

Introducción a la Ciencia de Datos

Maestría en Ciencias
de la Computación

Dr. Irvin Hussein López Nava

IS LOGISTIC REGRESSION



REGRESSION?

The slide features several decorative squares of various shades of blue and white. A large blue square with the number '03' is centered near the top. Other smaller squares are scattered around the slide, including a light blue square in the top left, a dark blue square in the top right, a 2x2 grid of squares (white, light blue, dark blue, light blue) in the middle right, a light blue square in the bottom left, a dark blue square in the bottom right, and a light blue square in the bottom right corner. A small white square is also visible in the middle right area.

03

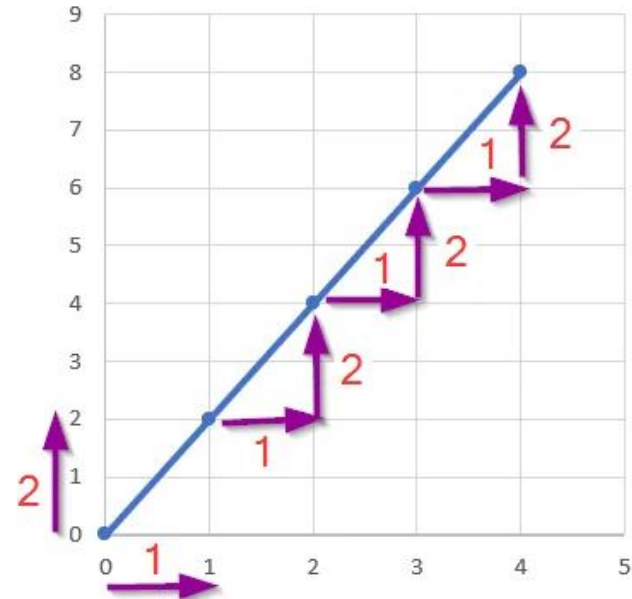
Regression

Recordando

- En regresión lineal, se intenta estimar la variable continua Y dependiendo de cualquier tipo de variable X cuando existe una **relación lineal** entre X e Y.
 - Relación lineal significa que hay un patrón entre X e Y, e.g. al aumentar X una cantidad, se encuentra un cambio constante en Y.
- Esto se representa como una ecuación lineal **$Y = mX + c$**
- Si **c** es 0, entonces la ecuación se convierte en **$Y = mX$**
 - Si **m** = 2, entonces $Y = 2X$, i.e., al aumentar 1 unidad de Y, hay 2 unidades de aumento en X.

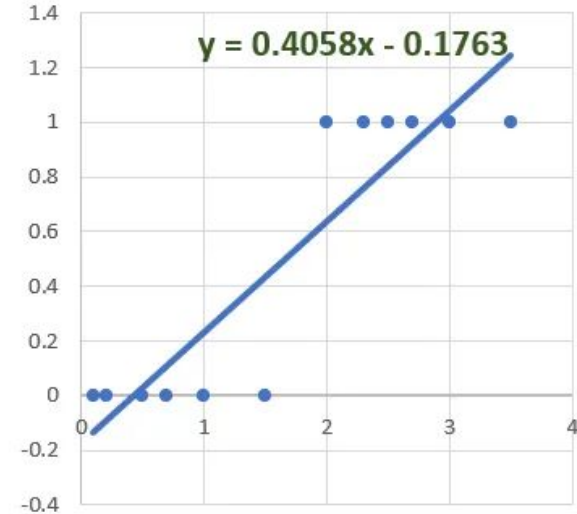
Visualmente

- Entonces, en regresión lineal, el cambio de Y es constante para cualquier valor de X.
- El valor de Y aumenta constantemente 2 veces el de X.



Ahora bien

- Considere que el valor de Y no es continuo. Si el valor de Y es **discreto**, y binomial, entonces el valor de Y sólo puede ser 0 o 1.
- El verdadero problema ocurre al pretender **predecir** nuevos valores.
- En este ejemplo, para un valor más alto de X, los valores de Y van más allá de 1.
 - Pero los posibles valores son 0 o 1.



