ntic master School

UNIVERSIDAD COMPLUTENSE DE MADRID

Regresión logística

Minería de Datos y Modelización Predictiva

Máster en Big Data, Data Science & Inteligencia Artificial
Universidad Complutense de Madrid

Recordemos que los modelos de regresión tienen por objetivo **predecir** una variable y (que recibe el nombre de dependiente) a partir de un conjunto de m variables x_i *independientes* a través de una **ecuación**:

$$y = f(x_1, x_2, \ldots, x_m) + \epsilon.$$

En el tema anterior tratamos el caso de la **regresión lineal**, donde la variable objetivo y es **continua** y viene dada por:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \epsilon.$$

En este tema vamos a trabajar con variables **objetivo de clase**, es decir, variables que toman un número finito de valores. Dado que estas variables no toman valores numéricos, **no es posible** plantear un modelo como el de regresión lineal.

Una alternativa sería **asignar un valor numérico a cada categoría** y aplicar un modelo de regresión lineal: por ejemplo, si se desea estimar qué producto financiero van a contratar los clientes de una entidad financiera, se puede asumir: 1=hipoteca, 2=cuenta corriente, 3=depósito, etc.

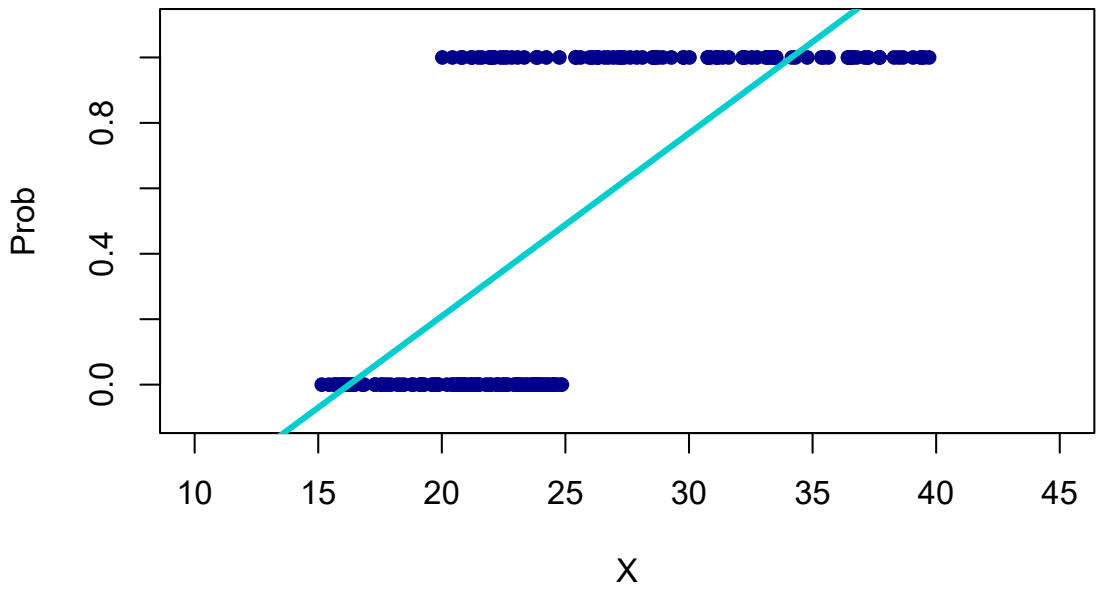
Sin embargo, eso implica asumir que la **distancia** entre todas las categorías es la misma y que existe cierto **orden** entre ellas, lo que generalmente no tiene sentido.

Si la variable a predecir es **binaria**, entonces se podría plantear construir un modelo de regresión lineal cuya **variable objetivo sea una variable *dummy*** obtenida a partir de la variable original. Por ejemplo, si se desea estudiar si ciertos clientes van a adquirir un producto o no: 1=sí y 0=no.

Se puede interpretar que, en ese caso, los valores de predicción obtenidos coinciden con la **probabilidad de que la variable tome el valor** 1. Por tanto, las variables regresoras aportan información sobre las probabilidades.

Esto nos da algunas pistas sobre cómo proceder para predecir variables objetivo de clase: ¿por qué no **predecir la probabilidad** de que la variable tome cada uno de los valores de la variable objetivo? ¿Por qué no relacionar las variables regresoras con dichas probabilidades?

