106}$$

Abbiamo visto che le donne in premenopausa hanno una probabilità di 12/118 di avere un tumore maligno. Detto in termini anglosassoni, le donne hanno un *odds* di 12 contro 106. Analogamente, le donne in postmenopausa hanno un *odds* di 27 contro 65. Il rapporto dei due *odds*, appunto, è:

$$O.R. = \frac{\frac{12}{106}}{\frac{27}{65}} = \frac{65 \cdot 12}{27 \cdot 106} = 0.273$$

In linea di principio, se due eventi non sono associati il loro odds ratio è pari ad 1.

Vogliamo infine ricordare che le probabilità marginali e le probabilità condizionate sono legate tra loro da una relazione detta **teorema di Bayes**:

$$P(maligno|premenopausa) = \frac{P(premenopausa|maligno)}{P(premenopausa)} \cdot P(maligno)$$

Limitiamoci dunque a verificare tale relazione utilizzando i dati della tabella:

	BENIGNO	MALIGNO	somma
POST	65	27	92
PRE	106	12	118
somma	171	39	210

$$\frac{12}{118} = \frac{(12/39)}{(118/210)} \cdot \frac{39}{210}$$

$$\frac{12}{118} = \frac{12}{39} \cdot \frac{210}{118} \cdot \frac{39}{210}$$

$$\frac{12}{118} = \frac{12}{118}$$

2.1 Rischio Relativo e Odds Ratio: facciamo chiarezza, una volta per tutte

Ci è capitato di sentire, parecchie volte, una questione che riguarda il fatto che O.R. e R.R sono sì due misure adatte per confrontare i rapporti di proporzioni, ma una è adatta agli studi di coorte, mentre l'altra è adatta negli studi caso-controllo. Forse, questa idea si è diffusa da un passaggio del celebre testo di Martin Bland, che riportiamo nella sua traduzione italiana:

> sintomi è 44/1046 = 0.04207. Per confrontare i rischi di persone esposte o non esposte ad un certo fattore di rischio, si calcola il rapporto tra il rischio di quelli esposti e il rischio di quelli non esposti, quantità che prende il nome di **rischio relativo**. Il rischio relativo nel caso del nostro esempio è 2.26. Per stimare il rischio relativo in maniera diretta, è necessario uno studio di coorte (§3.7) come quello riportato in Tabella 8.3. Per studi caso-controllo, invece, il rischio relativo si calcola in modo diverso (§13.7).

> L'odds ratio può essere usato come stima del rischio relativo negli studi caso-controllo. Il calcolo del rischio relativo in §8.6 dipendeva dalla possibilità di effettuare le stime dei rischi. In quel caso potevamo farlo perché si trattava di uno studio prospettico, quindi eravamo a conoscenza del numero di individui, all'interno del gruppo esposto al fattore di rischio, che aveva manifestato i sintomi. Questo non può essere fatto, invece, qualora si parta dall'outcome (in questo caso gli episodi di tosse) e si cerchi di risalire al fattore di rischio (la bronchite), cioè nel caso degli studi caso-controllo.

Qui vogliamo mostrare che da un punto di vista algebrico O.R. e R.R. sono 'legati' tra loro da una semplice relazione, che permette di scambiare l'uno con l'altro (a patto di conoscere, all'interno del gruppo dei malati, la proporzione di quelli che sono stati esposti al rischio).

