Regresión Logística

Licenciatura en Inteligencia Artificial y Ciencia de Datos, CUGDL, Universidad de Guadalajara.

Guadalajara, Jal., agosto de 2025

Modelos supervisados

Clasificación

Recordemos que en Machine Learning, los modelos supervisados pueden abordar dos tipos de problemas:

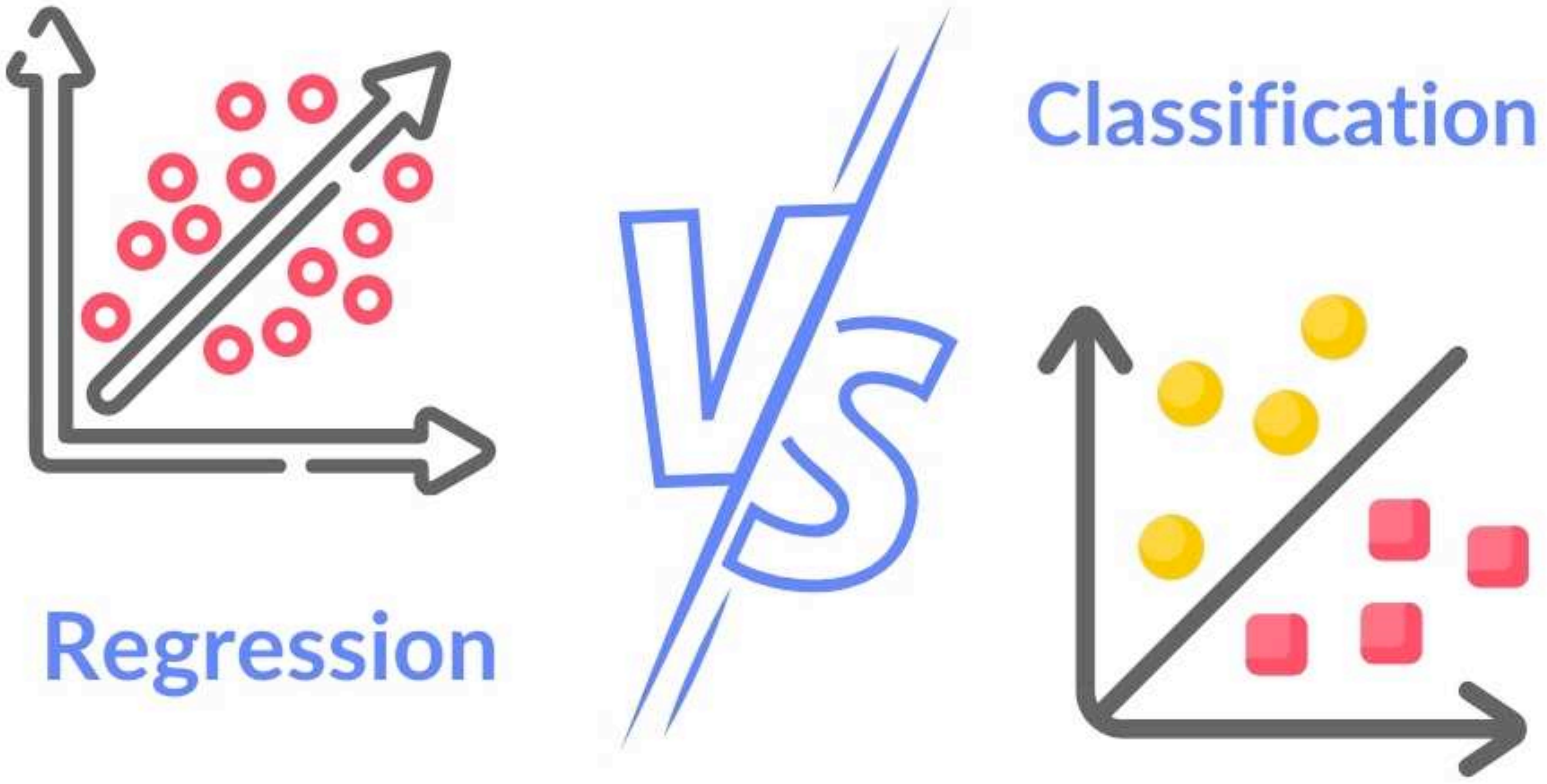
- **Regresión:** cuando la variable objetivo es continua.
- **Clasificación:** cuando la variable objetivo es categórica, es decir, toma valores discretos (e.g., si/no, spam/no spam, enfermo/sano, etc.)

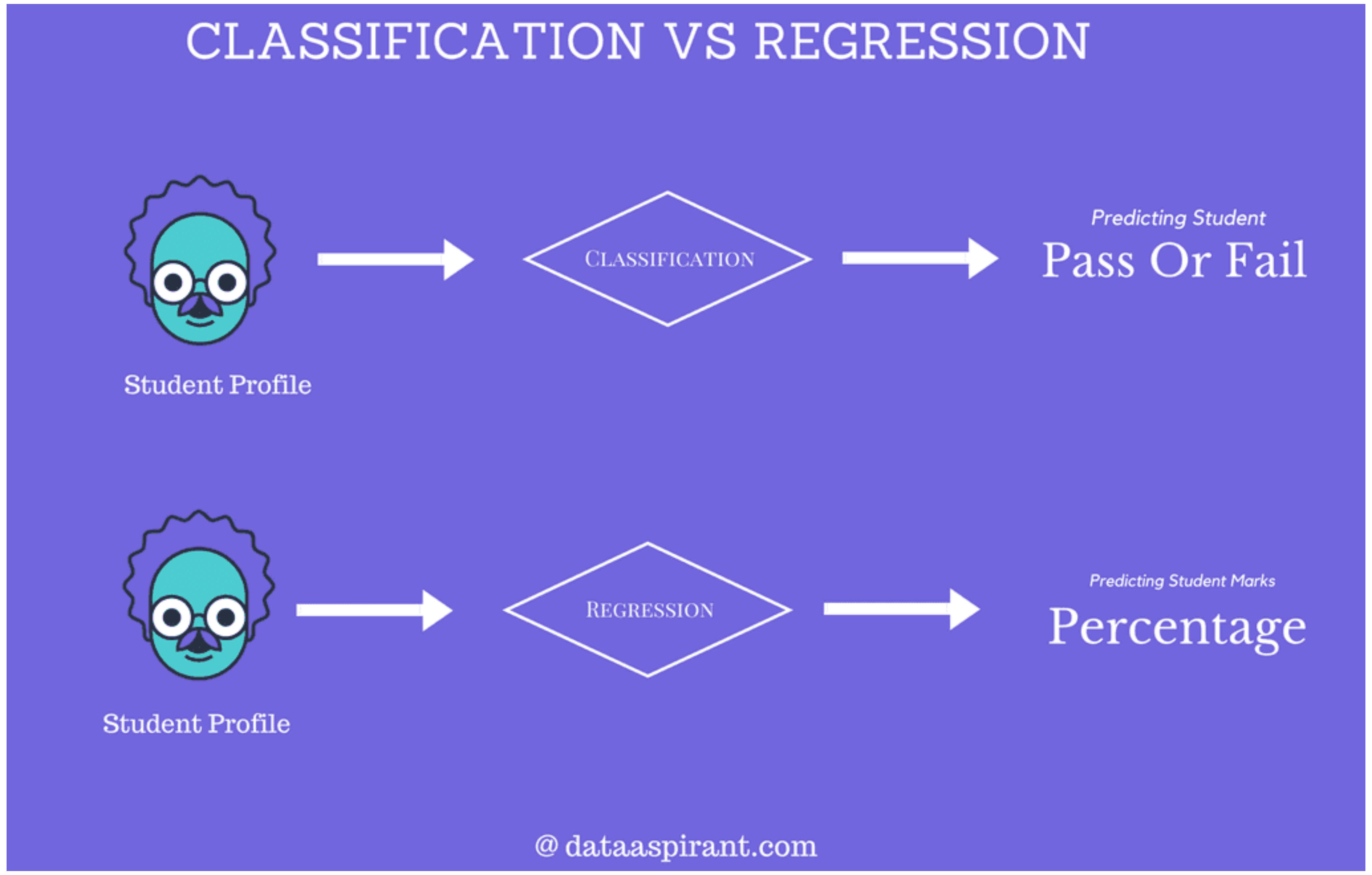
En un problema de **clasificación binaria**, el objetivo es predecir si un objeto pertenece a una de dos clases posibles,