¿qué se puede hacer?

- Entonces, ¿cómo podemos mapear/normalizar los valores de Y dentro de 0 y 1 para cualquier valor correspondiente de X ?
- Se requiere un efecto constante en Y para un valor de X , pero no debe ir más allá de 1 ni por debajo de 0.
- Esto se puede representar por una variable de Bernoulli donde las probabilidades están acotadas en ambos extremos (entre 0 y 1).

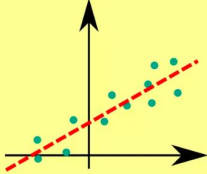
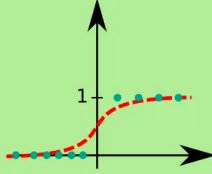
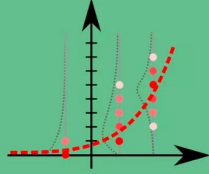
Limitaciones de la regresión lineal

- En regresión lineal las predicciones no son adecuadas para la clasificación, ya que la probabilidad verdadera debe estar entre 0 y 1, porque pueden ser mayores que 1 o menores que 0.
 - La clasificación no se distribuye normalmente, lo que viola el supuesto de **normalidad**.
 - Cualquier factor que afecte la probabilidad cambiará no solo la media sino también la varianza de las observaciones, lo que viola además el supuesto de **homocedasticidad**.
- Como resultado, no se puede aplicar directamente la regresión lineal.

Generalised linear models (GLMs)

- Un modelo de probabilidad para predecir valores en la escala de 0 a 1 es el **modelo lineal generalizado** (GLM), el cual permite variables de respuesta que tienen distribuciones arbitrarias (diferentes a distribuciones normales).
 - Emplea una función de enlace para variar linealmente con los valores predichos en lugar de asumir que la respuesta misma debe variar linealmente con el predictor.
- En nuestro caso anterior, se puede aplicar una función logística (también llamada '*inverse logit*' o '*sigmoid function*').
 - La **regresión logística** es un tipo de modelo lineal generalizado.

Tres tipos de regresiones

LINEAR REGRESSION	LOGISTIC REGRESSION	POISSON REGRESSION
<ul style="list-style-type: none"> ① Econometric modelling ② Marketing Mix Model ③ Customer Lifetime Value 	<ul style="list-style-type: none"> ① Customer Choice Model ② Click-through Rate ③ Conversion Rate ④ Credit Scoring 	<ul style="list-style-type: none"> ① Number of orders in lifetime ② Number of visits per user
		
Continuous \Rightarrow Continuous	Continuous \Rightarrow True/False	Continuous \Rightarrow 0,1,2,...
$y = \alpha_0 + \sum_{i=1}^N \alpha_i x_i$	$y = \frac{1}{1 + e^{-z}}$ $z = \alpha_0 + \sum_{i=1}^N \alpha_i x_i$	$y \sim \text{Poisson}(\lambda)$ $\ln \lambda = \alpha_0 + \sum_{i=1}^N \alpha_i x_i$
<code>lm(y ~ x1 + x2, data)</code>	<code>glm(y ~ x1 + x2, data, family=binomial())</code>	<code>glm(y ~ x1 + x2, data, family=poisson())</code>
1 unit increase in x increases y by α	1 unit increase in x increases log odds by α	1 unit increase in x multiplies y by e^α

3.2 Regresión logística

¿Qué es Regresión logística?

A woman with dark hair, wearing an orange sweater, is shown in profile, looking upwards and to the right. She is holding a tablet in her left hand and reaching out with her right hand towards a glowing point on a digital globe. The background is a complex, futuristic digital interface with a blue and white color scheme. It features a large globe in the center, surrounded by various data visualizations such as bar charts, line graphs, and circular progress indicators. The overall aesthetic is high-tech and data-driven.

Definiciones

ChatGPT

Es un método de análisis estadístico utilizado para modelar y predecir la probabilidad de un evento binario, es decir, un evento que tiene solo dos posibles resultados, como sí/no, 1/0, éxito/fracaso, o clases positivas/negativas.

