MMLM: Una función para construir modelos predictivos con mayor capacidad de discriminación

Luis Mariano Esteban[1], Gerardo Sanz[2] y Ángel Borque[3]

[1]Escuela Universitaria Politécnica La Almunia. Universidad de Zaragoza.
[2]Dpto. Métodos Estadísticos. Universidad de Zaragoza.
[3]Hospital Universitario Miguel Servet.

III Jornada de usuarios de R

Universidad Zaragoza 1542

Construcción del modelo óptimo bajo AUC criterio.

- El propósito del algoritmo es seleccionar una buena regla de discriminación entre dos estados de una enfermedad.
- Curva ROC: Es un gráfico de la Sensibilidad en función de la Especificidad obtenida para distintos puntos de corte.
- Área bajo la curva ROC (AUC): Mide la capacidad de un clasificador para discriminar entre dos estados de una enfermedad.
- Estimación no paramétrica del área
  n_D pacientes con estado enfermedad $D = 1$, $n_{\bar{D}}$ con $D = 0$
  Resultados del test: $\mathbf{Y}_{D_1}, \ldots, \mathbf{Y}_{Dn_D}$ y $\mathbf{Y}_{\bar{D}1}, \ldots, \mathbf{Y}_{\bar{D}n_{\bar{D}}}$.
  Regla de clasificación: $L(\mathbf{Y})$

$$\widehat{AUC} = \frac{\sum_{i=1}^{n_D} \sum_{j=1}^{n_{\bar{D}}} I(L(\mathbf{Y}_{Di}) > L(\mathbf{Y}_{\bar{D}j})) + \frac{1}{2} I(L(\mathbf{Y}_{Di}) = L(\mathbf{Y}_{\bar{D}j}))}{n_D \cdot n_{\bar{D}}}$$

Universidad Zaragoza 1542

Modelos lineales como regla de clasificación

Cuando un conjunto de variables $\mathbf{Y} = (Y_1, \ldots, Y_n)$ es medible para cada paciente, se puede considerar la regla

$$L(\mathbf{Y}) = Y_1 + \beta_2 \cdot Y_2 + \ldots + \beta_n \cdot Y_n$$

a la cual le corresponde el mismo AUC que el modelo más complejo

$$L_g(\mathbf{Y}) = g(\beta_0 + \beta_1 \cdot Y_1 + \ldots \beta_n \cdot Y_n)$$

donde la función link g es cualquier función monótona creciente

Universidad Zaragoza 1542

Estimación de los parámetros del clasificador lineal

- Cuando $\mathbf{Y}$ sigue una distribución normal multivariante en cada una de las poblaciones, enferma-sana, la función discriminante lineal maximiza el AUC (Su, Liu, 1993).
- La regresión logística estima los parámetros del modelo por máxima verosimilitud y da buenos resultados en diferentes escenarios.
- `SLM` es una función de R que estima los parámetros del modelo con un algoritmo paso a paso basado en una búsqueda extensiva.
- `SLM` actua sobre objetos de tipo data.frame y estima los parámetros del modelo lineal que corresponden a un valor de AUC máxima.

Universidad Zaragoza 1542

Descripción básica SLM

Objetivo: Estimar $(\beta_2, \ldots, \beta_n)$ in $L(\mathbf{Y}) = Y_1 + \beta_2 \cdot Y_2 + \ldots + \beta_n \cdot Y_n$ que corresponde al máximo AUC.

Método: Estimar la combinación de dos variables $L(\mathbf{Y}) = Y_i + \beta_j \cdot Y_j$ que corresponde al máximo AUC.

Buscar la nueva variable Y_k cuya combinación $L(\mathbf{Y}) = Y_i + \beta_j \cdot Y_j + \beta_k \cdot Y_k$ corresponde a el máximo AUC.

Repetir el proceso hasta que todas las variables sean incluidas en el modelo.

Universidad Zaragoza 1542

Selección de β

- β es elegido en un conjunto finito de valores.
- Seleccionar β en $[-1, 1]$ nos da todas las posibles combinaciones de (Y_i, Y_j), porque AUC en $(Y_i + \beta \cdot Y_j)$ para $\beta < -1$ y $\beta > +1$ es la misma que $\alpha \cdot Y_i + Y_j$ donde $\alpha = \frac{1}{\beta} \in [-1, +1]$.
- La función genérica usa 201 valores igualmente espaciados de β en $[-1, 1]$.
- Seleccionar un conjunto mayor de valores puede dar un número alto de modelos equivalentes (mismo AUC).
- Seleccionar un conjunto menor de valores puede dar lugar a pérdida de precisión en el cálculo del máximo AUC.