Ricordiamo le due definizioni:

	esposti	non esposti	somma
malati	a	b	a+b
sani	c	d	c+d
somma	a+c	b+d	N = a+b+c+d

$$OR = \frac{ad}{bc}$$

$$RR = \frac{a}{a+b} \cdot \frac{c+d}{c} = \frac{a}{c} \cdot \frac{(d+c)}{(b+a)}$$

Siccome risulta:

$$\frac{(d+c)}{(b+a)} = \frac{d}{b} + \frac{cb-ad}{b \cdot (a+b)}$$

sostituendo nella definizione di RR otteniamo:

$$RR = \frac{a}{c} \cdot \left(\frac{d}{b} + \frac{cb-ad}{b \cdot (a+b)} \right)$$

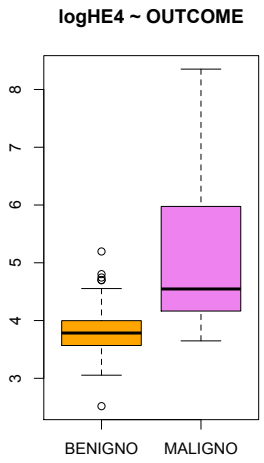
$$RR = OR + \frac{a}{c} \cdot \frac{cb-ad}{b \cdot (a+b)}$$

$$RR = OR + \frac{a}{c} \cdot \frac{bc \cdot (1-OR)}{b \cdot (a+b)}$$

$$RR = OR + \frac{a}{a+b} \cdot (1-OR)$$

$$RR = OR + P(esposti|malati) \cdot (1-OR)$$

3 Sensibilità, specificità e valori predittivi



La letteratura afferma che il marker HE4 è un predittore attendibile della malignità del tumore ovarico. Nella figura accanto osserviamo infatti che (in scala logaritmica) valori molto alti del marker sono sicuramente associati ad un esito maligno, e valori molto bassi sono associati con certezza all'esito benigno. Nella situazione intermedia, per esempio nella zona del valore 4, possono coesistere entrambi gli eventi. Siamo perciò interessati a fissare un cut-off in maniera tale che si minimizzino gli errori di previsione (anche se siamo ben consci dei limiti questa scelta: un falso positivo significa intal caso creare allarme in una paziente che poi risulterà essere non patologica, mentre un falso negativo significa, pericolosamente, dichiarare sana una ammalata).

Proviamo dunque a scegliere il cut-off pari a 4.1, dichiarando quindi positive al test le donne che hanno valori di `logHE4` maggiori o uguali a 4.1.

```
cutoff = 4.1
(tabella = table ( logHE4 < cutoff , OUTCOME ))
```

	BENIGNO	MALIGNO
≥ 4.1	30	32
< 4.1	141	7

Con questa scelta, commettiamo $30 + 7$ errori (in particolare, 7 false negative). Sappiamo che si definiscono rispettivamente la **sensibilità** e la **specificità** del test diagnostico le due probabilità condizionate:

$$Sens = Pr(positive|maligno) = \frac{32}{39}$$
$$Spec = Pr(negative|benigno) = \frac{141}{171}$$

Con questa scelta, i valori di sensibilità e specificità si attestano entrambi attorno all'82%:

```
(sensi = tabella[3] / ( tabella[3] + tabella[4]))
(speci = tabella[2] / ( tabella[1] + tabella[2]))
```

Dualmente, il **valore predittivo positivo** ed il **valore predittivo negativo** del test sono le probabilità condizionate:

$$VP+ = Pr(maligno|positive) = \frac{32}{62}$$
$$VP- = Pr(benigno|negative) = \frac{141}{148}$$

```
(vpredpos = tabella[3] / ( tabella[3] + tabella[1])) # 0.516129
(vpredneg = tabella[2] / ( tabella[2] + tabella[4])) # 0.9527027
```

Ecco invece cosa sarebbe successo se avessimo scelto un cut-off più alto, per esempio 5, oppure più basso, per esempio 3:

- cut-off = 5

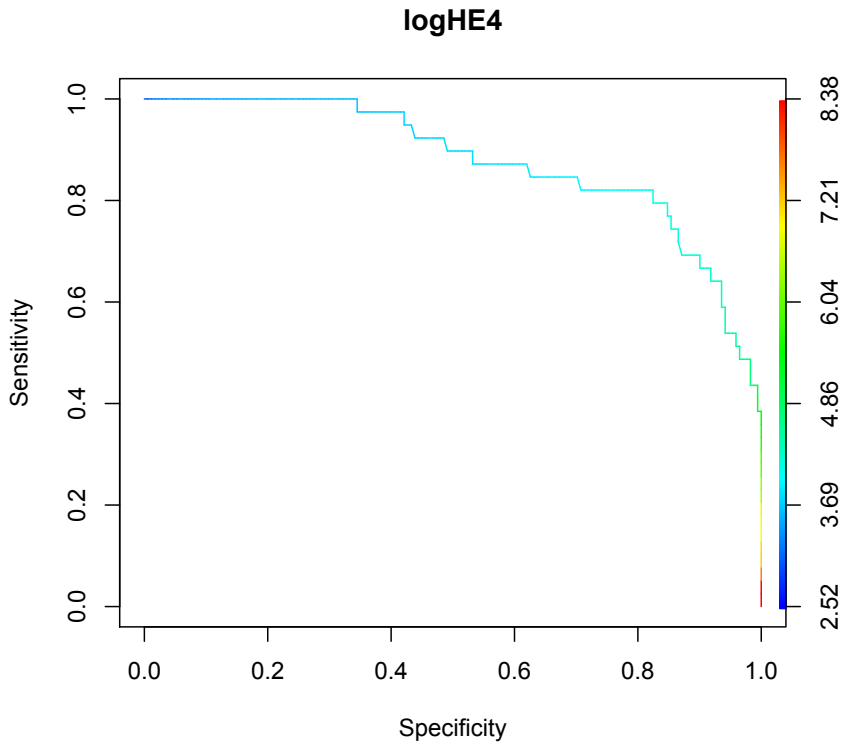
	BENIGNO	MALIGNO
positivo	1	16
negativo	170	23

- cut-off = 3

	BENIGNO	MALIGNO
positivo	170	39
negativo	1	0

Nel primo caso la specificità sarebbe cresciuta di molto, a 0.99, ma la sensibilità sarebbe scesa a 0.41; nel secondo caso, la specificità invece sarebbe scesa a 0, ma la sensibilità sarebbe salita ad 1. Questo fatto mostra come vi sia un trade-off tra il voler ottenere ottime sensibilità senza pagare in specificità, e viceversa.

La curva ROC descrive proprio questo fatto:




```
library(ROCR)
predizione <- prediction( logHE4, OUTCOME )
risultato <- performance( predizione , "sens", "spec")
plot(risultato , colorize=TRUE , main = "logHE4")
```

Abbiamo ottenuto la curva usando i comandi `prediction` e `performance` del pacchetto ROCR. I punti di colore azzurro sono relativi ad una scelta di cut-off molto bassa, prossima a valori di 2 e 3: garantiscono che non ci sono falsi negativi, condizione desiderabile. Al contrario, i punti di colore giallo aranciato provengono da scelte di cut-off molto elevate, che non sono di nostro interesse. Un compromesso potrebbe essere quello di cercare il punto della curva ROC che si avvicina maggiormente (in termini di distanza) all'angolo in alto a destra del grafico.

Per determinare tale punto ci conviene scrivere la funzione `migliorecutoff`:

```
migliorecutoff <- function(perf)
{Posizione_cut_off = which((risultato@x.values[[1]]-
```

```
risultato@y.values[[1]])==min((risultato@x.values[[1]]-
risultato@y.values[[1]])[risultato@x.values[[1]]-risultato@y.values[[1]]>0]))
return(risultato@alpha.values[[1]][Posizione_cut_off])  }
```

e scoprire che, per l'appunto, il miglior compromesso si ottiene fissando il cut-off pari a 4.1, come si diceva poco fa. Da ultimo si può quantificare, in un senso relativo (ad esempio per comparare due test diagnostici tra loro), la bontà di un test valutando l'area delimitata dalla curva ROC, che viene di solito indicata con la sigla AUC (*area under the curve*) e che in questo caso vale circa 0.88:

```
performance(predizione, "auc")@y.values[[1]]
```