Instituto Cántabro de ESTADÍSTICA

Estadística con R

Métodos de Clasificación

Métodos de Clasificación

- La clasificación supervisada es una de las tares que más frecuentemente son llevadas a cabo por los denominados Sistemas Inteligentes.
- Técnicas Estadísticas: Regresión Logística, Análisis Discriminante, Análisis de conglomerados ó Cluster.
- Inteligencia Artificial: Redes Neuronales, K vecinos próximos, los Árboles de Decisión, las Máquinas Soporte Vector y Redes Bayesianas.

Minería de datos aplicada a las técnicas de clasificación

- Partición del conjunto de datos en dos subconjuntos que serán utilizados
  - Entrenamiento (utilizado para estimar los parámetros del modelo)
  - Test (comprobar el comportamiento del modelo estimado)
- Dividir el conjunto de datos en ambos subconjuntos por un procedimiento de muestreo (muestreo aleatorio simple)
- Aplicar una métrica de evaluación

Métrica de evaluación

- Recuento de clasificación binaria:

		Valor real de Y_i	
		$Y_i = 0$	$Y_i = 1$
$\hat{Y}_i$	$\hat{Y}_i = 0$	P_{11}	P_{12}
	$\hat{Y}_i = 1$	P_{21}	P_{22}

- P_{11} y P_{22} corresponderán a predicciones correctas (valores 0 bien predichos en el primer caso y valores 1 bien predichos en el segundo caso), mientras que P_{12} y P_{22} corresponderán a predicciones erróneas (valores 1 mal predichos en el primer caso y valores 0 mal predichos en el segundo caso) .

Instituto Cántabro de ESTADÍSTICA

Bondad de Ajuste

Índices para medir la bondad del ajuste

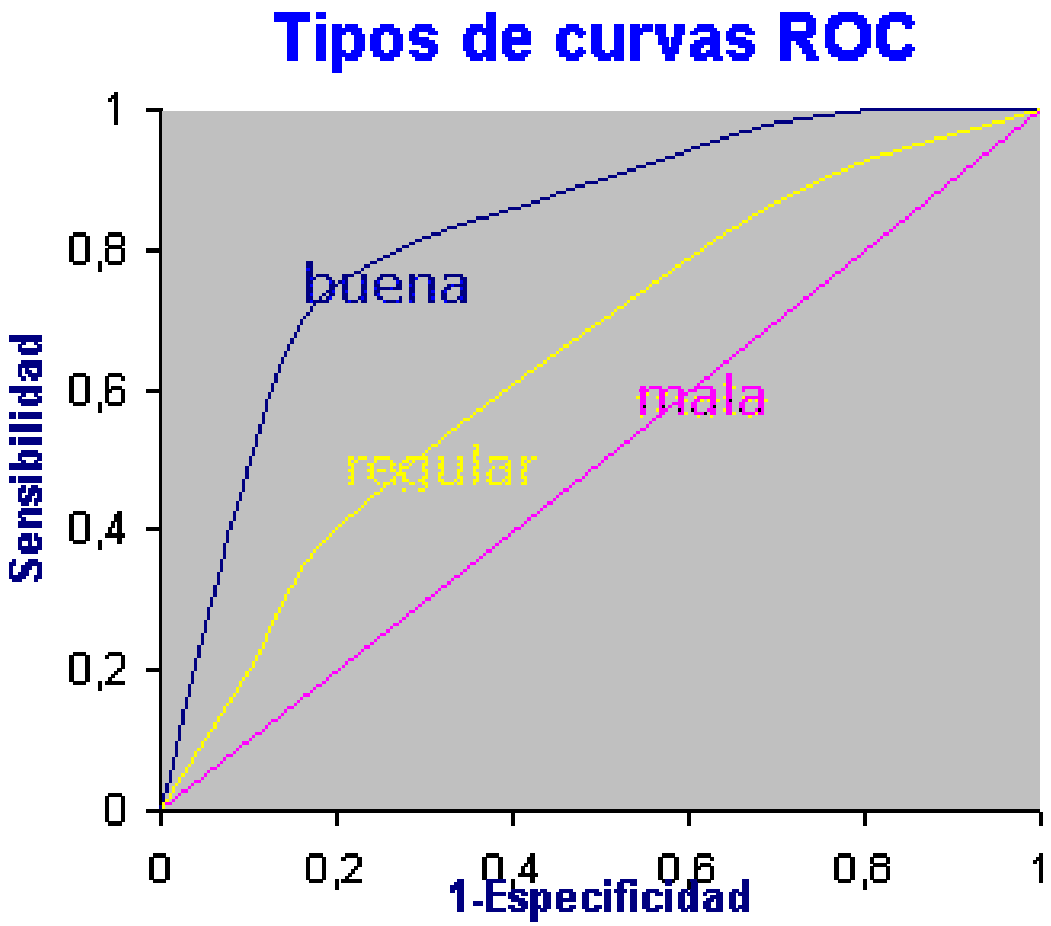
Indice	Definición	Expresión
Tasa de aciertos	Cociente entre las predicciones correctas y el total de predicciones	$\frac{P_{11}+P_{22}}{P_{11}+P_{12}+P_{21}+P_{22}}$
Tasa de errores	Cociente entre las predicciones incorrectas y el total de predicciones	$\frac{P_{12}+P_{21}}{P_{11}+P_{12}+P_{21}+P_{22}}$
Especificidad	Proporción entre la frecuencia de valores 0 correctos y el total de valores 0 observados	$\frac{P_{11}}{P_{11}+P_{21}}$
Sensibilidad	Razón entre los valores 1 correctos y el total de valores 1 observados	$\frac{P_{22}}{P_{12}+P_{22}}$
Tasa de falsos ceros	Proporción entre la frecuencia de valores 0 incorrectos y el total de valores 0 observados	$\frac{P_{21}}{P_{11}+P_{21}}$
Tasa de falsos unos	Razón entre los valores 1 incorrectos y el total de valores 1 observados	$\frac{P_{12}}{P_{12}+P_{22}}$

Curva ROC (Receiver Operating Characteristic)

- Representación gráfica del rendimiento del clasificador: muestra la distribución de las fracciones de verdaderos positivos y de falsos positivos

- La fracción de verdaderos positivos se conoce como sensibilidad: probabilidad de clasificar correctamente a un individuo cuyo estado real sea definido como positivo.
- La especificidad es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea clasificado como negativo (restar uno de la fracción de falsos positivos)
- La curva ROC también es conocida como la representación de sensibilidad frente a (1-especificidad)

Curva ROC


Tipos de curvas ROC
Sensibilidad
1
0,8
0,6
0,4
0,2
0
buena
regular
mala
0
0,2
0,4
0,6
0,8
1
1-Especificidad