$$y \in \{0,1\}$$

Modelos supervisados

Regresión vs clasificación





Regresión logística

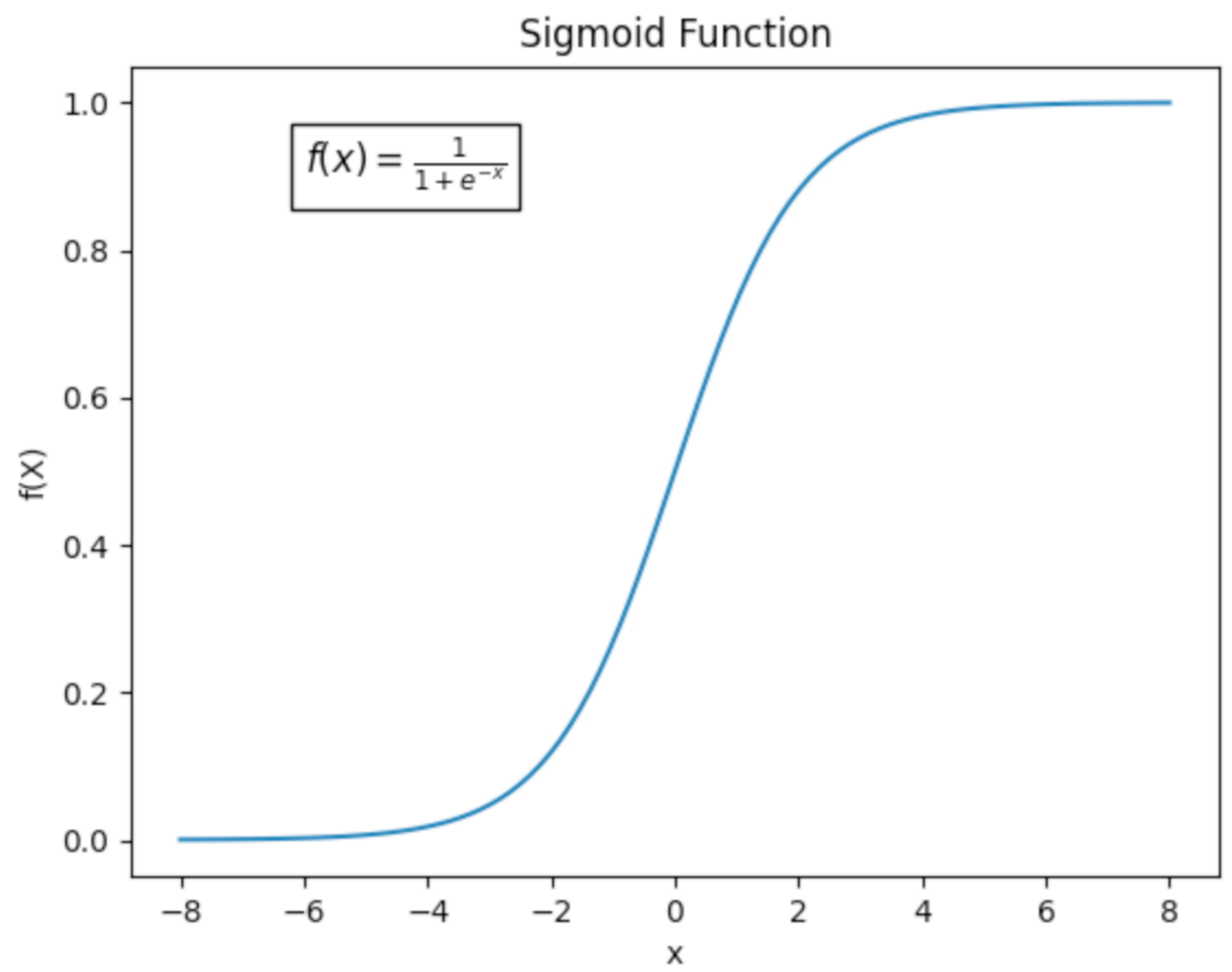
Cómo modelar una probabilidad

- Para clasificar ejemplos binarios necesitamos modelar la **probabilidad** de que una observación pertenezca a una clase,

$$P(y = 1 \,|\, x) = \hat{y}$$

- La **regresión logística** no predice directamente 1 ó 0, sino una probabilidad continua entre 0 y 1. Para lograr esto se usa la llamada **función sigmoide**, que comprime cualquier valor real a este rango,

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$



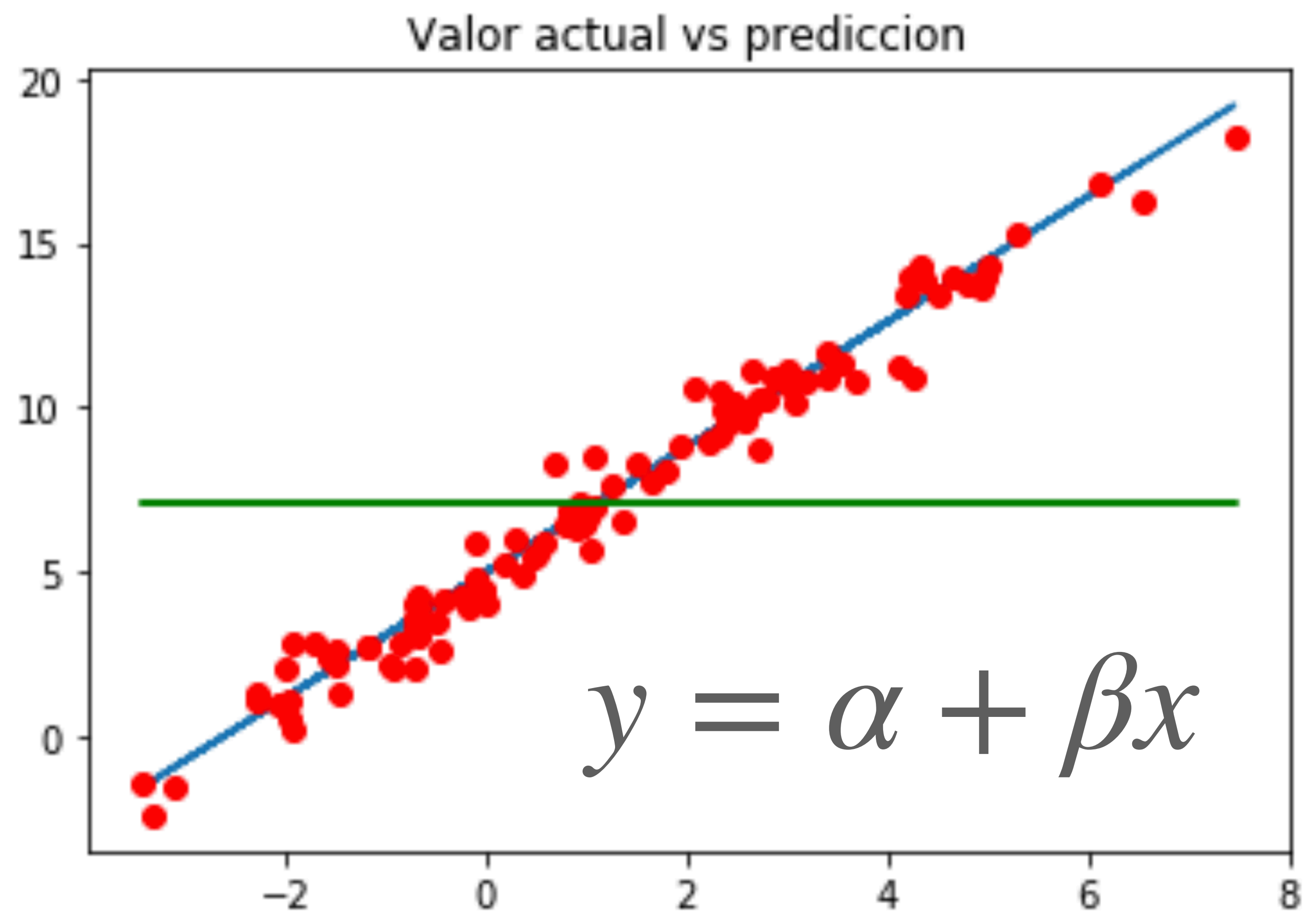
- Esta función tiene ciertas características importantes
- El valor de la función siempre está entre 0 y 1
- Cuando z=0, entonces $\sigma(z) = 0.5$
- Es suave, derivable, y tiene forma de S

