REGRESIONES LOGISTICAS

ING. JAIRO SALAZAR

REGRESION LOGISTICA

La regresión logística puede predecir con precisión un resultado binario.

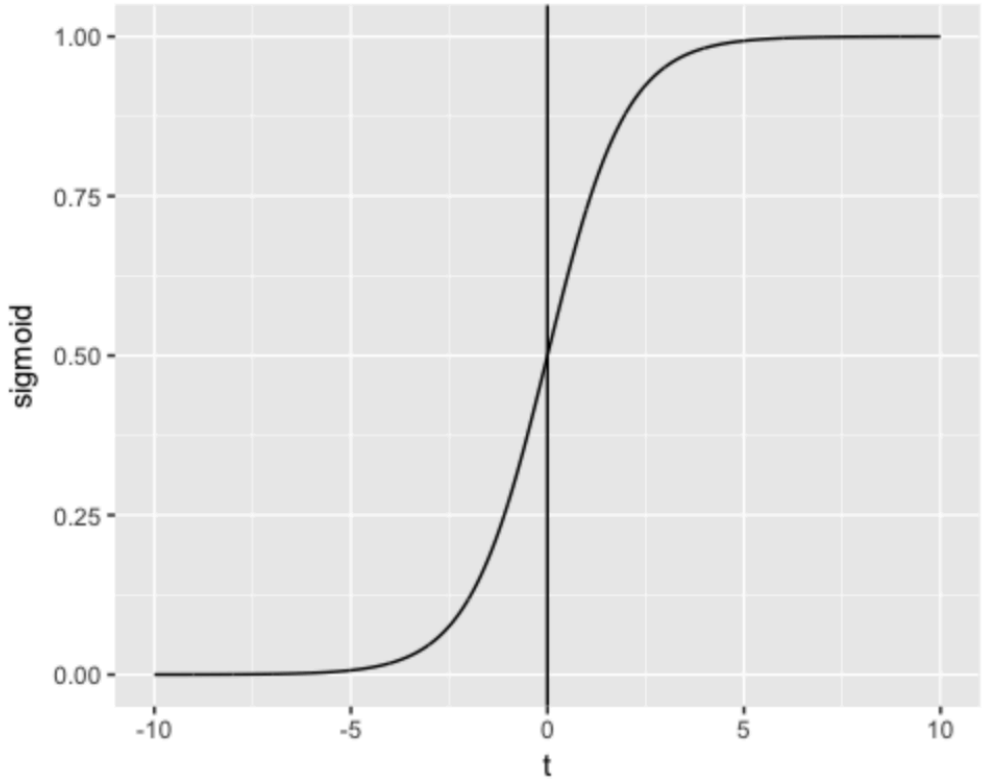
Imagine que desea predecir si un préstamo es denegado/aceptado en función de muchos atributos. La regresión logística es de la forma 0/1. y = 0 si se rechaza un préstamo, y = 1 si se acepta.

Un modelo de regresión logística difiere del modelo de regresión lineal de dos maneras.

- En primer lugar, la regresión logística solo acepta entradas binarias como variable dependiente (es decir, un vector de 0 y 1).
- En segundo lugar, el resultado se mide mediante la siguiente función de enlace probabilístico llamada **sigmoide** debido a su forma de S:

$$\sigma(t) = \frac{1}{1 + exp(-t)}$$

La salida de la función siempre está entre 0 y 1. Verifique la imagen a continuación



REGRESION LOGISTICA

La función sigmoid devuelve valores de 0 a 1. Para la tarea de clasificación, necesitamos una salida discreta de 0 o 1.

