Modelos para el aprendizaje automático

Silvia N. Pérez

Regresión logística múltiple

Bibliografía:

- Chaterjee, S.; Hadi, A.; Price, B. "Regression Analysis by Example". Wiley (Introductoria)
- Montgomery, Peck y Vining. "Introducción al Análisis de Regresión Lineal" (Introductoria)
- Hosmer, D.; Lemeshow, S.(2000). "Applied logistic regression" (Wiley Series in Probability)

Resumen

- Regresión logística múltiple
- Interpretación de los coeficientes
- Estimación máximo verosímil
- Inferencia en el modelo RL: significación de variables, IC, comparación de modelos.
- Medidas de ajuste: seudoR^2, test de Hosmer-Lemeshow
- Multicolinealidad
- Métodos de selección de variables
- Medidas Diagnósticas en RL
- RL para clasificación

Modelo de Regresión Logística múltiple

Para el caso de *p* predictores admite estas expresiones equivalentes:

- $$\pi(\mathbf{x}) = P(Y = 1/x_1, .. x_p) = \frac{e^{\beta_0 + \beta_1 x_1 + .. \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + .. \beta_p x_p}}$$

- $$\text{logit}(\mathbf{x}) = \log\left[\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}\right] = \beta_0 + \beta_1 x_1 + .. + \beta_p x_p$$

- $$\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} = e^{\beta_0 + \beta_1 x_1 + .. + \beta_p x_p}$$

Multicolinealidad en Reg Logística

- Produce efectos similares que en RLM: grandes desvíos para los estimadores de coeficientes.

- Los test de significación de coeficientes pueden dar "falsos aceptos", esto es, indicar que el coeficiente no es significativo cuando sí lo es.

Selección automática de variables

Están definidos los mismos métodos automáticos de selección de variables que se utilizaron en RLM:

- Backward
- Forward
- Stepwise

Estos comparan paso a paso modelos a partir de AIC (lo más común).

Vamos a R

RLogistica para clasificación

El modelo logístico **predice la probabilidad** de que se presente el evento Y=1 en cada nivel de covariables.

¿Cómo lo usamos para predecir la ocurrencia del evento?

La manera más común de clasificar observaciones usando RL es considerar que si la probabilidad pronosticada para un caso es mayor a 0.5 entonces el caso pertenece a la clase de interés (Y=1). Pero…. puede que esto no sea lo más adecuado!

Los resultados de la clasificación los vemos en una tabla 2x2 llamada "**tabla de clasificación o matriz de confusión**".

Matriz de confusión o Tabla de clasificación

		Real	
		1 (positivo)	0 (negativo)
Predicho	1 (positivo)	**VP**	**FP**
	0 (negativo)	**FN**	**VN**

Recall o **sensitividad** $= \frac{VP}{VP + FN} = \frac{VP}{total +}$

Especificidad $= \frac{VN}{FP + VN} = \frac{VN}{total -}$

Precisión $= \frac{VP}{VP + FP} = \frac{VP}{total\ clasif +}$

$$\mathrm{F1} = 2 * \frac{Precis * Sens}{Precis + Sens}$$

Accuracy del modelo

Para valorar la exactitud del modelo podemos definir:

$$\textbf{Accuracy} = \frac{\#decisiones\ correctas}{\#total\ de\ casos}$$

$$\textbf{Accuracy} = \ \mathrm{S}\ .\ \frac{total(+)}{total.casos} + \mathrm{E}.\ \frac{total(-)}{total.casos}$$

Esto es, accuracy es un promedio ponderado entre S y E.

Equivalentemente se puede ver

PMC = prob de mal clasificados =1-Exactitud

Ejemplo

		Real o Referencia		
		1 (positivo)	0 (negativo)	Totales
Predicho	1 (positivo)	37	16	**53**
	0 (negativo)	35	184	**219**
	Totales	**72**	**200**	**272**

Accu=(184+37)/272

Sens=37 / 72

Esp=184 / 200

Precisión= 37/53

$$\text{Accuracy} = (37/72)*(72/272)+(184/200)*(200/272) = 0.8125$$

RL para clasificación

La tabla de clasificación depende fuertemente de las probabilidades predichas. S y E dependen de la "mezcla" de casos, no sólo de la superioridad del modelo.