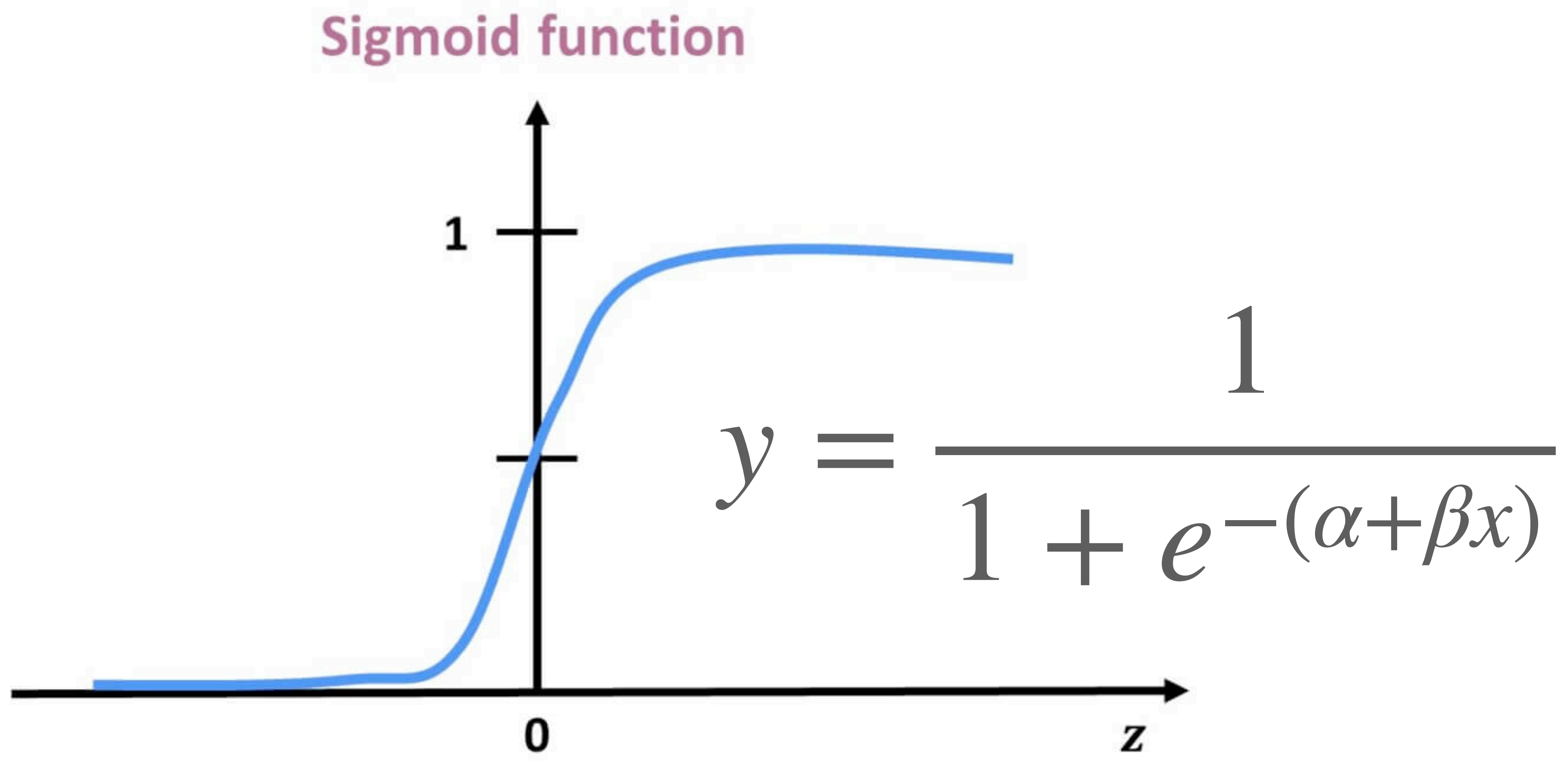
Regresión lineal a regresión logística

Variable continua a variable categórica

- Recordemos que una regresión lineal nos ayudaba a predecir valores continuos.

- Una regresión logística, al igual que el modelo lineal, requiere que las variables predictoras y objetivo tengan una relación lineal.





Regresión lineal a regresión logística

Variable continua a variable categórica

- En la regresión lineal, los valores a predecir y son función de las variables predictoras x, dentro de la ecuación de una recta

$$y = \alpha + \beta \cdot x$$

- En este caso, teóricamente, las variables pueden tomar los siguientes rangos,

$$y \in [-\infty, \infty], \ x \in [-\infty, \infty]$$

- Sin embargo, en la regresión logística lo que buscamos es clasificación, queremos que $y \in \{0,1\}$, o sea, binario.
- Aquí hay que recordar que ambos lados de la ecuación tienen que ser consistentes. Si de un lado tenemos un rango, del otro lado también tenemos que tener el mismo rango.

Regresión lineal a regresión logística
1er parte

- En lugar de buscar un valor y binario, buscamos una probabilidad (variable continua) que se encuentra $P \in [0,1]$, de tal forma que

- $P = \alpha + \beta \cdot x$

- Sin embargo, todavía no coinciden los rangos de ambos lados de la ecuación.

Regresión lineal a regresión logística

2da parte

- Entonces, en lugar de P, podríamos hacer la regresión sobre el cociente de probabilidades,

$$\frac{P}{1-P} = \alpha + \beta \cdot x$$

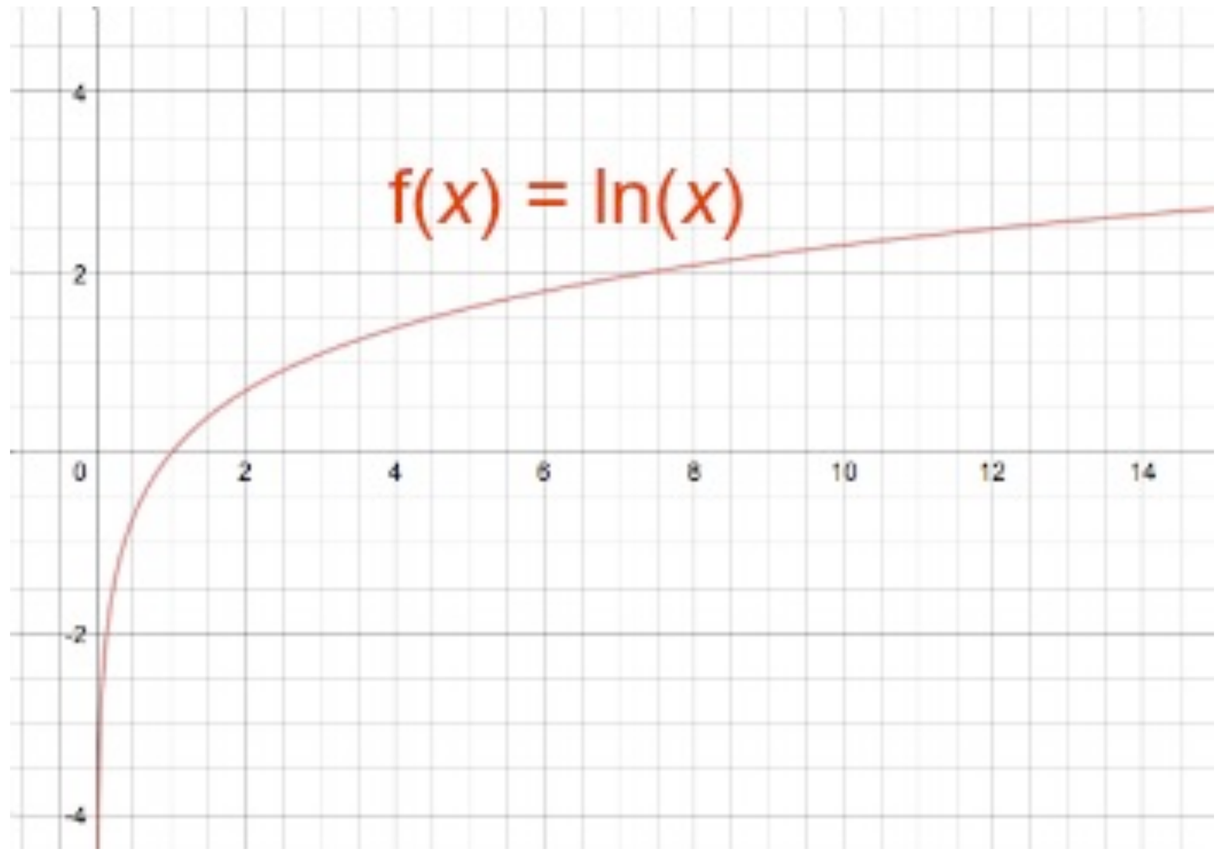
- Es el cociente entre los casos de éxito (categoría 0) sobre los de fracaso (categoría 1) en el suceso estudiado y para cada grupo.