Universidad Zaragoza 1542

Max-min combination

- Consideramos k biomarcadores cuyas componentes denotamos como $M = (M_1, \ldots, M_k)$.
- Sean $M_{\text{máx}} = \underset{1 \leq i \leq k}{\text{máx}} M_i$ y $M_{\text{mín}} = \underset{1 \leq i \leq k}{\text{mín}} M_i$.
- La combinación max-min de los biomarcadores es la combinación lineal $M_\lambda = M_{\text{máx}} + \lambda M_{\text{mín}}$, donde λ es el parámetro que debe ser estimado por la función `MMLM`.

Universidad Zaragoza 1542

Max-min combination

- Si suponemos una variable D, que define los estados de una enfermedad, siendo $\mathbf{X} = (\mathbf{X_1}, \ldots, \mathbf{X_k})$ los valores de los biomarcadores para un individuo con estado $D = 1$ y $\mathbf{Y} = (\mathbf{Y_1}, \ldots, \mathbf{Y_k})$ los valores de los biomarcadores para un individuo con estado $D = 0$
- La combinación Max-min viene dada por el valor λ_{opt} que maximiza el área, es decir

$$A(\lambda_{opt}) = \max_{\lambda} Pr\{(Y_{\text{máx}} - X_{\text{máx}}) + \lambda(Y_{\text{mín}} - X_{\text{mín}}) > 0\}$$

- En la función `MMLM` los valores de λ se calculan numéricamente con la función `SLM` anteriormente descrita.

Universidad Zaragoza 1542

Simulaciones en un problema de estadíficación de cáncer de próstata

Para analizar el comportamiento de la función `MMLM` se ha tomado dos variables sobre una base de datos real de cáncer de próstata y se han simulado otro tipo de marcadores continuos.

Staging prostate cancer database (n=621)				
PSA				
	Mean	Median	SD	Q1-Q3
All cases (n=621)	15.70	8.64	32.718	5.90-15.10
Organ-confined (n=369)	8.782	7.200	5.635	5.32-10.63
NonOrgan-confined (n=252)	25.840	13.800	49.236	7.49-25.62
Rate of cylinder affected				
	Mean	Median	SD	Q1-Q3
All cases (n=621)	38.230	33.330	26.169	16.67-50.00
Organ-confined (n=369)	28.210	25.000	18.368	12.50-37.50
Nonorgan-confined (n=252)	52.900	50.000	28.869	30.00-75.00

Simulation results in a Staging prostate cancer database (1000 sim.)

M_1 =PSA, M_2 = Rate of cylinder affected in the Biopsy

M_3, M_4, M_5 under multivariate normality

($\mathbf{X}$: Non Organ-confined, $\mathbf{Y}$: Organ Confined) with:

$\mathbf{X} = (0{,}5, 0{,}7, 1)$; $\mathbf{Y} = (0, 0, 0)$; $n_X = 252$, $n_Y = 369$

$$\sigma_X = \begin{pmatrix} 1 & 0{,}5 & 0{,}5 \\ 0{,}5 & 1 & 0{,}5 \\ 0{,}5 & 0{,}5 & 1 \end{pmatrix}; \sigma_Y = \begin{pmatrix} 1 & 0{,}5 & 0{,}5 \\ 0{,}5 & 1 & 0{,}5 \\ 0{,}5 & 0{,}5 & 1 \end{pmatrix}$$

AUC	AUC Mean	AUC Median	AUC SD	AUC 95 % C.I.
SLM	0.8730	0.8734	0.0113	0.8508-0.8945
Max-Min	0.8184	0.8194	0.0163	0.7860-0.8493

$$\sigma_X = \begin{pmatrix} 1 & 0{,}5 & 0{,}5 \\ 0{,}5 & 1 & 0{,}5 \\ 0{,}5 & 0{,}5 & 1 \end{pmatrix}; \sigma_Y = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

AUC	AUC Mean	AUC Median	AUC SD	AUC 95 % C.I.
SLM	0.8657	0.8652	0.0082	0.8505-0.8830
Max-Min	0.8412	0.8424	0.0143	0.8110-0.8662