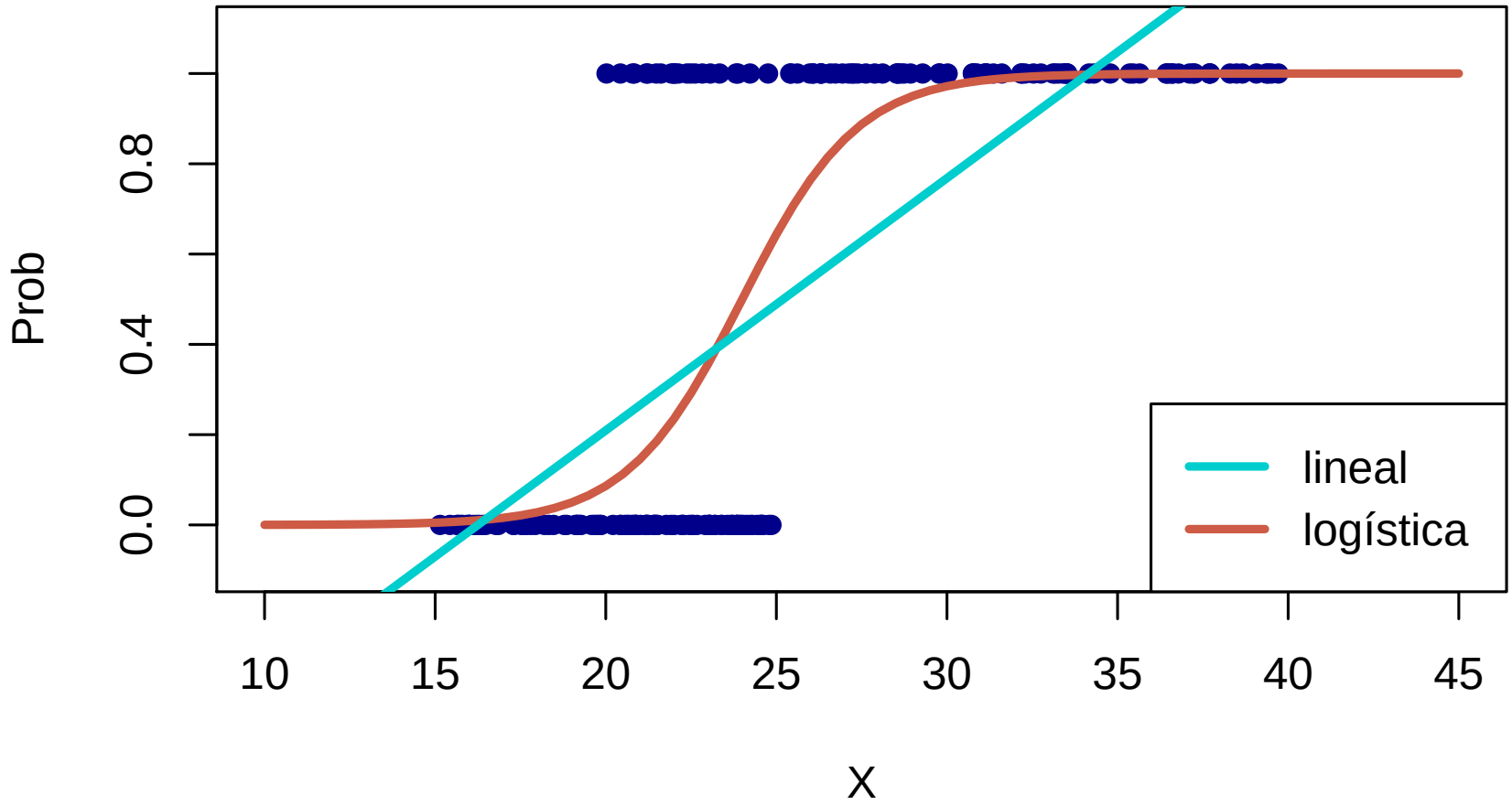
Sin embargo, el planteamiento anterior tiene un **problema**: con el modelo de regresión lineal se pueden obtener valores de predicción **fuera del intervalo** $(0, 1)$, lo que es incompatible con el concepto de probabilidad.