- Este cociente puede tener un rango de valores $\dfrac{P}{1-P} \in [0,\infty]$.

- Sin embargo, todavía no coinciden los rangos en ambos lados de la ecuación

Regresión lineal a regresión logística

3ra parte



- Entonces, para que los rangos coincidan, podemos aplicar un logaritmo natural.

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta \cdot x$$

- De esta manera, la variable $\ln\left(\frac{P}{1-P}\right) \in [-\infty, \infty]$, los rangos coinciden en ambos lados de la ecuación.

- Finalmente, nuestro objetivo es obtener la probabilidad P de pertenencia a cierta categoría, por lo que podemos despejar y obtener la función sigmoide,

$$P = \frac{1}{1 + e^{-(\alpha+\beta\cdot x)}}$$

- Normalmente, una $P \geq 0.5$ se toma como categoría 1 $(y = 1)$ y $P < 0.5$ se toma como categoría 0 $(y = 0)$.

Regresión lineal a regresión logística

Interpretación de los coeficientes

- Como teníamos antes, en la regresión logística

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta \cdot x$$

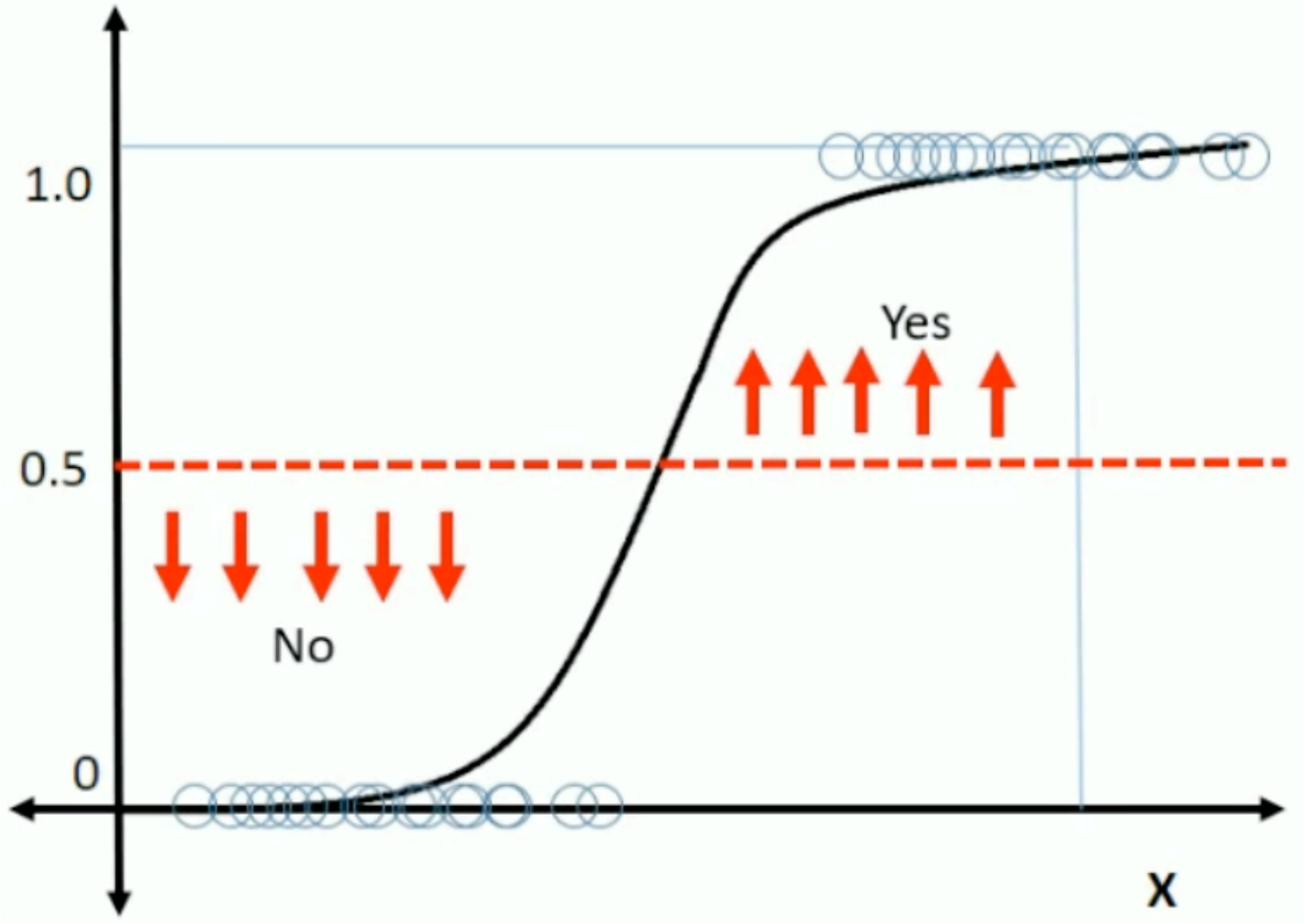
- Por lo que, los parámetros β no se interpretan directamente en términos de unidades de y, sino en términos de **log-odds.** En este contexto, odds se refiere al cociente de probabilidades dentro de la ecuación.

- Por lo tanto, la interpretación sería, si la variable x_i aumenta en una unidad, el log-odds ó $\ln\left(\frac{P}{1-P}\right)$ asociado a que el valor de y predicho sea 1, aumenta en β_1 unidades.

- Puesto de otra forma, si tuviéramos un modelo $\ln\left(\frac{P}{1-P}\right) = -2 + 0.7x_1$, y x_1 aumenta en una unidad, entonces, en términos de verosimilitud, éste aumenta $e^{0.7} \sim 2.01$ veces, o sea, se duplica por cada unidad extra de x_1.