Simulaciones sobre otras distribuciones(I)

Simulaciones sobre otras distribuciones (1000 simulaciones)

Log-transformed levels of M_1, M_2, M_3, M_4, M_5 under multivariate norma
($\mathbf{X}$: Non Organ-confined, $\mathbf{Y}$: Organ Confined) with:

$$\mathbf{X} = (0{,}5, 0{,}6, 0{,}7, 0{,}8, 1);\ \mathbf{Y} = (0, 0, 0, 0, 0);\ n_X = 250,\ n_Y = 250$$

$$\sigma_X = \begin{pmatrix} 1 & 0{,}5 & 0{,}5 & 0{,}5 & 0{,}5 \\ 0{,}5 & 1 & 0{,}5 & 0{,}5 & 0{,}5 \\ 0{,}5 & 0{,}5 & 1 & 0{,}5 & 0{,}5 \\ 0{,}5 & 0{,}5 & 0{,}5 & 1 & 0{,}5 \\ 0{,}5 & 0{,}5 & 0{,}5 & 0{,}5 & 1 \end{pmatrix};\ \sigma_Y = \begin{pmatrix} 1{,}5 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2{,}5 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix}$$

AUC	AUC Mean	AUC Median	AUC SD	AUC 95 % C.I.
SLM	0.7096	0.7087	0.0187	0.6745-0.7490
Max-Min	0.9106	0.9112	0.0121	0.8855-0.9331

Universidad Zaragoza 1542

Simulaciones sobre otras distribuciones(II)

Simulaciones sobre otras distribuciones (1000 simulaciones)				
M_1, M_2, M_3, M_4, M_5				
($\mathbf{X}$: Non Organ-confined, $\mathbf{Y}$: Organ Confined) with:				
$\mu_{X_i} = N(1,1)^{-3}$; $\mu_{Y_i} = N(0,1)$; $n_X = 250$, $n_Y = 250$				
AUC	AUC Mean	AUC Median	AUC SD	AUC 95 % C.I.
SLM	0.7115	0.7115	0.0020	0.6754-0.7530
Max-Min	0.9337	0.9345	0.0126	0.9085-0.9566

Universidad Zaragoza 1542

Simulaciones sobre otras distribuciones(III)

Simulaciones sobre otras distribuciones (1000 simulaciones)

M_1, M_2, M_3, M_4, M_5 under multivariate normality

($\mathbf{X}$: Non Organ-confined, $\mathbf{Y}$: Organ Confined) with:

$\mathbf{X} = (0{,}5, 0{,}5, 0{,}5, 0{,}5, 0{,}5)$; $\mathbf{Y} = (0, 0, 0, 0, 0)$; $n_X = 250$, $n_Y = 250$

$$\sigma_X = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}; \sigma_Y = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

AUC	AUC Mean	AUC Median	AUC SD	AUC 95 % C.I.
SLM	0.7883	0.7890	0.0199	0.7487-0.8266
Max-Min	0.7569	0.7567	0.0210	0.7165-0.7963

$$AUC_{max} = 0{,}7854$$

Universidad Zaragoza 1542

Conclusiones

- Cuando la distribución de los biomarcadores no esta claramente definida, los procedimientos no paramétricos son una buena alternativa para alcanzar modelos predictivos con una capacidad de discriminación alta.
- El algoritmo paso a paso que utiliza la función `SLM`, y la combinación Max-min que puede obtenerse con la función `MMLM` son modelos que proporcionan buenos modelos predictivos en el campo de la medicina.
- La combinación Max-min funciona mejor que otros modelos cuando la verdadera relación entre las variables no es del tipo lineal.

Universidad Zaragoza 1542

Bibliografía

- Su, J.Q. and Liu, J.S., 1993. Linear combinations of multiple diagnostic markers. J. Amer. Statist. Assoc., 88, 1350-1355.
- Pepe, M.S., Cai, T. and Longton, G., 2006. Combining Predictors for Classification Using The Area under the Receiver Operating Characteristic Curve. Biometrics, 62, 221-229.
- Esteban LM, Sanz G, Borque A, 2011. A step-by-step algorithm for combining diagnostic tests. Journal of Applied Statistics, 38(5), 899-911.
- Liu C, Liu A, Halabi S, 2011. A min-max combination of biomarkers to improve diagnostic accuracy. Stat Med. 30(16), 2005-2014.

Universidad Zaragoza 1542