Por ejemplo: con p_corte=0.5

Si son 100 casos con probabilidad predicha de 0.51 → asigno Y=1

Pero dado que # positivos en 100 es una variable Bi(100, p), se estima que habrá 51 casos positivos y 49 negativos!!! Muchos mal clasificados.

Esto es, si hay muchos sujetos con probabilidad estimada cerca de p_corte, se esperan muchos mal clasificados.

Ejemplo: ¿cuál clasificador es mejor?

en un conjunto de test de 100.000 instancias, se tienen 3 clasificadores:

REAL

PRED	+	-
+	300	500
-	200	99000

S = 300 / 500 -> 60%

E= 99000 / 99500 -> 99,5%

Acc= 99300 / 100000 -> 99,3%

Error=PMC -> 0,7%

REAL

	+	-
+	0	0
-	500	99500

S = 0 / 500 -> 0%

E= 99500 / 99500 -> 100%

Acc= 99500/ 100000 -> 99,5%

Error=PMC -> 0,5%

REAL

	+	-
+	400	5400
-	100	94100

S = 400 / 500 -> 80%

E = 94100 / 99500 -> 94,6%

Acc= 94500 / 100000 -> 94,5%

Error=PMC -> 5,5%

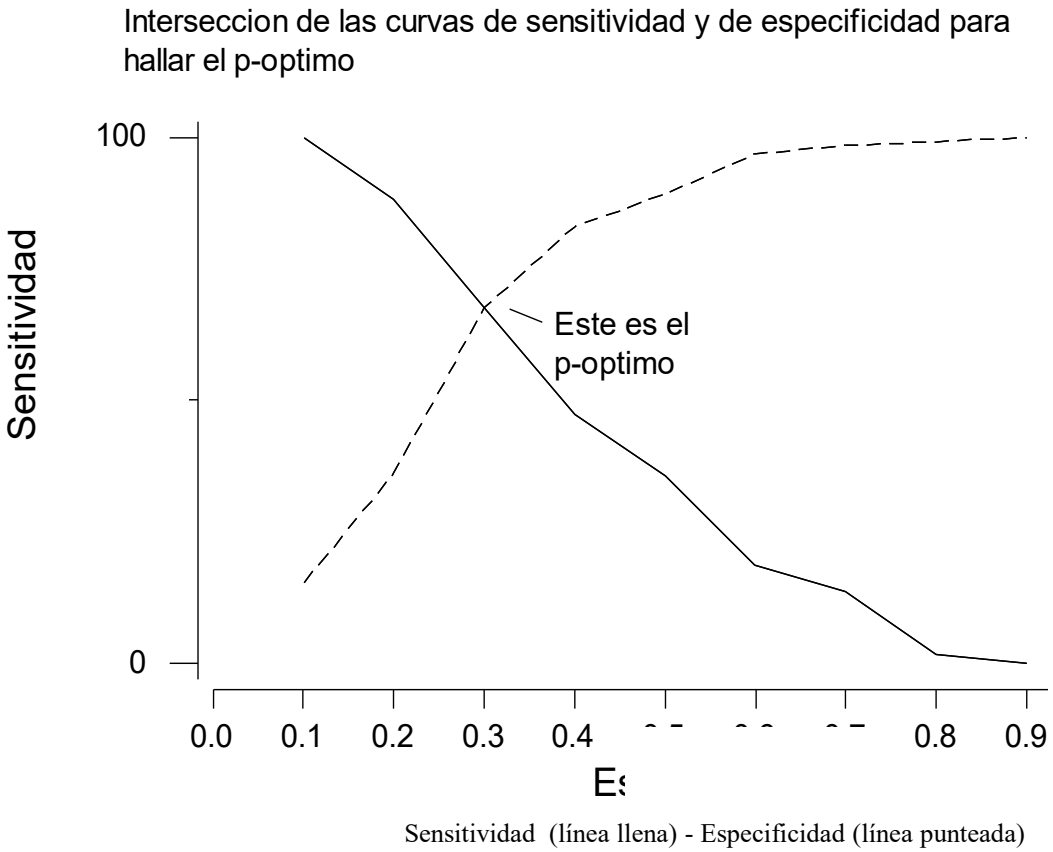
RL para clasificación

Un método para elegir el punto de corte puede ser graficar las curvas de Sensitividad y Especifidad versus los mismos niveles de probabilidad en una misma gráfica. La probabilidad que se usará para clasificar las observaciones se obtiene intersecando las dos curvas.