Función Logit

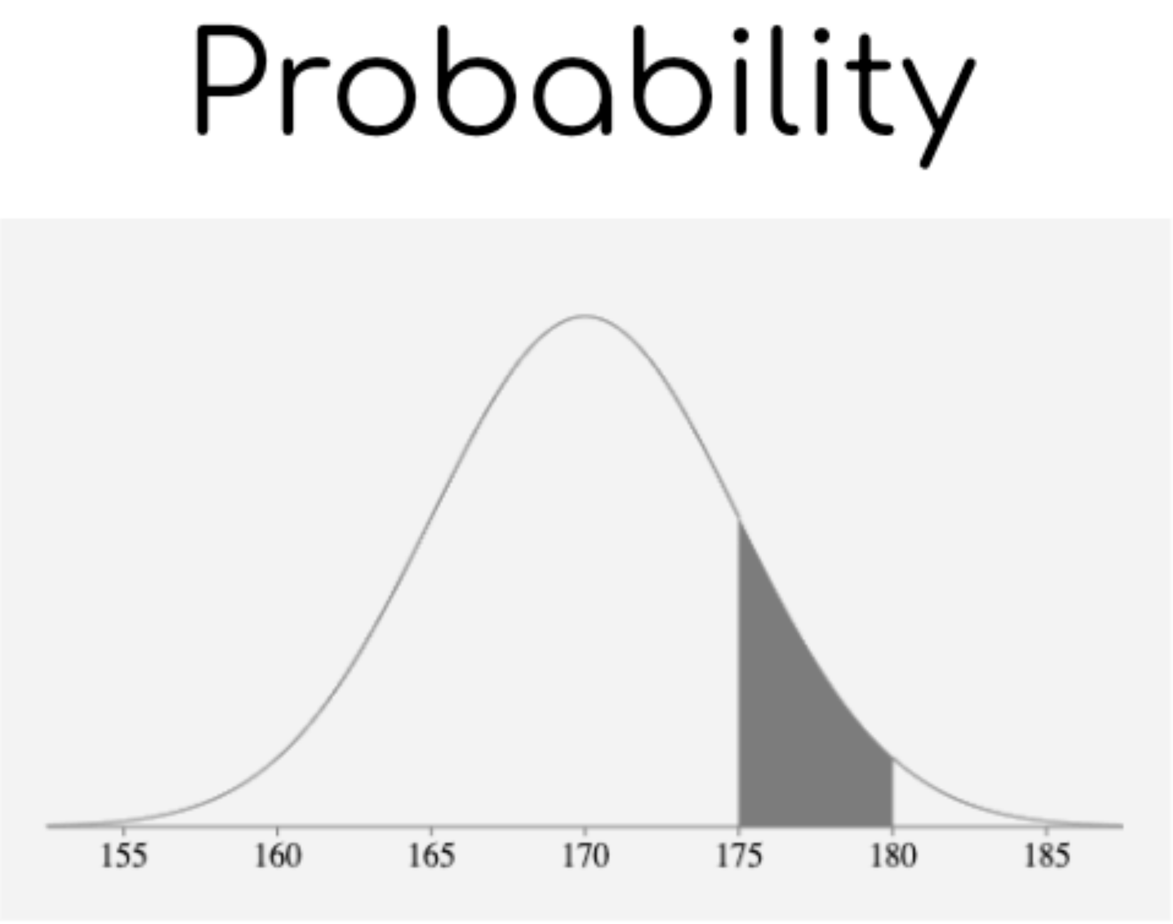
Umbral de decisión

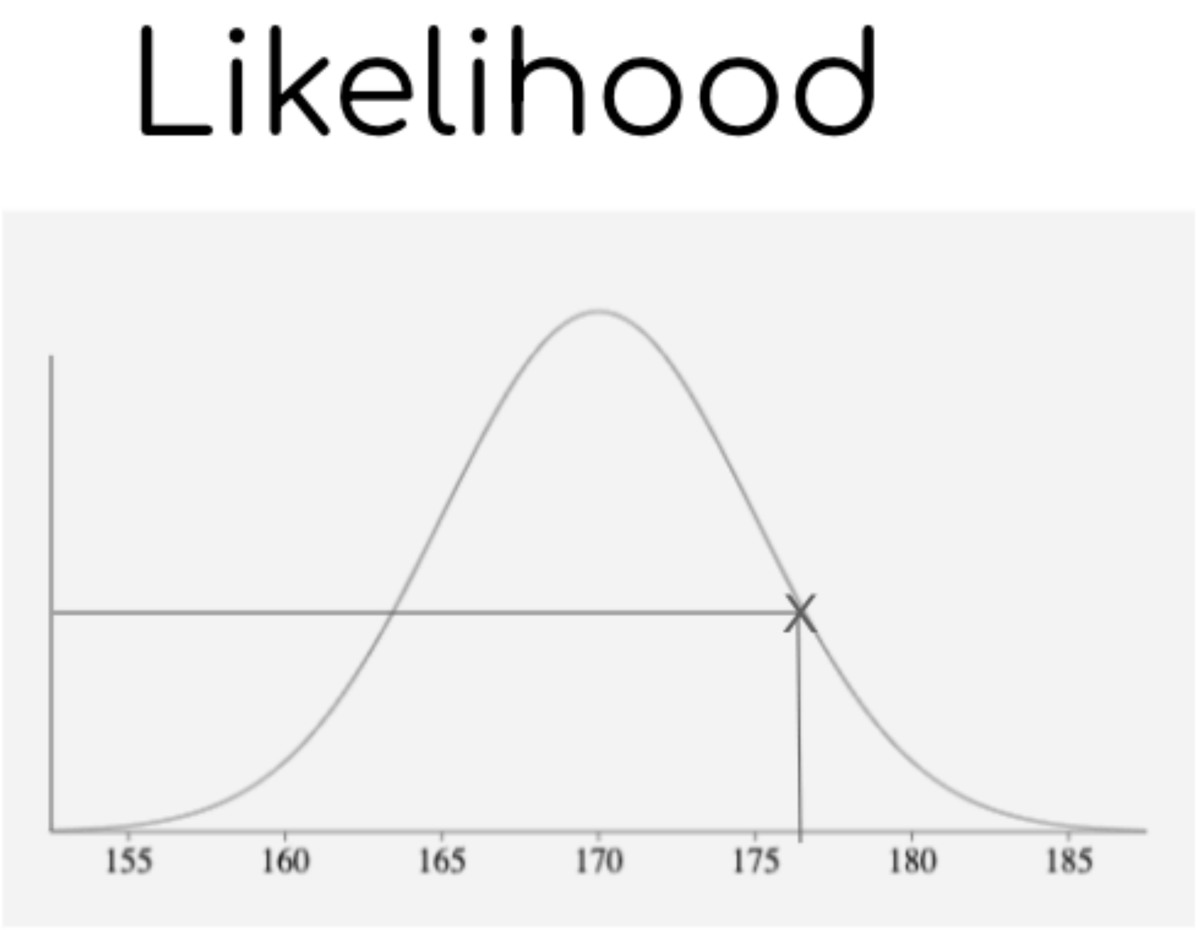


- Normalmente, una $P \geq 0.5$ se toma como categoría 1$(y = 1)$ y $P < 0.5$ se toma como categoría 0 $(y = 0)$.
- Sin embargo, esto no es fijo, se puede modificar según sea el caso.

Distribucion de probabilidad continua

Probabilidad vs Verosimilitud





- **Probabilidad**
- Los parámetros se suponen fijos.
- Los resultados son variables.
- Se considera el espacio de todos los posibles resultados (espacio muestral).
- La suma/integral de la probabilidad sobre todos los resultados posibles es 1.

- **Likelihood / Verosimilitud**
- Se permite que los parámetros que la generan varíen.
- Los resultados son fijos.
- Se considera el espacio de todos los posibles parámetros.
- La suma/integral sobre todos los posibles valores de los parámetros no es igual a 1.

Regresión logística

Método de máxima verosimilitud

- En la regresión logística, **no usamos** el MSE como en la regresión lineal como función de costo, porque estamos modelando probabilidades, no variables continuas.

- En su lugar, se usa el llamado **método de máxima verosimilitud** para encontrar los parámetros del modelo que **hacen más probable observar los datos que tenemos.**

- Supón que tienes un conjunto de datos $\{(x_i, y_i)\}_{i=1}^{m}$; $\quad x_i = \{x_{i1}, x_{i2}, \ldots, x_{ij}\}$

- El modelo quiere predecir una probabilidad con la función sigmoide $P = \hat{y}_i = \sigma(\beta \cdot x)$, evaluada sobre la ecuación de la recta que ya teníamos.

Regresión logística

Estimador de máxima verosimilitud

- Se define el entorno como la probabilidad conjunta de que se cumpla p_1 para cada una de las y's, que son 1's o son 0's.
- $L = \prod_{i=1}^{n} P_i^{y_i} \cdot (1 - P_i)^{1-y_i}$, donde y_i puede valer de cero a uno, y básicamente aparecerá uno de los dos términos. El primero para $y_i = 1$ y el segundo para $y_i = 0$.
- Este término es la **verosimilitud de observar todos los datos.**
- Se quiere encontrar la pertenencia o la no pertenencia de un conjunto a una sola categoría.
- En teoría, este término es el que quisiéramos optimizar. Pero hay un detalle ...

Regresión logística

Función de costo: entropía cruzada

- Pero hay cosas que no conocemos en el método, como lo parámetros α y β de la recta.
- $$L(\alpha,\beta) = \prod_{i=1}^{m} P_i^{y_i}(1-P_i)^{1-y_i}; \quad P_i = P_i(x_i)$$
- Todas las α y β están en productos dentro de la función, lo que complica demasiado su manejo. Sin embargo, aplicando un logaritmo natural, este se vuelve más fácil de trabajar.
- $$l = \ln L(\alpha,\beta) = \ln\left(\prod_{i=1}^{m} P_i^{y_i}(1-P_i)^{1-y_i}\right) = \sum_{i=1}^{m} y_i \ln P_i + (1-y_i)\ln\left(1-P_i\right)$$
- Esta función de costo que queremos optimizar es llamada **entropía cruzada.**

Regresión logística

Optimización: reescribir la función de costo

$$l = \ln L(\alpha, \beta) = \ln\left(\prod_{i=1}^{m} P_i^{y_i}(1-P_i)^{1-y_i}\right) = \sum_{i=1}^{m} y_i \ln P_i + (1-y_i)\ln\left(1-P_i\right)$$

$$= \sum_{i=1}^{m} \ln(1-P_i) + y_i(\ln P_i - \ln(1-P_i)) = \sum_{i=1}^{m} \ln(1-P_i) + y_i\left(\ln\left(\frac{P_i}{1-P_i}\right)\right)*$$

$$= \sum_{i=1}^{m}\left(\ln\left(1-P_i\right) + y_i\left(\alpha + \sum_{j=1}^{k} \beta_j x_{ij}\right)\right)$$