Lo ideal en este caso sería, entonces, **encontrar una función** que nos permita relacionar las variables regresoras con las probabilidades y que asegurara que éstas se encuentran en el **intervalo** $(0, 1)$. Este tipo de funciones reciben el nombre de **funciones de enlace** (o link).

Las **funciones de distribución** son un claro ejemplo de funciones de enlace. Las más utilizadas son: distribución logística, distribución normal y distribución gumbel.

Debido a la mayor facilidad de **interpretación** de los parámetros, la función **logística** es la más frecuente y el objetivo de este tema.


Prob
0.0
0.4
0.8
10
15
20
25
30
35
40
45
X
lineal
logística

REGRESIÓN LOGÍSTICA

El modelo de regresión logística asume la siguiente **relación**:

$$p_1 = P(Y = 1|x_1, x_2, \ldots, x_m) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m)}},$$

lo que implica que:

$$p_0 = 1 - p_1 = P(Y = 0|x_1, x_2, \ldots, x_m) = \frac{e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m)}}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m)}}$$

y

$$\log\left(\frac{p_1}{1 - p_1}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m.$$

El término de la izquierda recibe el nombre de *logit* y es el logaritmo de la **razón de probabilidades** u *odds* (que explicaremos a continuación).

Nótese que si un determinado parámetro es positivo (o negativo), aumentar una unidad la variable correspondiente se traduce en un **aumento** (o disminución) **de** p_1 y, correspondientemente, del **logit** (en este último en β_i unidades).

Además, se puede comprobar que usando esta función de enlace **las probabilidades resultantes están restringidas** al intervalo $(0, 1)$.

Para poder **interpretar correctamente** los parámetros del modelo de regresión logística es necesario definir dos **conceptos nuevos**: Odds y Odds-Ratio.

ODDS

Los *odds* asociados a un determinado suceso A se definen como el **cociente entre la probabilidad** de que ocurra el suceso y la probabilidad de que no ocurra:

$$odds(A) = \frac{P(A)}{1 - P(A)} \Leftrightarrow P(A) = \frac{odds(A)}{odds(A) + 1}$$

Obsérvese que si **aumenta la probabilidad** del suceso, también **aumenta el odds** de dicho suceso. Además, conociendo uno de los valores, se puede obtener el otro por lo que ambos miden, en distinta escala, cómo de probable es un suceso.

Nótese que el logit definido anteriormente no es más que el **logaritmo de los odds del suceso de interés** $(Y = 1)$.

ODDS-RATIO

Asociado al concepto de Odds, se puede definir el Odds-ratio como el **cociente entre los odds** de un suceso bajo una determinada condición y el odds de ese mismo suceso bajo otra condición, lo que permitirá evaluar el **efecto de dichas condiciones** sobre las probabilidades del suceso.

INTERPRETACIÓN DE LOS PARÁMETROS: VARIABLE REGRESORA DICOTÓMICA

Para interpretar correctamente el parámetro de una variable regresora dicotómica debemos tener en cuenta que la **inclusión de variables cualitativas** se lleva a cabo de la misma forma en regresión logística y lineal y, por tanto, se basa en la **creación de variables *dummy***.

Supongamos un modelo univariante donde la única variable regresora sea **dicotómica**:

$$\log\left(\frac{p_1}{1-p_1}\right) = \beta_0 + \beta_1(x=1) \Leftrightarrow p_1 = \frac{1}{1+e^{-(\beta_0+\beta_1(x=1))}}$$

Calculemos los **odds** de aquellos individuos para los que $x=1$ y para los que $x=0$:

$$odds(evento|x=1) = \frac{P(evento|x=1)}{1-P(evento|x=1)} = \frac{\frac{1}{1+e^{-(\beta_0+\beta_1)}}}{\frac{e^{-(\beta_0+\beta_1)}}{1+e^{-(\beta_0+\beta_1)}}} = e^{\beta_0+\beta_1}$$

$$odds(evento|x=0) = \frac{P(evento|x=0)}{1-P(evento|x=0)} = \frac{\frac{1}{1+e^{-\beta_0}}}{\frac{e^{-\beta_0}}{1+e^{-\beta_0}}} = e^{\beta_0}$$

El **odds-ratio** se define como el cociente entre los odds con $x=1$ y los odds con $x=0$:

$$OR = \frac{odds(evento|x=1)}{odds(evento|x=0)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

INTERPRETACIÓN DE LOS PARÁMETROS: VARIABLE REGRESORA DICOTÓMICA

Por lo tanto, la **exponencial del parámetro** representa el odds-ratio asociado a la variable y, en consecuencia, permite aproximar **cuánto más probable** (o improbable) es que se de el **evento** entre los individuos con $x = 1$ frente a los individuos con $x = 0$.