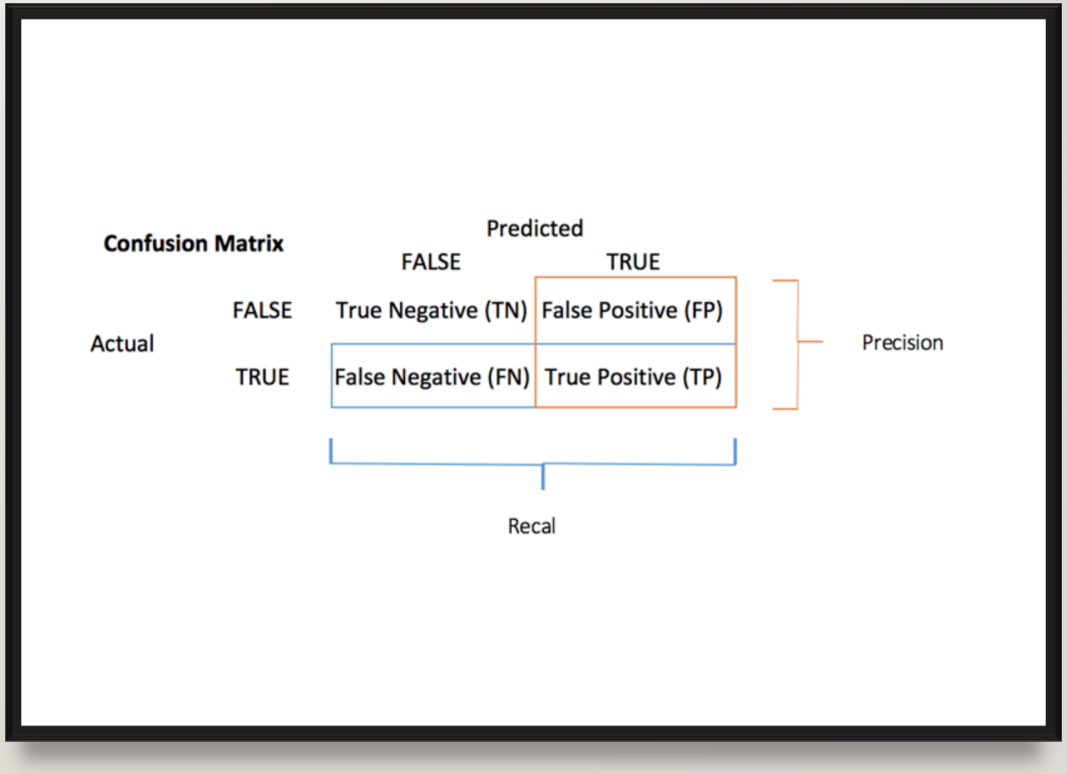
Para convertir un flujo continuo en un valor discreto, podemos establecer un límite de decisión en 0,5. Todos los valores por encima de este umbral se clasifican como 1

$$\hat{y} = \begin{cases} 0 & if \quad \hat{p} < .5 \\ 1 & if \quad \hat{p} \geq .5 \end{cases}$$

EVALUAR EL DESEMPEÑO

Matriz de confusión

La **matriz de confusión** es una mejor opción para evaluar el rendimiento de la clasificación. La idea general es contar el número de veces que las instancias verdaderas se clasifican como falsas.



$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

ACCURACY

Puede calcular la **precisión** del modelo sumando el verdadero positivo + el verdadero negativo sobre la observación total

PROBLEMAS CON EL ACCURACY

A veces puede suceder la **paradoja de la prueba de Accuracy** . Dijimos que la Accuracy es la relación entre las predicciones correctas y el número total de casos. Podemos tener una precisión relativamente alta pero un modelo inútil. Ocurre cuando hay una clase dominante. Si mira hacia atrás en la matriz de confusión, puede ver que la mayoría de los casos se clasifican como verdaderos negativos. Imagínese ahora, el modelo clasificó todas las clases como negativas. Tendría un accuracy alto Su modelo funciona mejor, pero tiene dificultades para distinguir el verdadero positivo del verdadero negativo.

- En tal situación, es preferible tener una métrica más concisa. Podemos mirar:
- Precision=TP/(TP+FP)
- recall=TP/(TP+FN)

La precisión analiza la exactitud de la predicción positiva. **El recall** es la proporción de instancias positivas que el clasificador detecta correctamente.

F1 SCORE

Se puede crear una puntuación basada en la precisión y la recuperación. Esta es una media armónica de estas dos métricas, lo que significa que da más peso a los valores más bajos.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$