En realidad, muchas veces se quiere mejor Sensitividad (detecto mejor el %VP entre los P encontrados) que Especificidad (%VN entre los N encontrados).

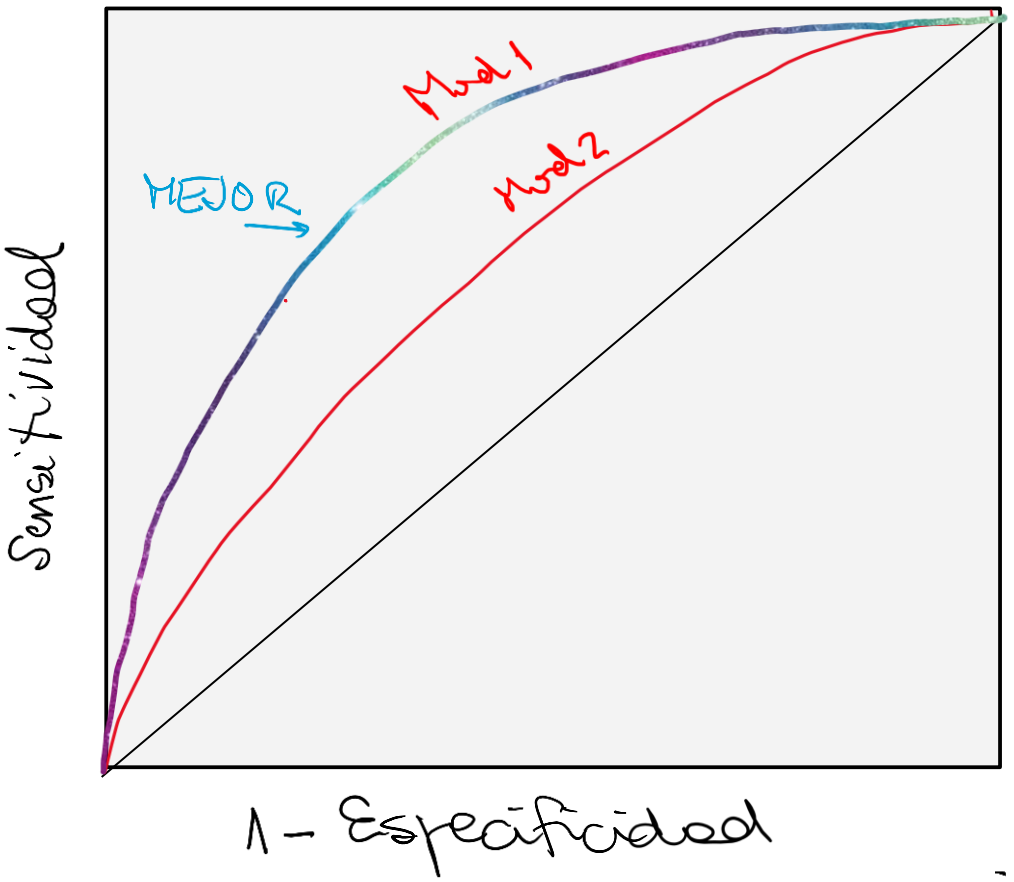
Lo más frecuente: usar la curva ROC (receiver operating characteristic) para elegir modelo. Hay métodos para encontrar p_corte a partir de esta.

Ejemplo



Esto muestra que un punto de corte aconsejable es p = 0.3

Curva Roc



El área bajo la curva ROC mide la **capacidad predictiva** del modelo.