INTERPRETACIÓN DE LOS PARÁMETROS: VARIABLE REGRESORA CATEGÓRICA

Cuando se desea analizar el efecto de una variable regresora categórica es necesario tener en cuenta que el modelo tendrá **tantos parámetros β como categoría menos uno**.

La interpretación de cada uno de estos parámetros **coincide con la de las variables dicotómicas** pero teniendo en cuenta que la comparación a la que se refiere el odds ratio es la de la **categoría correspondiente y la de referencia**.

INTERPRETACIÓN DE LOS PARÁMETROS: VARIABLE REGRESORA CONTINUA

En el caso de variables regresoras continuas, la exponencial del parámetro concide con el odds ratio asociado al **efecto de un incremento unitario** en la variable:

$$OR = \frac{odds(evento|x = a+1)}{odds(evento|x = a)} = \frac{e^{\beta_0+\beta_1(a+1)}}{e^{\beta_0+\beta_1 a}} = e^{\beta_1}$$

En ocasiones (si la variable toma valores pequeños o grandes) evaluar el efecto de un incremento unitario puede **no tener sentido**. En ese caso, se puede evaluar el odds ratio asociado a un incremento de c **unidades** (donde c puede ser tan grande o pequeño como queramos; por ejemplo, una decima o un millón):

$$OR = \frac{odds(evento|x = a+c)}{odds(evento|x = a)} = \frac{e^{\beta_0+\beta_1(a+c)}}{e^{\beta_0+\beta_1 a}} = e^{c\beta_1}$$

Es importante recordar que cuando hay varias variables en el modelo la interpretación de los parámetros ha de hacerse asumiendo que el resto de variables se **mantienen constantes**. Es decir, se evalúa el efecto de cierta variable x_i entre dos individuos que poseen exactamente las mismas características restantes.

En regresión logística también se pueden añadir **interacciones** (como se estudió en el tema anterior). Su interpretación va en línea con las interpretaciones recien vistas.

INTERPRETACIÓN DE LOS PARÁMETROS: VARIABLES CUALITATIVAS DICOTÓMICAS

$$OR = \frac{odds(evento|x_i = 1)}{odds(evento|x_i = 0)} = e^{\beta_i}$$

- Si e^{β_i} **es mayor que 1**, la ODD del evento si la categoría de la variable i-ésima es la 1, entonces es e^{β_i} veces mayor que la ODD del evento si la categoría de la variable i-ésima es la 0.
- Si e^{β_i} **es menor que 1**, la ODD del evento si la categoría de la variable i-ésima es la 0, entonces es $\frac{1}{e^{\beta_i}}$ veces mayor que la ODD del evento si la categoría de la variable i-ésima es la 1.

INTERPRETACIÓN DE LOS PARÁMETROS: VARIABLES CUALITATIVAS (MÁS DE DOS CATEGORÍAS)

La interpretación es la misma que en el caso anterior (variables cualitativas dicotómicas), con la única diferencia que la comparación en lugar de hacerlo con la categoría 0 de la variable dicotómica se hace con la categoría de referencia de la variable cualitativa.

INTERPRETACIÓN DE LOS PARÁMETROS: VARIABLES CUANTITATIVAS

$$OR = \frac{odds(evento|x_j = a+1)}{odds(evento|x_j = a)} = e^{\beta_j}$$

- Si e^{β_j} **es mayor que 1**, la ODD del evento añadiendo una unidad a la variable X_j es de e^{β_j} veces mayor que la ODD del evento sin añadir esta unidad.
- Si e^{β_j} **es menor que 1**, la ODD del evento sin añadir una unidad a la variable X_j es de $\frac{1}{e^{\beta_j}}$ veces mayor que la ODD del evento añadiendo esta unidad.