<https://chat.openai.com/>

Wikipedia

Es un modelo estadístico que modela la probabilidad de que ocurra un evento haciendo que las probabilidades logarítmicas del evento sean una combinación lineal de una o más variables independientes.

https://en.wikipedia.org/wiki/Logistic_regression

Gemini

Es un tipo de análisis de regresión que se utiliza para predecir el resultado de una variable categórica, como si un paciente tiene cáncer o no, si un cliente comprará un producto o no, o si un candidato ganará una elección o no.

<https://gemini.google.com/>

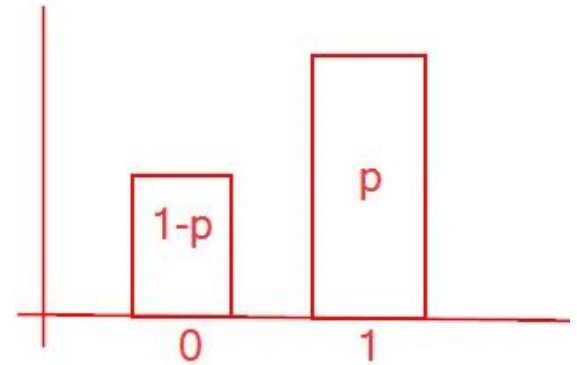
Tipos de regresión logística

- **Regresión logística binomial** o binaria – dos resultados posibles "0" y "1"
Ejemplo: Éxito/Fracaso, Sí/No.
- **Regresión logística multinomial** – más de dos resultados que no pueden ordenarse. **Ejemplo:** Selección de curso: Curso A, Curso B, Curso C.
- **Regresión logística ordinal** – similar a la multinomial, pero los resultados pueden ordenarse. **Ejemplo:** Clasificación de películas: Excelente, Buena, Regular, Mala.

Cuando hay dos resultados posibles, puede compararse con la distribución de Bernoulli.

Distribución de Bernoulli

- Es una distribución de probabilidad de una variable aleatoria que usa 0 o 1. Como usa solo 2 clases, es un modelo binario/binomial.
- La distribución de Bernoulli, que lleva el nombre del matemático suizo Jacques Bernoulli (1654–1705), describe un experimento probabilístico en el que una prueba tiene dos resultados posibles, un éxito o un fracaso.



Probabilidad

- Si la probabilidad de éxito es p , entonces la probabilidad de fracaso es la de todas las demás posibilidades restantes.
- Entonces, la probabilidad de falla se puede escribir como $1-p$, al denotarla con una variable diferente, $q = 1-p$.
- Sin embargo, tenerlo como $1-p$ facilita la comprensión cuando se trata de derivaciones.

$$p(r;p) = \begin{cases} 1 - p = q & \text{if } r = 0 \text{ (failure)} \\ p & \text{if } r = 1 \text{ (success)} \end{cases}$$

Generalización

- La distribución anterior se generaliza a la siguiente forma:

$$P(X = x) = p^x(1 - p)^{1-x}$$

- Para entender la ecuación previa, se puede establecer x con los valores 0 y 1.
 - Para x= 1 (éxito), $P(X = 1) = p^1(1 - p)^{1-1} = p$
 - Para x= 0 (fracaso), $P(X = 1) = p^0(1 - p)^{1-0} = 1 - p$
- Ahora se obtienen los valores de éxito y fracaso de la ecuación $p(r; p)$.

Ejemplo

- La **probabilidad** puede entenderse como el número de veces que se produjo el éxito en comparación con el número total de intentos.
- Para 10 eventos, y un número de éxitos de 8, entonces
 - Probabilidad de éxito = $8 / 10 = 0.8$
 - Probabilidad de fracaso = $1 - 0.8 = 0.2$
- Ahora descubrimos cuál es la naturaleza de nuestro valor Y. Es un valor binomial y puede describirse como probabilidad de éxito y fracaso.