- Donde el término $(1 - P_i)$ se puede reescribir,

$$1 - P_i = 1 - \frac{1}{1 - e^{-(\alpha+\beta_i x_i)}} = \frac{e^{-(\alpha+\beta_i x_i)}}{1 + e^{-(\alpha+\beta_i x_i)}} = \frac{1}{1 + e^{(\alpha+\beta_i x_i)}}$$

- Entonces, su logaritmo natural, $\ln\left(1-P_i\right) = \ln\left(\frac{1}{1+e^{(\alpha+\beta_i x_i)}}\right) = \ln(1) - \ln\left(1 + e^{(\alpha+\beta_i x_i)}\right)$

Regresión logística

Optimización: reescribir la función de costo

- Finalmente tenemos la siguiente expresión para la función a optimizar, todo en términos de los parámetros de ajuste α y β:

$$l = \sum_{i=1}^{m} -\ln\left(1 + e^{\alpha + \beta x_i}\right) + y_i \left(\alpha + \sum_{j=1}^{k} \beta_j x_{ij}\right)$$

- Uno lo que hace aquí es nuevamente buscar los mínimos (derivadas igualadas a cero). Buscamos la máxima verosimilitud.
- $\dfrac{\partial l}{\partial \alpha} = \dfrac{\partial l}{\partial \beta_j} = 0$

Regresión logística

Optimización: primera derivada

- **Término α**

$$\frac{\partial l}{\partial \alpha} = \sum_{i=1}^{m} - \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} \cdot 1 + y_i \cdot 1 = \sum_{i=1}^{m} - P_i + y_i$$

- **Término β**

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{m} - \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} \cdot x_{ij} + y_i x_{ij} = \sum_{i=1}^{m} x_{ij}(y_i - P_i)$$

Regresión logística

Optimización: primera derivada

- Si definimos $\beta_0 = \alpha$ y que todos los puntos $x_{i0} = 1$ para $i = 1,2,\ldots,m$.
- Tenemos la expresión general
- $\dfrac{\partial l}{\partial \beta_j} = \sum_{i=1}^{m} x_{ij} \left(y_i - P_i \right)$
- Esto da una idea de como simplificar la función de máxima verosimilitud. En forma de producto escalar
- $\dfrac{\partial l}{\partial \beta_j} = x_j \cdot (y_i - P_i)$
- Incluso puedo hacer todas las derivadas de golpe y simplificar la expresión
- $\dfrac{\partial l}{\partial \beta} = X \cdot (Y_i - P_i)$

Regresión logística

Optimización: segunda derivada

- Al derivar por segunda vez,

$$\frac{\partial^2 l}{\partial \beta_k^2} = \sum_{i=1}^{m} - x_{ij} \frac{\partial P_i}{\partial \beta_k}$$

- Donde $P_i = \dfrac{1}{1 + e^{-(\alpha + \beta x_i)}}$

$$\frac{\partial P_i}{\partial \beta_j} = \frac{0 * (*) - 1 * e^{-(\alpha+\beta x_i)}}{\left(1 + e^{-(\alpha+\beta x_i)}\right)^2} x_{ij} = \frac{1}{1 + e^{-(\alpha+\beta x_i)}} \cdot \frac{e^{-(\alpha+\beta x_i)}}{1 + e^{-(\alpha+\beta x_i)}} x_{ij}$$

Regresión logística

Optimización: segunda derivada

- Por lo que finalmente,

$$\frac{\partial P_i}{\partial \beta_j} = P_i \cdot (1 - P_i) \cdot x_{ij}$$

- Entonces, la segunda derivada de l,

$$\frac{\partial^2 l}{\partial \beta_j^2} = \sum_{i=1}^{m} x_{ij} P_i (1 - P_i) x_{ij}$$

- Y en forma de vectores

$$\frac{\partial^2 l}{\partial \beta^2} = \sum_{i=1}^{m} x_i \cdot P_i (1 - P_i) x_i^t = X \cdot W \cdot X^t$$

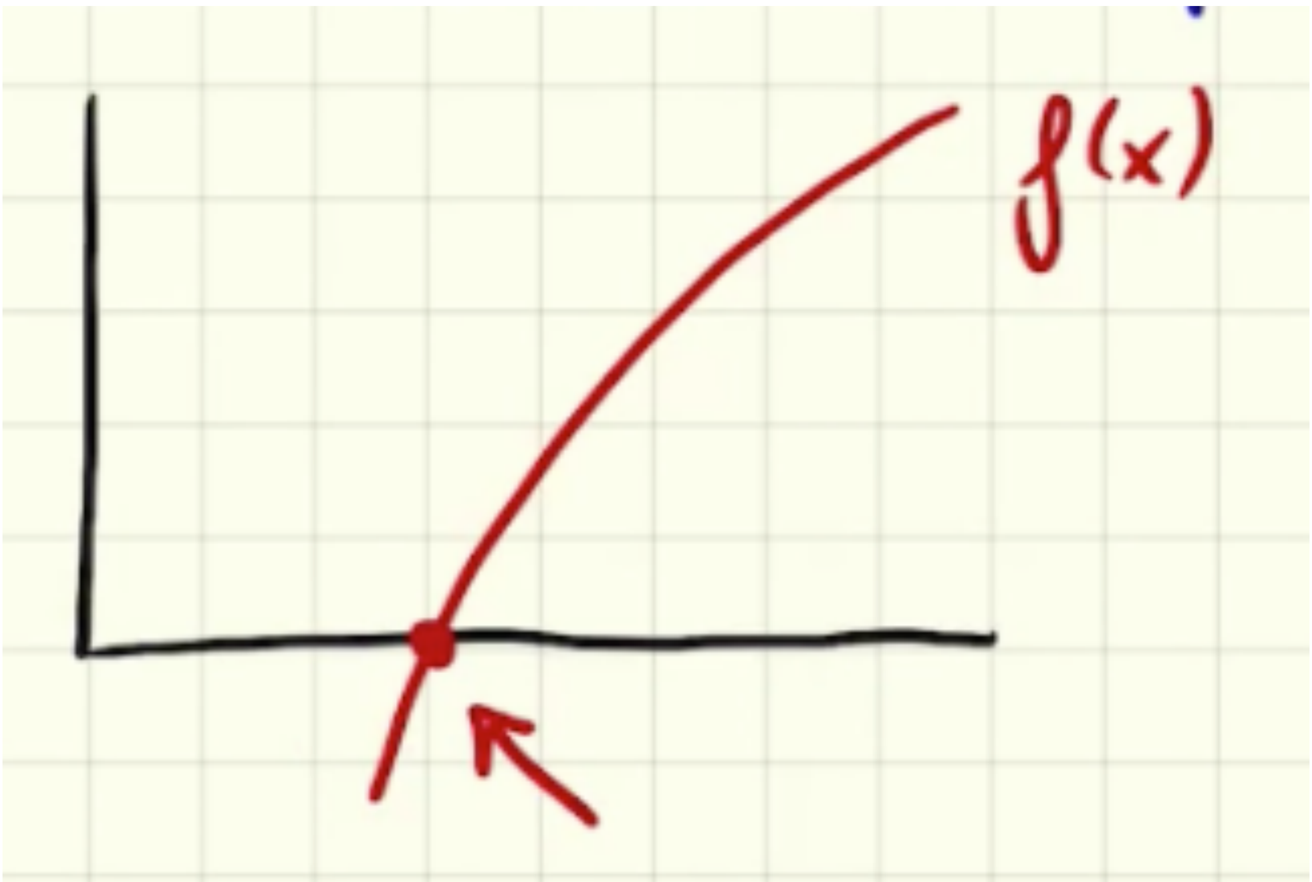
- Donde $W(\beta) = diag\left(P_i(1 - P_i)\right)_{i=1}^{m}$

