Umbral de Clasificación

Como vimos la regresión logística devuelve un una probabilidad estimada y podríamos dejarla como está o bien transformarla en un valor binario (0 ó 1, o bien P o N). ¿Con qué criterio?

Umbral de Clasificación

Como vimos la regresión logística devuelve un una probabilidad estimada y podríamos dejarla como está o bien transformarla en un valor binario (0 ó 1, o bien P o N). ¿Con qué criterio?

Pensemos en la situación con covariable edad: si obtuviéramos para una determinada edad una probabilidad estimada de 0.99, seguramente asignaríamos a CUI, mientras que si fuera de 0.005 no lo haríamos....

Umbral de Clasificación

Como vimos la regresión logística devuelve un una probabilidad estimada y podríamos dejarla como está o bien transformarla en un valor binario (0 ó 1, o bien P o N). ¿Con qué criterio?

Pensemos en la situación con covariable edad: si obtuviéramos para una determinada edad una probabilidad estimada de 0.99, seguramente asignaríamos a CUI, mientras que si fuera de 0.005 no lo haríamos....

...Y si para cierta edad la probabilidad estimada fuera 0.6 ¿a qué categoría la asignaríamos?

Umbral de Clasificación

Como vimos la regresión logística devuelve un una probabilidad estimada y podríamos dejarla como está o bien transformarla en un valor binario (0 ó 1, o bien P o N). ¿Con qué criterio?

Pensemos en la situación con covariable edad: si obtuviéramos para una determinada edad una probabilidad estimada de 0.99, seguramente asignaríamos a CUI, mientras que si fuera de 0.005 no lo haríamos....

...Y si para cierta edad la probabilidad estimada fuera 0.6 ¿a qué categoría la asignaríamos?

Para asignar debemos tener un **umbral (threshold) de clasificación o decisión** por encima del cual asignamos a Positivo o 1 y por debajo del cual asignamos a N o 0.

Umbral de Clasificación

Como vimos la regresión logística devuelve un una probabilidad estimada y podríamos dejarla como está o bien transformarla en un valor binario (0 ó 1, o bien P o N). ¿Con qué criterio?

Pensemos en la situación con covariable edad: si obtuviéramos para una determinada edad una probabilidad estimada de 0.99, seguramente asignaríamos a CUI, mientras que si fuera de 0.005 no lo haríamos....

...Y si para cierta edad la probabilidad estimada fuera 0.6 ¿a qué categoría la asignaríamos?

Para asignar debemos tener un **umbral (threshold) de clasificación o decisión** por encima del cual asignamos a Positivo o 1 y por debajo del cual asignamos a N o 0.

Muchas veces se supone a ese umbral como 0.5, pero este valor se debe ajustar en muchos problemas.

Verdadero Positivo y Falso Positivo

En un problema de aprendizaje supervisado, tenemos las etiquetas de cada individuo, por lo tanto podemos contrastar la asignación con la etiqueta.

De este contraste nos daremos cuenta de que existen asignaciones correctas y otras que no lo son. En base a esta información podemos computar distintas medidas para evaluar la performance de una clasificación.

Evaluación de la Clasificación: Métricas

- Matriz de Confusión
- Accuracy o Exactitud
- Sensibilidad
- Especificidad
- Curvas ROC

Matriz de confusión: Clasificación Binaria

Se trata de una tabla en la que se resume el rendimiento de un modelo de clasificación en los datos de prueba.
Su nombre se refiere a que permite identificar dónde el clasificador está confundiendo dos clases.

- **Verdaderos Positivos (VP)**: la clase real es Positiva (=1) y la predicha también es Positiva (=1)
- **Verdaderos Negativos (VN)**: la clase real es Negativa (=0) y la pronosticada también es Negativa (=0).
- **Falsos Positivos (FP)**: la clase real es Negativa (=0) y la pronosticada es Positiva (=1).
- **Falsos Negativos (FN)**: la clase real es Positiva (=1) y el valor predicho es Negativo (=0).

Matriz de Confusión

	Predicción	
Observación	+	-
• +	Verdaderos Positivos (VP)	Falsos Negativos (FN)
• −	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Exactitud o Accuracy

La exactitud o accuracy es la proporción de clasificaciones correctas.

$$\frac{VP + VN}{VP + VN + FP + VN}$$

Es decir, es el número de predicciones correctas sobre el número de predicciones totales.