ESTIMACIÓN DE LOS PARÁMETROS

Dado que los parámetros que definen la regresión logística son **desconocidos**, es necesario estimarlos utilizando la información del conjunto de datos del que dispongamos.

A diferencia de la regresión lineal, no es posible estimar los parámetros mediante minimos cuadrados, por lo que lo habitual es utilizar el método de **máxima verosimilitud**.

La verosimiltud es un concepto muy similar al de probabilidad. La idea de este método es estimar los parámetros como aquellos valores que **maximicen la probabilidad** del conjunto de datos con el que estamos trabajando. La razón detrás de este argumento es que si hemos observado dichos datos en la realidad esto debe deberse a que era **lo más probable**.

Dado que las observaciones son independientes, la función de verosimilitud viene dada por:

$$L(\boldsymbol{\beta}) = \prod_{i/y_i=1} p_{1i} \prod_{j/y_j=0} (1 - p_{1j}),$$

donde p_{1i} y p_{1j} representan las probabilidades del suceso que vienen dadas por:

$$p_{1i} = P(Y = 1|x_{1i}, x_{2i}, \ldots, x_{mi}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_m x_{mi})}}.$$

No existe una fórmula explicita para la obtención de los parámetros que maximizan la verosimilitud, por lo que será necesario recurrir a **métodos iterativos de optimización**.

Una vez estimados los valores de los parámetros, se debe **evaluar la utilidad** de las variables, así como la bondad del modelo en su conjunto.

Al igual que ocurre con el modelo de regresión lineal, en minería de datos no vamos a darle credibilidad a los contrastes de hipótesis debido principalmente al gran número de observaciones con el que contamos.

El contraste de **significación global** del modelo se basa en **comparar** las verosimilitudes del modelo a analizar y el modelo "básico" en el que sólo hay un parámetro (β_0) que da lugar a un **valor de predicción único** para todas las observaciones que coincide con la **proporción de eventos** del conjunto de datos.

La idea que subyace a este contraste es que si las verosimilitudes de ambos modelos son **similares**, el modelo analizado no es mejor que el "no modelo" y, por tanto, **no resultará útil**. Por el contrario, si la verosimilitud del modelo analizado es mucho mayor (recordemos que el método de estimación utilizado es el de **máxima** verosimilitud), el modelo permitirá obtener estimaciones más precisas.

Es habitual obtener el valor de $-2(\log(L))$ en lugar de la propia verosimulitud L. Nótese que en este caso, cuanto menor sea esta cantidad, **mejor será el modelo**.

Siguiendo con esta idea, podemos obtener una medida de la calidad de los modelos de regresión logística a partir del $Pseudo - R^2$ **de McFadden**, de la siguiente forma:

$$Pseudo - R^2 = 1 - \frac{-2(\log(L_{modelo}))}{-2(\log(L_{no\_modelo}))}.$$

Como se puede observar, **valores altos** del $Pseudo - R^2$ son indicativos de grandes diferencias entre la verosimilitud del modelo y el "no modelo" por lo que el modelo será mucho **mejor** que "no hacer nada".

Es importante destacar que este indicador no suele tomar valores tan grandes como el R^2 de regresión lineal. Estudios en la literatura indican que valores del $Pseudo - R^2$ en el intervalo $(0.2; 0.4)$ serían equivalentes a valores de $(0.7; 0.9)$ del R^2 en modelos lineales.

Además del contraste de hipótesis global, se puede evaluar la **significatividad de los parámetros** a través del denominado como **test de Wald** cuyo estadístico viene dado por el cociente entre la estimación del parámetro y su desviación típica.

Recordemos que ignoraremos el resultado de los p-valores, aunque la magnitud de los estadísticos nos podrán indicar la importancia relativa de los parámetros.

Así mismo, podemos calcular la **importancia de las variables** obteniendo los modelos sin cada una de ellas y observando la **reducción** en el $Pseudo - R^2$ que se produce.