Regresión logística

Optimización: minimización de las funciones derivadas

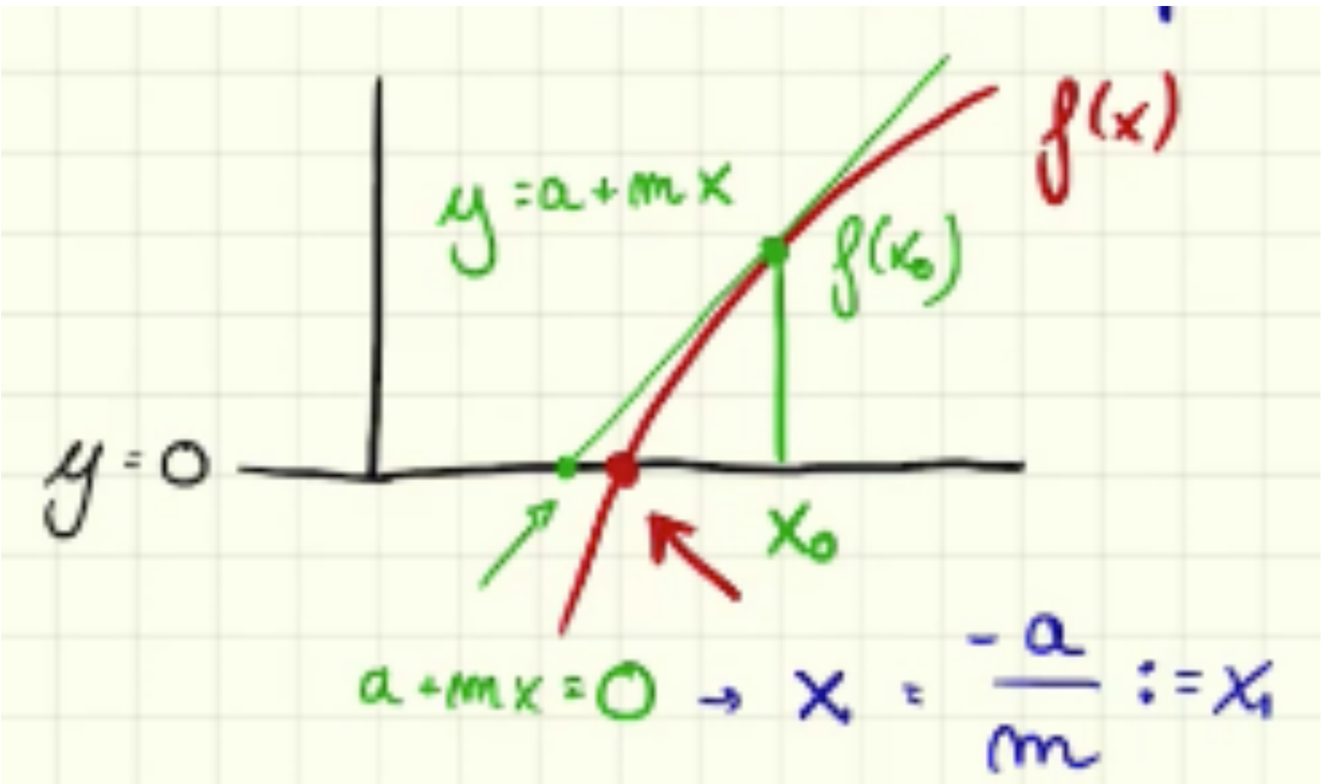
- En la primera derivada tendríamos que separar un término bastante complicado.

- Si la expresión no es analítica, si no tienen una expresión que sea relativamente sencilla de obtener, obtener esta intersección puede ser muy complicado.



Regresión logística

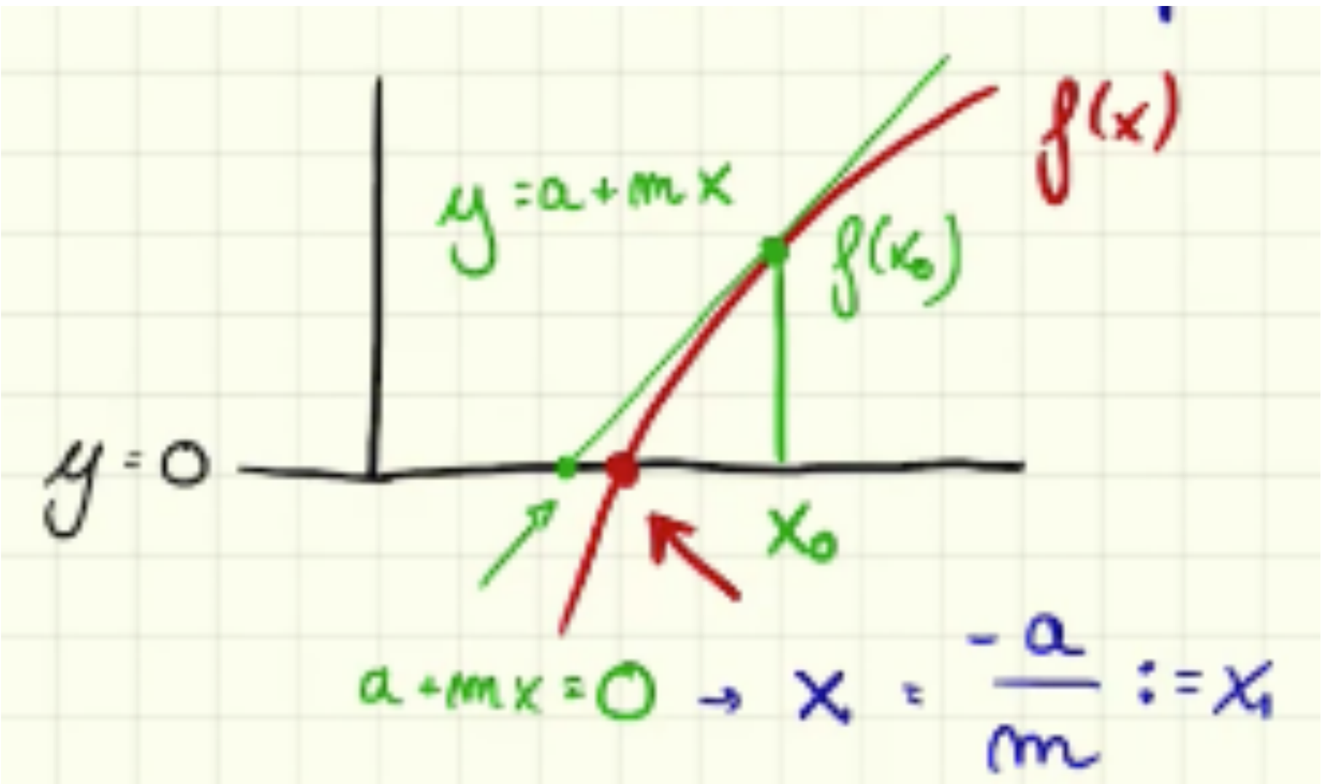
Optimización: método de Newton - Raphson



- Si decidimos empezar con un punto relativamente cerca del cero (ejemplo, todas las $\beta's$ como cero).
- A esa función sí le puedes calcular un valor, donde la derivada la puedes poner como su ecuación de recta $y = a + mx$ correspondiente.
- Después puede uno calcular la intersección de esta recta con el eje $y = 0$, con solución $x = -\dfrac{a}{m} = x_1$, que nos daría un punto x más cercano al cero que estamos buscando.
- Es un método aproximado.

Regresión logística

Optimización: método de Newton - Raphson



- Esta recta al punto x_0 ya lo conocemos, donde $m = f^{'}(x_0)$. En las formas que ya tenemos,

$$\beta_{n+1} = \beta_n - \frac{f(\beta_n)}{f^{'}(\beta_n)} \rightarrow \Delta\beta = -\frac{f(\beta)}{f^{'}(\beta)}$$