Exactitud o Accuracy

La exactitud o accuracy es la proporción de clasificaciones correctas.

$$\frac{VP + VN}{VP + VN + FP + VN}$$

Es decir, es el número de predicciones correctas sobre el número de predicciones totales.

No caracteriza a un buen clasificador necesariamente. Supongamos que tenemos 10 enfermos y 90 sanos y el clasificador, asigna a cualquier individuo a la clase sano. Entonces, la Accuracy sería 0.9. No sirve si las clases están muy desbalanceadas.

Sensibilidad y Especificidad

- **Sensibilidad** (Recall)

$$\frac{VP}{VP + FN}$$

entre los positivos cuántos fueron predichos como positivos. ¿Qué proporción de positivos reales se identificó correctamente? Alta sensibilidad implica que no se deja a ningún positivo mal clasificado, pero posiblemente se clasificó a alguno de la otra clase. ojo!: el clasificador que simplemente predice a todos como positivos tiene alta sensibilidad.

Sensibilidad y Especificidad

- **Sensibilidad** (Recall)

$$\frac{VP}{VP + FN}$$

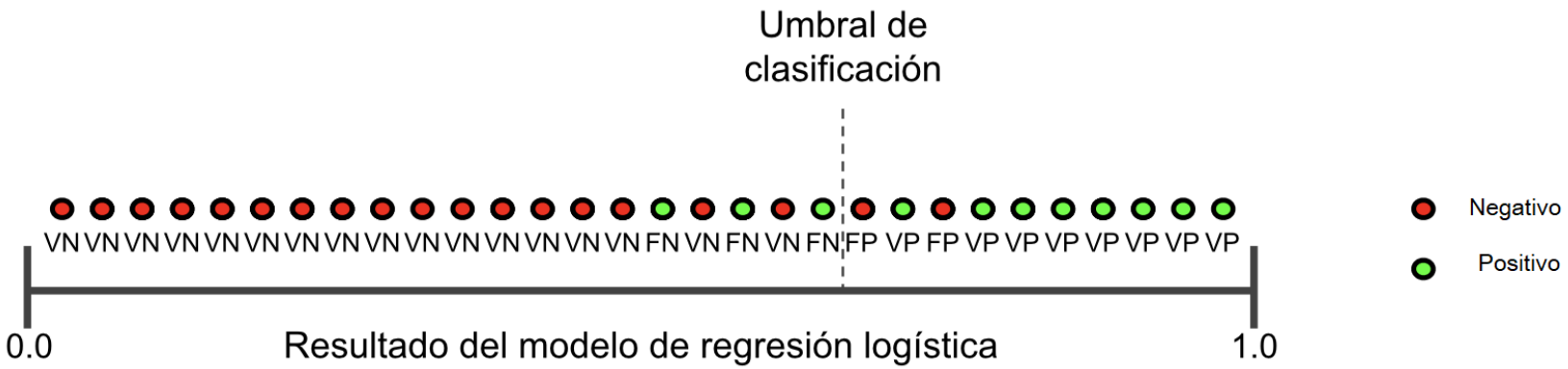
  entre los positivos cuántos fueron predichos como positivos. ¿Qué proporción de positivos reales se identificó correctamente? Alta sensibilidad implica que no se deja a ningún positivo mal clasificado, pero posiblemente se clasificó a alguno de la otra clase. ojo!: el clasificador que simplemente predice a todos como positivos tiene alta sensibilidad.