Odds

- *Odds Ratio* es otra medida para saber qué tan probable es que algo ocurra. Las cuotas, o momios, son la cantidad de veces que ocurrió el éxito en comparación con la cantidad de veces que ocurrió el fracaso.
 - Odds Ratio (Éxito) = Probabilidad de éxito / Probabilidad de fracaso
 - Odds Ratio (Fracaso) = Probabilidad de fracaso / Probabilidad de éxito
- Es posible determinar los *odds ratio* a partir de las probabilidades. Si la probabilidad de éxito es p ,
 - Odds Ratio = $p / (1-p)$

Ejemplo

- Los *odds ratio* varían de 0 a infinito.
- Si la probabilidad de éxito es 0.8, entonces el *odds ratio* es
 - $Odds\ Ratio = 0.8 / (1-0.8) = 0.8 / 0.2 = 4$
 - el *Odds Ratio* es 4:1, i.e., por cada 4 éxitos hay 1 fracaso.
- En otro caso, si p es 0.25, el $OR = 0.25 / 0.75 = 1/3$
 - Entonces por cada éxito hay 3 fracasos.
- Pero esta *Odds Ratio* comienza desde 0, pero llega hasta el infinito.
- ¿Cómo cambiar el rango? Tenemos al logaritmo.

Logaritmo

- $\log(x)$ es un logaritmo en base 10. También se puede escribir como $\log_{10}(x)$.
- $\ln(x)$ significa el logaritmo en base e . También se puede escribir como $\log_e(x)$.
- Logaritmo y exponencial son inversos entre sí; e^x es la inversa de $\ln(x)$.
- $\text{Log}_e(a) = e^x$

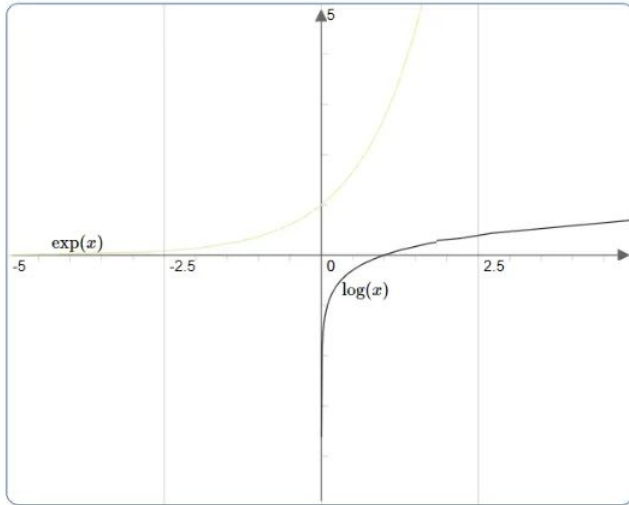
$$\left. \begin{array}{l} \ln x \text{ and } e^x \text{ are} \\ \text{inverses of each other} \end{array} \right\} \iff \left\{ \begin{array}{l} \ln(e^x) = x, \text{ for all } x \\ e^{\ln x} = x, \text{ for all } x > 0 \end{array} \right.$$

$$e^0 = 1 \iff \ln 1 = 0$$

$$e^1 = e \iff \ln e = 1$$

$$e^{(\ln x)} = x \iff \ln e^x = x$$

Ejemplo



A la transformación de los valores de probabilidad a probabilidades logarítmicas se le llama **función de enlace**.

S.No.	Probability of Success	Odds PS/(1-PS)	LogOdds Log(Odds)
1	0.001	0.001001001	-2.999565488
2	0.01	0.01010101	-1.995635195
3	0.15	0.176470588	-0.753327667
4	0.2	0.25	-0.602059991
5	0.25	0.333333333	-0.477121255
6	0.3	0.428571429	-0.367976785
7	0.35	0.538461538	-0.268845312
8	0.4	0.666666667	-0.176091259
9	0.45	0.818181818	-0.087150176
10	0.5	1	0
11	0.55	1.222222222	0.087150176
12	0.6	1.5	0.176091259
13	0.65	1.857142857	0.268845312
14	0.7	2.333333333	0.367976785
15	0.75	3	0.477121255
16	0.8	4	0.602059991
17	0.85	5.666666667	0.753327667
18	0.9	9	0.954242509
19	0.999	999	2.999565488
20	0.9999	9999	3.999956568