Como armamos una curva ROC?

eje X : tasa de falsos positivos (FP/N)

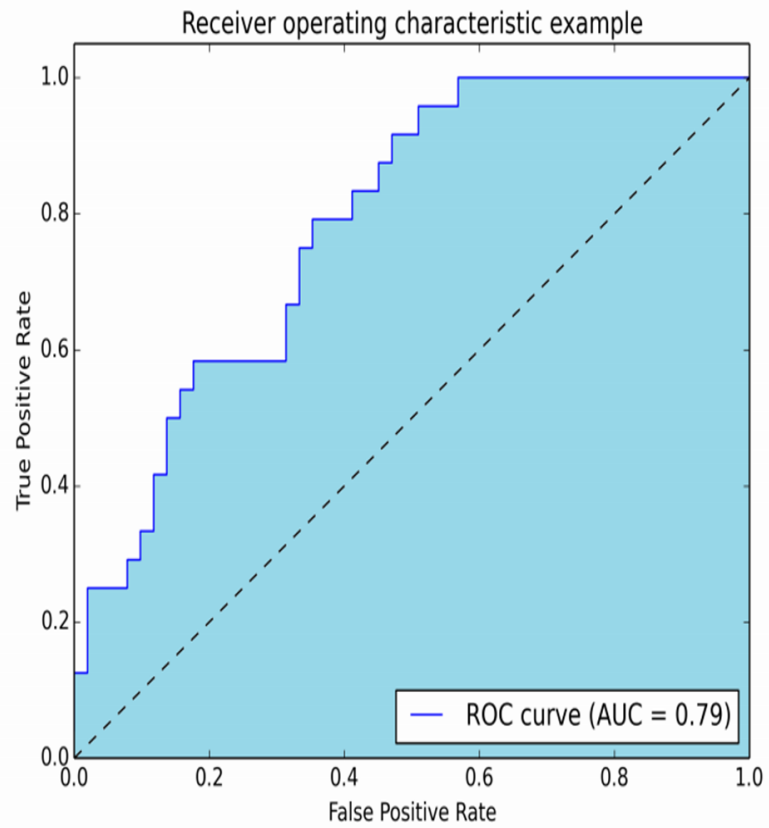
eje Y: tasa de verdaderos positivos (VP/P)

Los casos se ordenan en forma decreciente por su probabilidad de pertenencia a la clase P

La “curva” es una composición secuencial de segmentos horizontales (de izquierda a derecha y verticales de abajo a arriba)

Si el próximo caso es positivo la curva aumenta en el eje Y en una proporción de 1/P

Si el próximo caso es negativo (falso positivo) la curva se desplaza a derecha en una proporción de 1/N



	Pred +	Pred-
+	VP	FN
-	FP	VN

Ejemplo bobo

P=# positivos reales (Y=1) = 4
N=# negativos reales (Y=0) = 6

10 casos

Ajusto una regresión logística y obtuve

1º) Prob. predichas orden descendente

Punto_i= (FP_i/N; VP_i/P)