- **Especificidad**

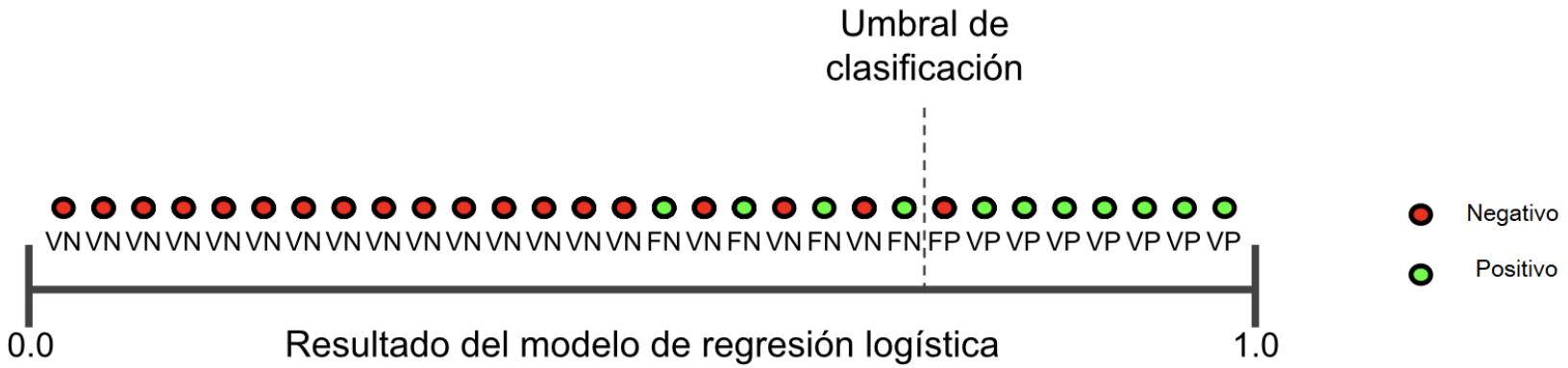
$$\frac{VN}{VN + FP}$$

  entre los negativos cuántos fueron predichos negativos. Es la capacidad de detectar negativos.

Especificidad vs. Sensibilidad



Especificidad vs. Sensibilidad



Curvas ROC

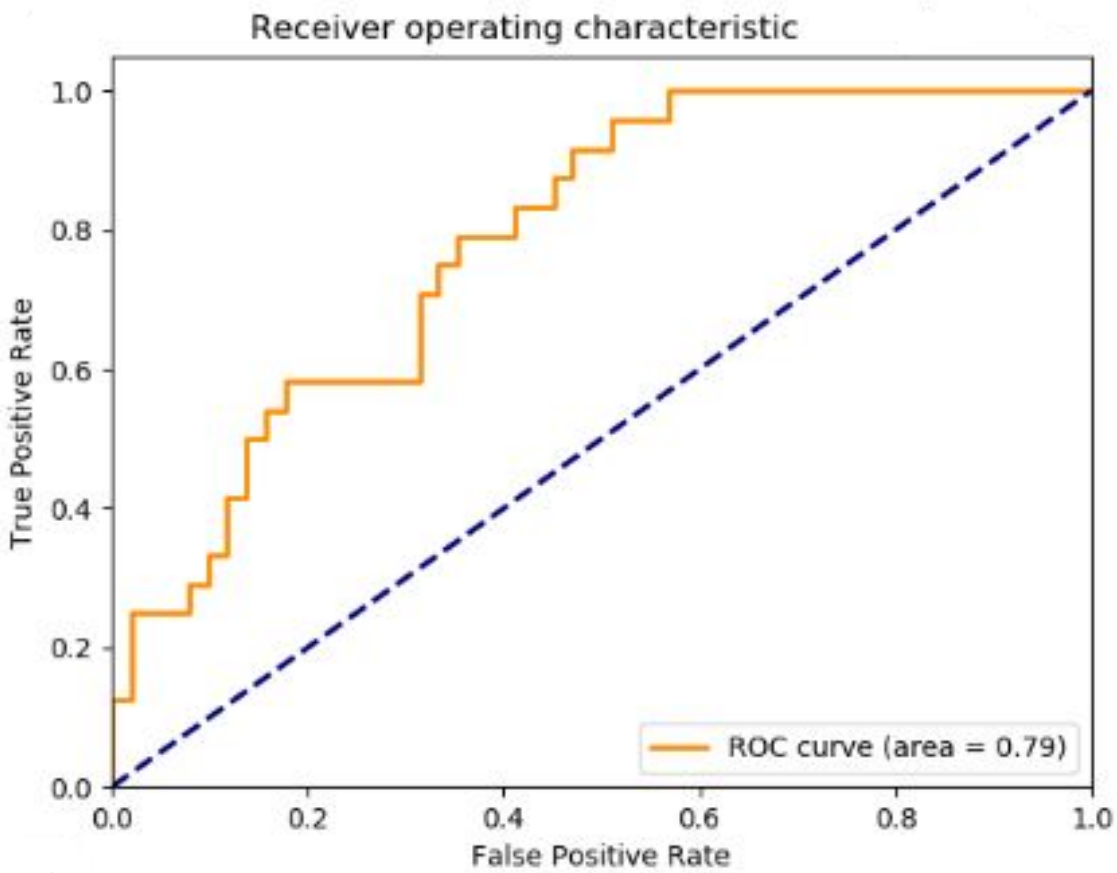
(receiver operating characteristics)

Es un gráfico que muestra el rendimiento de un modelo de clasificación a medida que se varían los umbrales de clasificación.