De regresión lineal a regresión logística

- Una función de enlace transforma las probabilidades de los niveles de una **variable de respuesta** categórica a una escala continua ilimitada.
 - Una vez completada la transformación, la relación entre los **predictores** y la respuesta se puede modelar con regresión lineal.
- Por ejemplo, una variable de respuesta binaria puede tener dos valores únicos. La conversión de estos valores a probabilidades hace que la **variable de respuesta** oscile de 0 a 1.
 - Cuando aplica una función de enlace adecuada a las probabilidades, los números resultantes oscilarán entre $-\infty$ y $+\infty$.

Función logística

Y value is a Probability value P here.

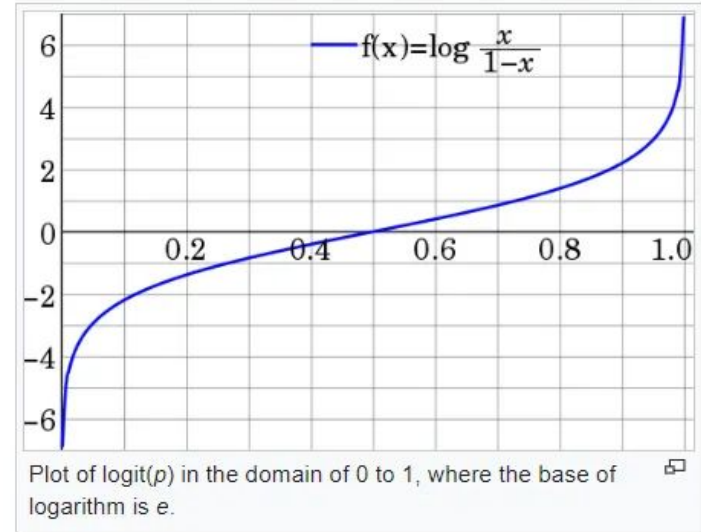
$$\text{So, } y = \ln\left(\frac{p}{1-p}\right)$$

From Linear reg. $Y = \beta_0 + \beta_1 X_1$

Let's denote it as $X\beta$, $Y = \beta_0 + \beta_1 X_1 = X\beta$.

$$X\beta = \ln\left(\frac{p}{1-p}\right)$$

This function is a logistic function, its graph is given below

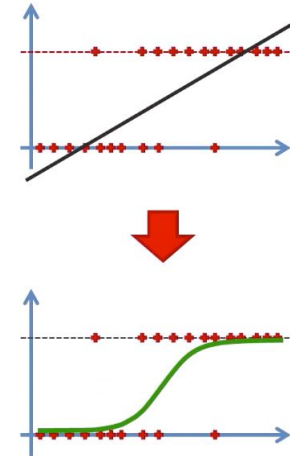
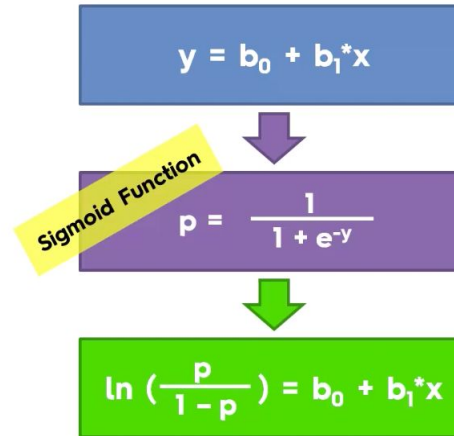


De regresión lineal a regresión logística

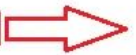
$$\ln\left(\frac{P}{1-P}\right) = b_0 + b_1 x$$