OTRAS MEDIDAS DE EVALUACIÓN

Cuando se contruye un modelo de predicción para una variable objetivo binaria, se obtiene lo que se conoce como **matriz de confusión**, que contiene el número de observaciones que han sido **bien/mal clasificadas**:

	Predicción = 0	Predicción = 1
Realidad = 0	**VN** Verdadero negativo	**FP** Falso positivo
Realidad = 1	**FN** Falso negativo	**VP** Verdadero positivo

Utilizando esta información, se pueden definir las siguientes medidas de clasificación:

- Tasa de acierto: $\frac{VN+VP}{VN+FP+FN+VP}$
- Tasa de fallo: $\frac{FP+FN}{VN+FP+FN+VP}$
- Sensibilidad o tasa de verdaderos positivos: $\frac{VP}{FN+VP}$
- Especificidad o tasa de verdaderos negativos: $\frac{VN}{VN+FP}$
- Valor predictivo positivo (VPP): $\frac{VP}{FP+VP}$
- Valor predictivo negativo (VPN): $\frac{VN}{VN+FN}$

CURVA ROC

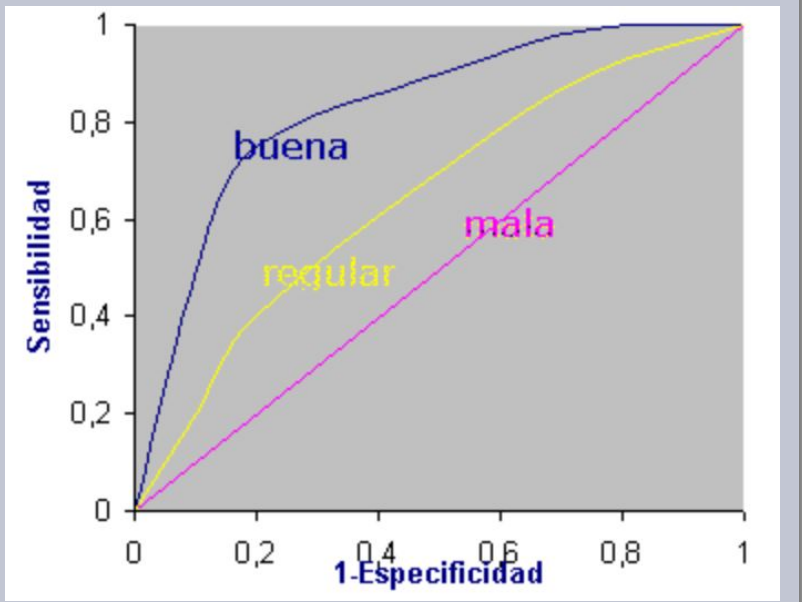
Para obtener la matriz de confusión, es necesario determinar una **probabilidad** a partir de la cuál se considera una observación como "evento" o "no evento". Generalmente, esta probabilidad de referencia es 0.5 pero, como veremos más adelante, no tiene porqué ser así.

La siguiente gráfica permite obtener una medida de la evaluación del modelo **sin necesidar de clasificar cada observación** como evento o no evento.

La curva ROC muestra la **sensibilidad frente a 1-especificidad** (tasa de falsos positivos) para todos los posibles valores de corte de la probabilidad a posteriori.

Un modelo **perfecto** es aquel que otorga probabilidad 0 a los no eventos y 1, a los eventos. Por lo tanto, existe un corte de la probabilidad tal que clasifica correctamente todas las observaciones y que por tanto da lugar a una sensibilidad y una especificidad de 1. Esa curva tendrá forma de **ángulo recto**.

Por el contrario, si asignamos **aleatoriamente** un valor entre 0 y 1 a cada observación (sería un modelo totalmente inútil), obtendremos una curva ROC que concide con la **bisectriz**.



Curva ROC

Por lo tanto, cuánto **más cóncava** sea una curva, **mejor** será el método que estamos evaluando.

Dado que hay ocasiones en las que visualmente es **dificil discernir** qué curva es más cóncava, se define el **área bajo la curva** ROC (AUROC) como medida de evaluación del modelo. Un modelo **perfecto tendrá una AUROC de 1**, dado que define un cuadrado de lado unidad. Por el contrario, un "mal" modelo tendrá un AUROC de 0.5, dado que define un triángulo equivalente a la mitad del cuadrado anterior.

Por tanto, **cuánto mayor** sea el área bajo la curva ROC, **más poder predictivo** tendrá el modelo.

Técnicas de remuestreo

Las técnicas de remuestreo estudiadas en el tema de regresión lineal, **Validación Cruzada**, también son aplicables a regresión logística.