Estas curvas surgieron en la segunda guerra mundial, fueron desarrolladas por ingenieros militares para el análisis de señales en un radar y la capacidad del operador para determinar la naturaleza del objeto detectado.

Se usa tanto la curva ROC como el área que queda debajo de ella, llamada AUC.

Curva ROC


Receiver operating characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.79)

Curvas ROC

Tasa de Verdaderos Positivos (TPR) $= \frac{VP}{VP + FN}$ =Sensibilidad

(TPR: True Positive Ratio)

Curvas ROC

Tasa de Verdaderos Positivos (TPR) $= \frac{VP}{VP+FN}$ =Sensibilidad

(TPR: True Positive Ratio)

Tasa de Falsos Positivos (FPR) $= \frac{FP}{FP+VN}$ = 1- Especificidad

(FPR: False Positive Ratio)

Curvas ROC

Tasa de Verdaderos Positivos (TPR) $= \frac{VP}{VP+FN}$ =Sensibilidad

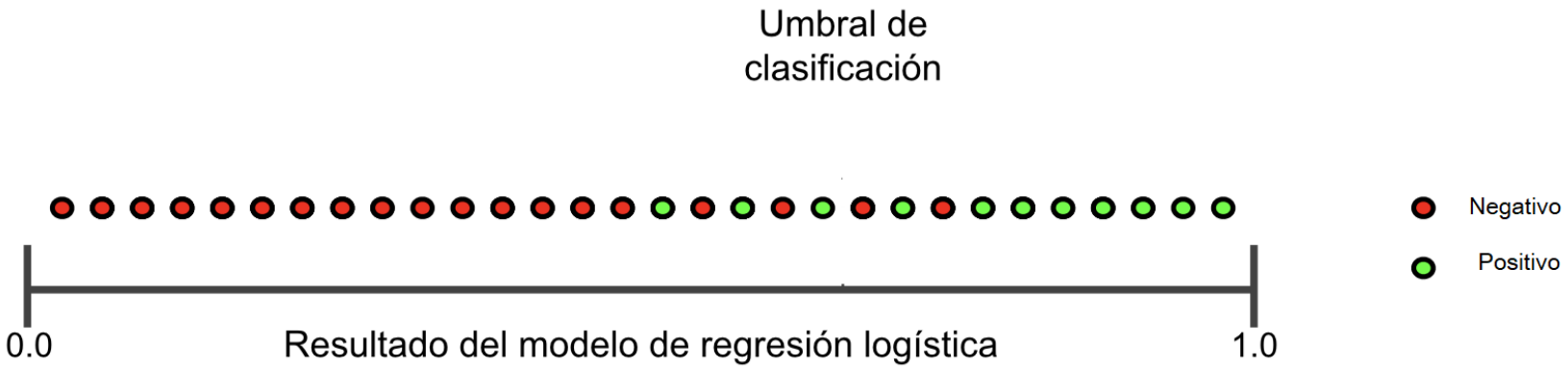
(TPR: True Positive Ratio)

Tasa de Falsos Positivos (FPR) $= \frac{FP}{FP+VN}$ = 1- Especificidad

(FPR: False Positive Ratio)

Una curva ROC representa TPR vs. FPR en diferentes umbrales de clasificación.