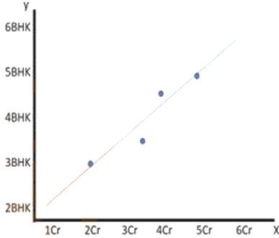
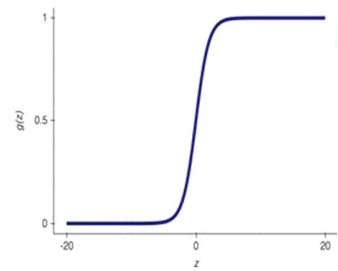
$$\frac{P}{1-P} = e^{b_0 + b_1 x}$$

$$P = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$$

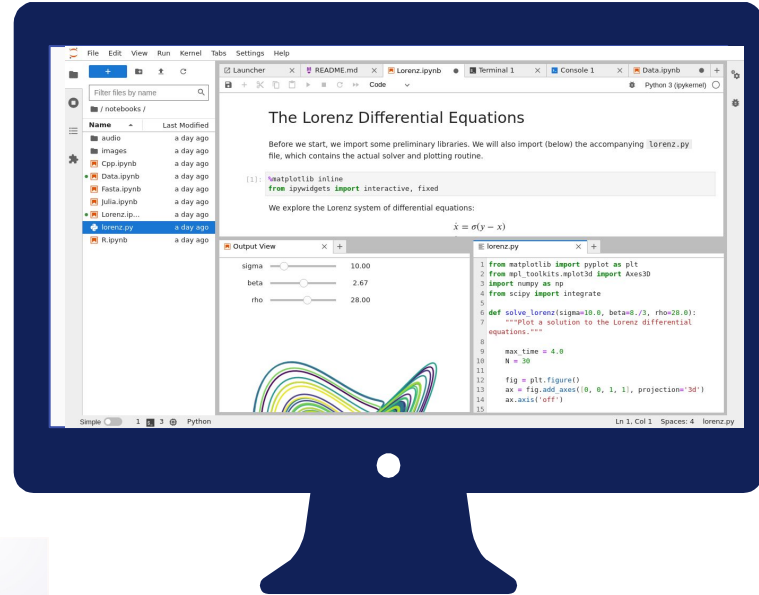


Finalmente

- Si se conocen los coeficientes x y las pendientes, se pueden convertir a los valores de probabilidad requeridos.
$$p = \frac{1}{1 + e^{-(b_0 + b_1 x)}}$$

$$y = b_0 + b_1 x$$
- La función sigmoidea convierte un valor x en una probabilidad de un evento.
 - Si el valor de probabilidad, i.e., el resultado del valor de la función sigmoidea $p \geq 0.5$, entonces la salida se marca como 1 (verdadero/positivo/éxito).
 - Si $p < 0.5$, entonces se marca como 0 (falso/negativo/fracaso).
- Por lo tanto, para cualquier valor de x , se tiene un efecto constante en y y podría oscilar entre 0 y 1 y también podría **clasificarse** en un valor discreto.

Regresión Lineal	Regresión Logística	Explicación
La variable objetivo es de intervalo	La variable objetivo es discreta (binaria u ordinal)	La regresión lineal predice valores continuos (como temperatura, altura, etc.), mientras que la regresión logística predice la probabilidad de que ocurra un evento discreto (como si va a llover o no, si un cliente comprará o no, etc.).
Valores predichos son la media	Valores predichos son probabilidades	La regresión lineal predice el valor promedio de la variable objetivo, mientras que la regresión logística predice la probabilidad de que la variable objetivo pertenezca a una determinada categoría.
Resuelve problemas de regresión	Resuelve problemas de clasificación	La regresión lineal se utiliza para modelar relaciones entre variables continuas, mientras que la regresión logística se utiliza para clasificar datos en categorías.
Ejemplo: ¿Cuál es la temperatura?	Ejemplo: ¿Va a llover o no?	La regresión lineal podría utilizarse para predecir la temperatura en función de otros factores (como la hora del día, la estación del año, etc.), mientras que la regresión logística podría utilizarse para predecir si lloverá o no en función de variables como la humedad, la presión atmosférica, etc.
		<p>La relación entre las variables en la regresión lineal se representa mediante una línea recta, mientras que en la regresión logística se representa mediante una curva en forma de S (curva logística).</p>

(Go to live notebook)