Y_i	$\hat{\pi}_i$
1	0,9
1	0,8
0	0,75
1	0,6
0	0,55
1	0,5
0	0,4
0	0,3
0	0,15
0	0,1

Punto 1: corto en 1º punto arriba

$P_1 = \left(\frac{0}{6}; \frac{1}{4}\right) = (0, \frac{1}{4})$

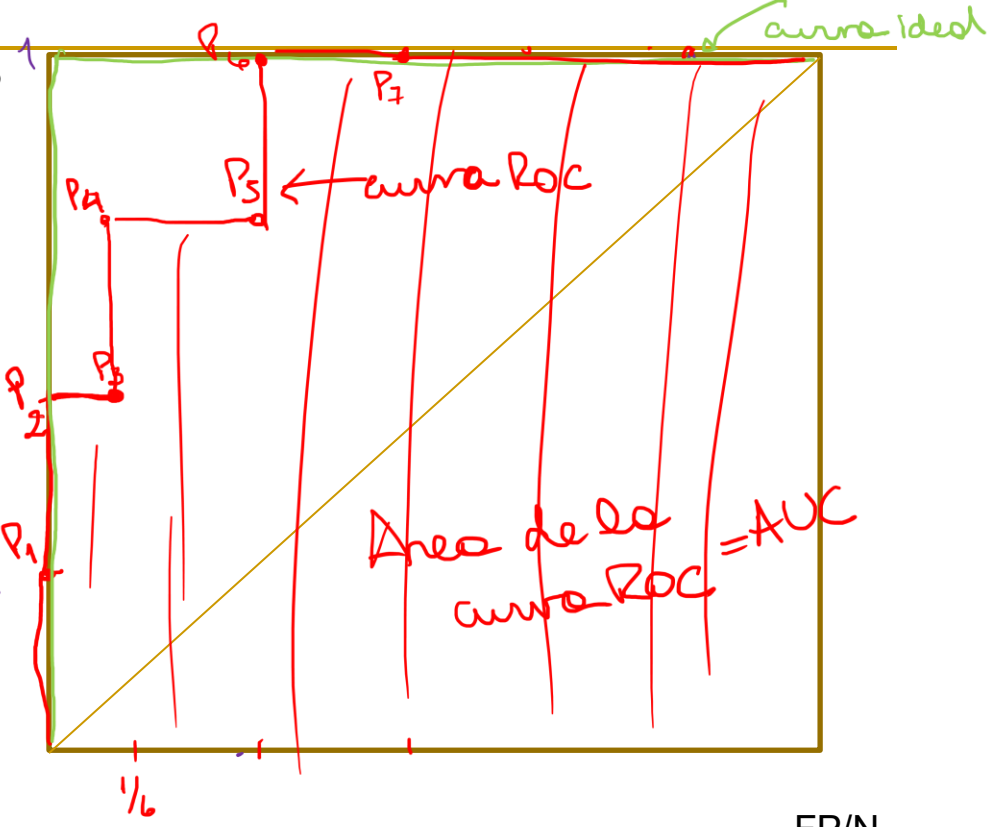
$P_2 = \left(\frac{0}{6}; \frac{2}{4}\right) = (0, \frac{1}{2})$

$P_3 = \left(\frac{1}{6}; \frac{2}{4}\right) = (\frac{1}{6}, \frac{1}{2})$

$P_4 = \left(\frac{1}{6}; \frac{3}{4}\right) = (\frac{1}{6}, \frac{3}{4})$

$P_5 = \left(\frac{2}{6}, \frac{3}{4}\right) = (\frac{1}{3}, \frac{3}{4})$ $P_7 = \left(\frac{3}{6}, \frac{4}{4}\right) = (\frac{1}{2}, 1)$

$P_6 = \left(\frac{2}{6}, \frac{4}{4}\right) = (\frac{1}{3}, 1)$



AUC ideal = 1

Modelo de alta capacidad predictiva tiene AUC → 1

Vamos a R

Ejemplo “hipoteca25”

Se quiere modelar la situación crediticia de un cliente a partir de ciertas variables relevadas.

1. Partir el conjunto de datos en train/test.
2. En el conjunto de train:
   - Proponer un modelo con todas las variables,
   - seleccionar variables, proponer nuevos modelos
3. Comparar modelos y elegir algunos para evaluar en test:
   - Matriz de confusión,
   - Sensitividad, especificidad, precisión.
   - Área bajo la curva ROC.

Para implementar un análisis de RL

1. ¿El tamaño de muestra es adecuado? (mín de 10 casos por variable en el nivel de "*Y*" con menos casos).

2.Transformar/limpiar datos.

3. Partir el conjunto de datos en train/test.

4. En el conjunto de train:

- Proponer un modelo con todas las variables,
- analizar multicolinealidad
- ver influyentes, H-L, R2, etc
- seleccionar variables, proponer nuevos modelos

5. Comparar modelos y elegir algunos para evaluar en test

6. Evaluar modelos en el conjunto de test (Matriz de confusión, sensitividad, especificidad. Área bajo la curva ROC).

7. Elegir modelo, interpretar coeficientes, ¡predecir nuevos datos!