Especificidad vs. Sensibilidad



Curvas ROC

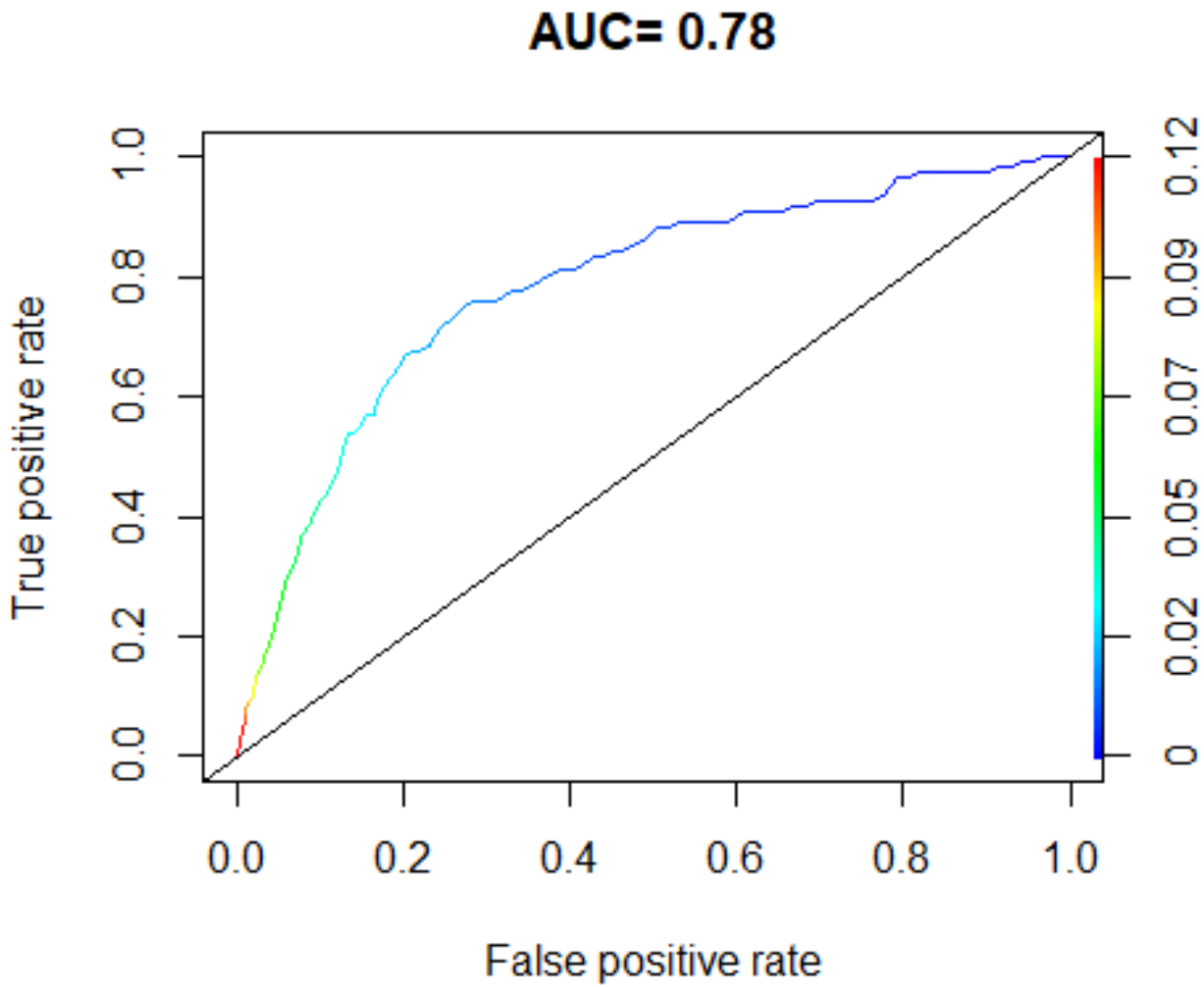
¿Con qué umbral nos quedamos?
Depende cada problema.

Se suele elegir el Indice de Youden. Gráficamente es el el punto de la curva ROC más cercano al ángulo superior-izquierdo del gráfico (0,1), corresponde al máximo índice de Youden que es (sensibilidad + especificidad - 1).

ROC: usando Edad

```
#fije semilla
dima<-dim(datos)
grupo.train<- sample(c(0,1), dima[1], replace = TRUE,prob=c(0.10,0.90))
indices<- 1:dima[1]
indices.train<- (indices[grupo.train==1])

datos.train<- datos[indices.train,]
datos.test<- datos[-indices.train,]

mean(datos.train$cui)
mean(datos.test$cui)
mean(datos$cui)

salida.train<- glm(cui~edad.dec+genero,data=datos.train,family=binomial)
predicciones<-predict.glm(salida.train,datos.test,type="response")

library(ROCR)
predict.rocr<- prediction(predicciones,datos.val$cui)
pref.rocr<- performance(predict.rocr,"tpr","fpr")

auc.cui<- as.numeric(performance(predict.rocr,"auc")@y.values)
auc.cui

plot(pref.rocr,colorize=TRUE,type="l",main=paste("AUC=",round(auc.cui,2)))
abline(a=0,b=1)
```

Curva ROC


AUC= 0.78
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0
0.02
0.05
0.07
0.09
0.12