Implementaciones de Scikit-learn



1.1. Linear Models > Logistic regression

https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

Logistic Regression 3-class Classifier

https://scikit-learn.org/stable/auto_examples/linear_model/plot_iris_logistic.html#logistic-regression-3-class-classifier

Métricas

Accuracy

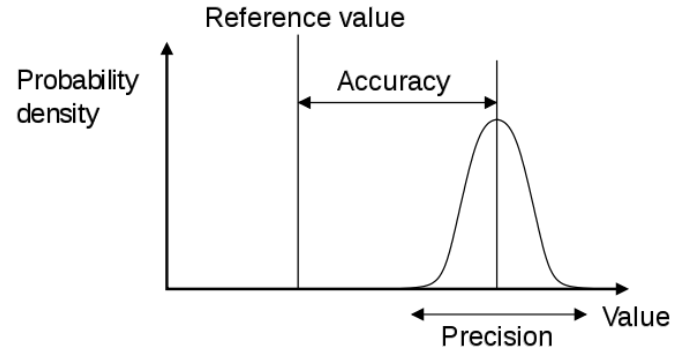
- La exactitud (en inglés "*accuracy*") es una métrica utilizada para evaluar el rendimiento de un modelo de clasificación.
- Representa la proporción de predicciones correctas realizadas por el modelo en relación con el total de predicciones realizadas.
 - En otras palabras, mide qué tan acertado es un modelo en clasificar correctamente los ejemplos en un conjunto de datos.

$$\text{Exactitud} = \frac{\text{Número de predicciones correctas}}{\text{Total de predicciones}}$$

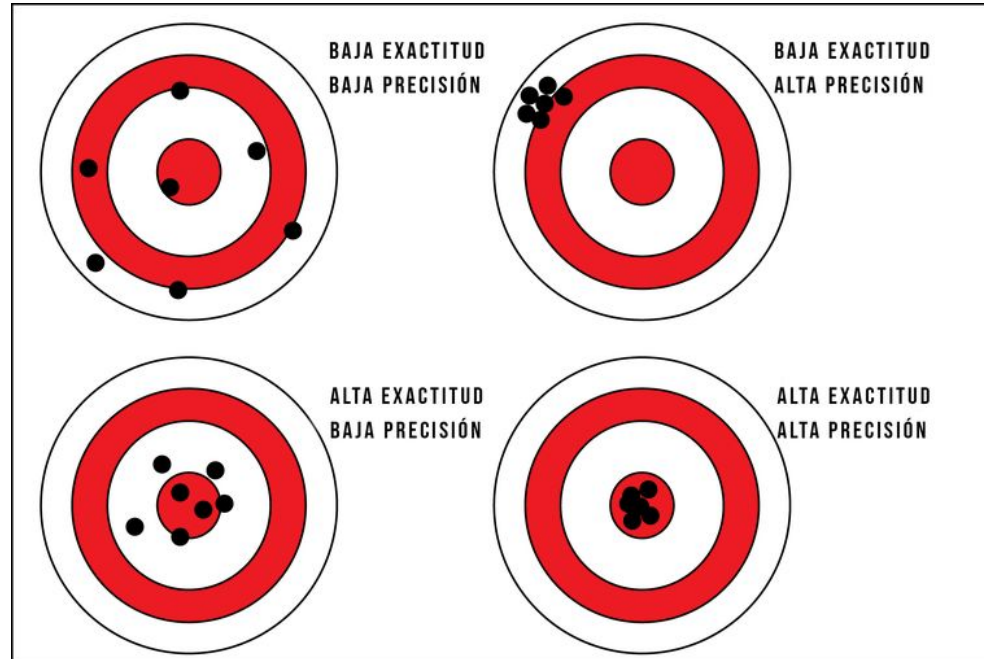
NO confundir
con la métrica
precisión!!

Accuracy vs Precisión

- Son dos métricas de error de observación.
- La **exactitud** es qué tan cerca está un conjunto determinado de mediciones (observaciones) de su valor real, mientras que la **precisión** es qué tan cerca están las mediciones entre sí.