La principal diferencia es la medida que se utilice para **evaluar** pues, en este caso, se deben utilizar **medidas específicas** de modelos con variable objetivo de clase, como son la tasa de clasificación errónea, el AIC, el BIC, el área bajo la curva ROC, etc.

Las predicciones obtenidas a partir de la regresión logística son las probabilidades del evento.

Sin embargo, el objetivo principal es clasificar las observaciones como **"evento" o "no evento"**. Recordemos que los "no eventos" deberían tener probabilidades pequeñas y, los eventos, probabilidades grandes. Por lo tanto, será necesario decidir **a partir de qué probabilidad** consideraremos una observación como "evento".

SELECCIÓN DEL PUNTO DE CORTE

A la hora de seleccionar la probabilidad anterior (o punto de corte), existen **distintas estrategias**. No obstante, todas ellas consisten en **evaluar**, desde distintos puntos de vista, la **matriz de confusión** que se deriva de **todos** los posibles puntos de corte.

Recordemos que una matriz de confusión viene dada por:

	Predicción = 0	Predicción = 1
Realidad = 0	**A** Verdadero negativo	**B** Falso positivo
Realidad = 1	**C** Falso negativo	**D** Verdadero positivo

- La estrategia más frecuente es **minimizar** la tasa de **mal clasificados**, lo que se consigue habitualmente fijando el punto de corte en **0.5**.

SELECCIÓN DEL PUNTO DE CORTE

Otras estrategias habituales, para las que habrá que evaluar todos los posibles puntos de corte y seleccionar el que ofrezca mejores resultados, son:

- Seleccionar el punto de corte que **maximiza la tasa de aciertos**.
- Seleccionar el punto de corte que **maximiza el índice de Youden**:

$$Youden = sensibilidad + especificidad - 1.$$

- Fijar una **especificidad mínima** y maximizar la sensibilidad.
- Fijar una **sensibilidad mínima** y maximizar la especificidad.

El punto de corte debe seleccionarse a partir del conjunto de datos *test*, lo que ofrecerá resultados más realistas.

A la hora de comparar modelos existen dos estrategias:

- Obtener el mejor punto de corte para todos los modelos y comparar los resultados para dicho punto de corte.
- Comparar los modelos a partir de estadísticos que no dependan del punto de corte (como el área bajo la curva ROC o el AIC) y, a continuación, obtener el mejor punto de corte únicamente para el modelo seleccionado.

Cuando la variable objetivo toma **más de dos valores** diferentes (es categórica, pero no binaria), también es posible ajustar un modelo de regresión logística:

Para generalizar el modelo anterior, simplemente tendremos que **generar tantas** funciones *logit* como número de clases k menos una:

$$g_1(\boldsymbol{x}) = \log\left(\frac{p_1(\boldsymbol{x})}{p_k(\boldsymbol{x})}\right) = \beta_{10} + \beta_{11}x_1 + \cdots + \beta_{1m}x_m$$

$$g_2(\boldsymbol{x}) = \log\left(\frac{p_2(\boldsymbol{x})}{p_k(\boldsymbol{x})}\right) = \beta_{20} + \beta_{21}x_1 + \cdots + \beta_{2m}x_m$$

$$\vdots$$

$$g_{k-1}(\boldsymbol{x}) = \log\left(\frac{p_{k-1}(\boldsymbol{x})}{p_k(\boldsymbol{x})}\right) = \beta_{(k-1)0} + \beta_{(k-1)1}x_1 + \cdots + \beta_{(k-1)m}x_m,$$

donde k es la **clase de referencia**, cuya elección es arbitraria pues da lugar a las mismas probabilidades de pertenencia, que se calculan como:

$$p_i(\boldsymbol{x}) = \frac{e^{g_i(\boldsymbol{x})}}{1 + \sum_{j=1}^{k-1} e^{g_j(\boldsymbol{x})}}, \forall i = 1, \ldots, k-1 \qquad \text{y} \qquad p_k(\boldsymbol{x}) = 1 - \sum_{i=1}^{k-1} p_i(\boldsymbol{x}).$$

El modelo de regresión logística para variables categóricas consiste, por tanto, en $k-1$ modelos de regresión logística 'simples' que **comparan cada categoría con la de referencia**.