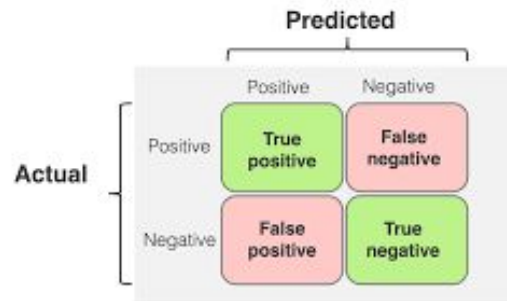


Gráficamente

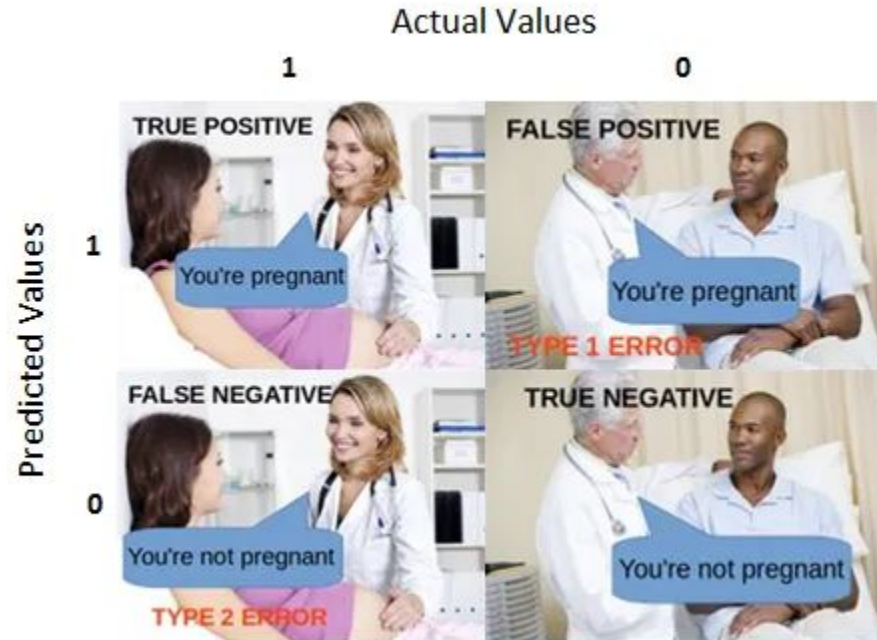


Matriz de confusión (caso binario)

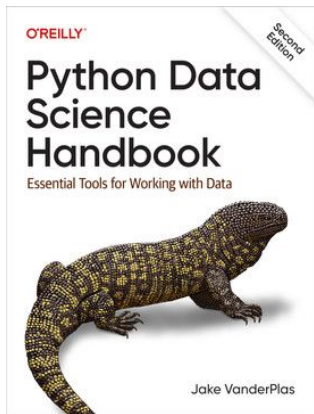
- Es una herramienta que permite la visualización del desempeño de un modelo de clasificación.
- Las columnas pueden representar el número de predicciones de cada clase, mientras que las filas representan las instancias en la clase real, o viceversa.



Pero...



Extra Libro?



Logistic Regression is a classification algorithm and not Regression



Logistic Regression is Regression and models continuous outcome



Logistic Regression is a Binomial regression with logit link



Logistic Regression is a special case of Generalized linear Model (GLM) like Poisson, Beta and Gamma



Logistic Regression can be used as classification only when a probability cut off (e.g. 0.4, 0.5, 0.6 etc) is set externally.



Gracias!

¿Alguna pregunta?

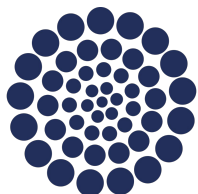
hussein@cicese.mx

<https://sites.google.com/view/husseinlopeznava>

RG

in

X



CONAHCYT

CONSEJO NACIONAL DE HUMANIDADES
CIENCIAS Y TECNOLOGÍAS

CREDITS: This presentation was based on a template by Slidesgo, and includes icons by Flaticon.