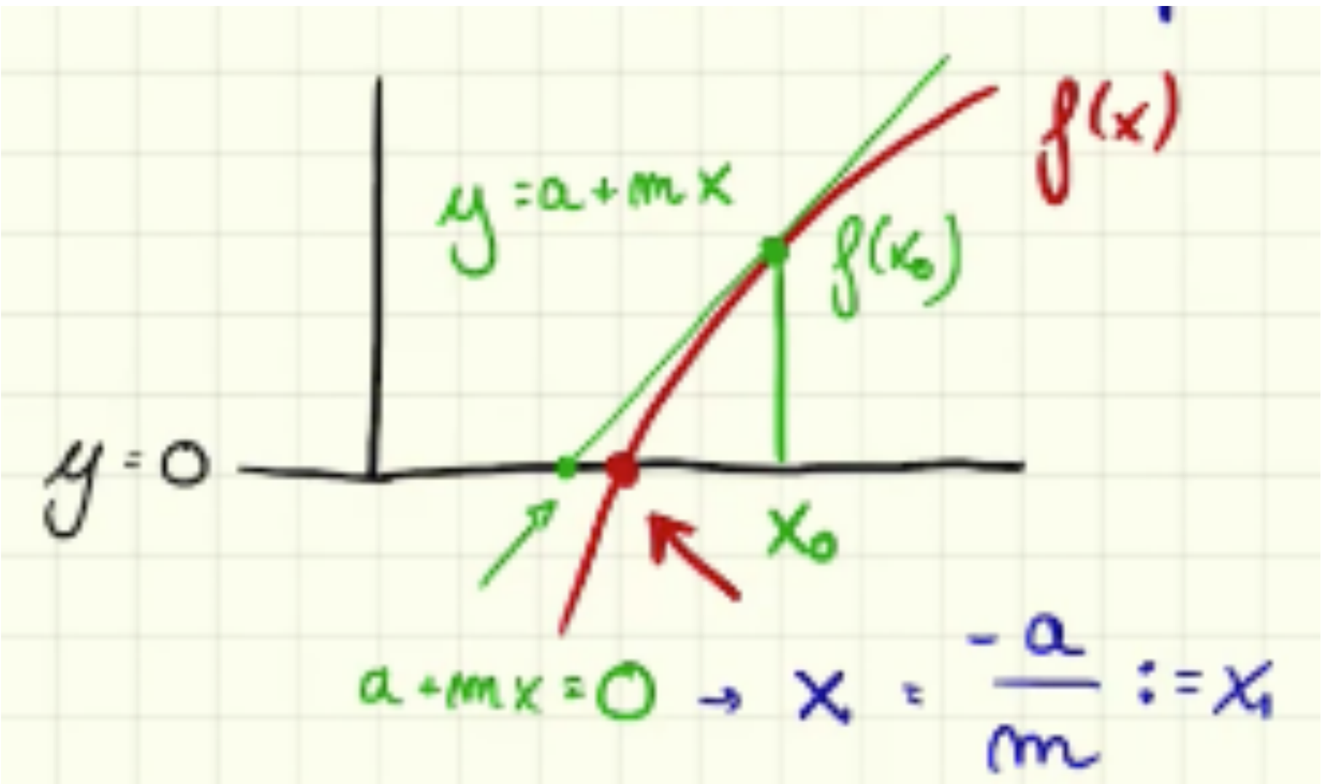
- La única función de β que tenemos es la P, con las derivadas

$$f(\beta) = \frac{\partial l}{\partial \beta} = X \cdot (Y - P(\beta))$$

$$f^{'}(\beta) = \frac{\partial^2 l}{\partial \beta^2} = X \cdot W(\beta) \cdot X^t$$

Regresión logística

Optimización: método de Newton - Raphson



- De tal forma que,
- $\Delta\beta = \left(XWX^t\right)^{-1} \cdot \left(X \cdot (Y - P)\right)$
- Y esto se hace de manera iterada, empezando con un vector $\overrightarrow{\beta} = 0$ hasta converger a $\overrightarrow{\beta}^*$, que serían los parámetros de la regresión logística.
- Existen otros algoritmos de optimización, como podría ser el **gradiente descendiente.**

Evaluación de un problema de clasificación

Matriz de confusión

- La **matriz de confusión** es una herramienta para evaluar el desempeño de un modelo de clasificación. Permite comparar las predicciones del modelo con los valores reales del conjunto de prueba.

- En clasificación binaria se organiza como en la siguiente tabla,

	Predicción: Positivo (1)	**Predicción: Negativo (0)**
Real: Positivo (1)	Verdadero Positivo (TP)	Falso Negativo (FN)
Real: Negativo 0)	Falso Positivo (FP)	Verdadero Negativo (TN)

Evaluación de un problema de clasificación

Métricas de evaluación (exactitud)

- La exactitud es una de las métricas básicas para evaluar problemas de clasificación, obtenida a partir de los valores de la matriz de confusión. Para un caso binario,
- $$\text{Accuracy} = \frac{\text{Numero de aciertos}}{\text{Total de observaciones}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

	Pred 1	**Pred 0**
Real 1	TP	FN
Real 0	FP	TN

Evaluación de un problema de clasificación

Métricas de evaluación

- Existen otro tipo de métricas de referencia, igualmente basadas en los valores de la matriz de confusión.
- **Precisión**: $$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$
- **Sensibilidad**: $$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
- **F1 score**: $$\text{F1} = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

	Pred 1	**Pred 0**
Real 1	TP	FN
Real 0	FP	TN

Evaluación de un problema de clasificación

Curva de ROC

- La **curva de ROC** (Receiver Operating Characteristic) es una herramienta gráfica para evaluar el rendimiento de un modelo de clasificación **binaria** al variar **el umbral de decisión.**

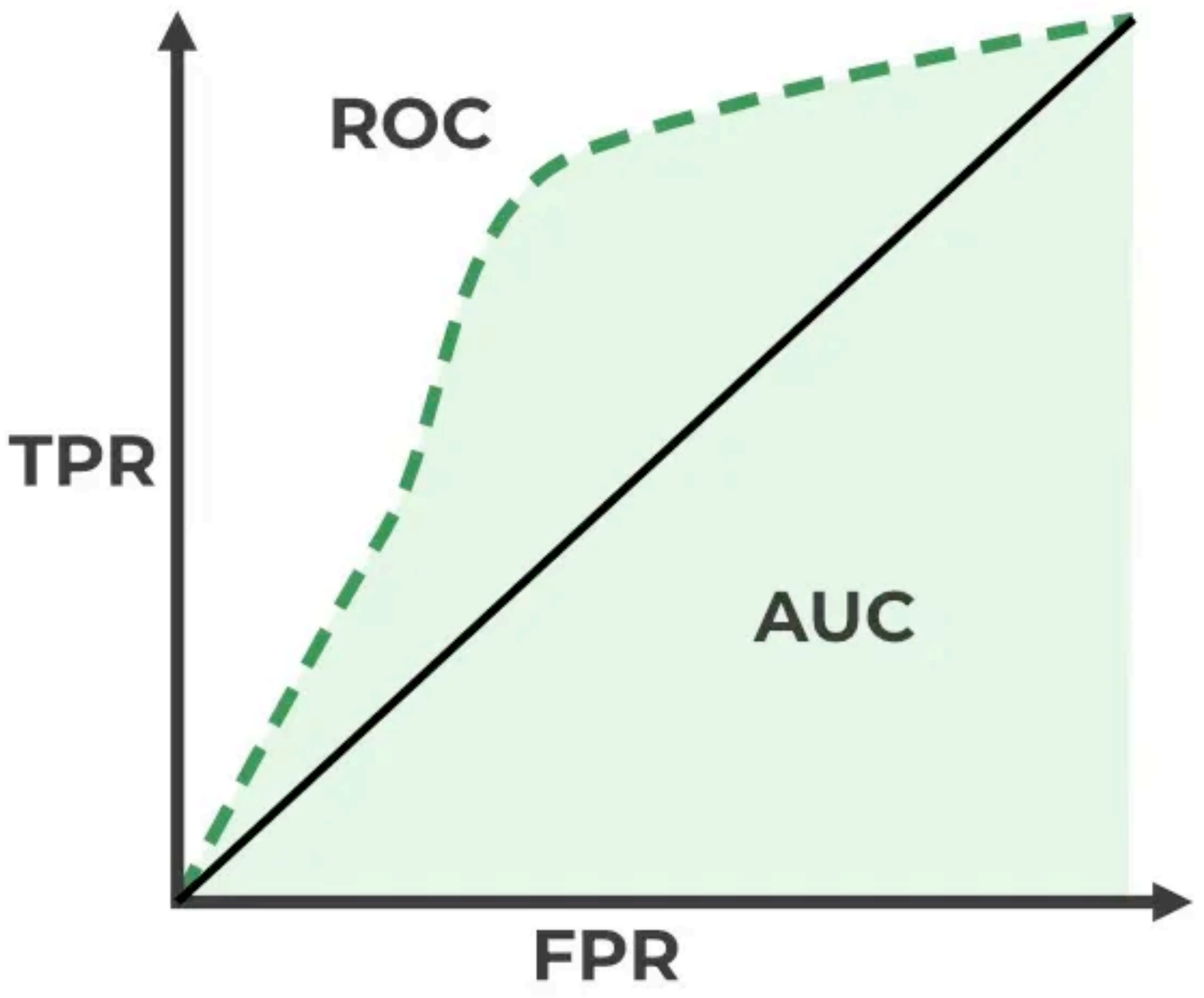
- En lugar de fijar un sólo umbral (e.g., default 0.5), la curva de ROC muestra cómo se comporta el modelo para todos los posibles umbrales de clasificación.

- En el eje X se tiene la tasa de falsos positivos $\mathrm{FPR} = \dfrac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$

- En el eje Y se tiene la tasa de verdaderos positivos (recall/sensibilidad) $\mathrm{TPR} = \dfrac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$.

Evaluación de un problema de clasificación

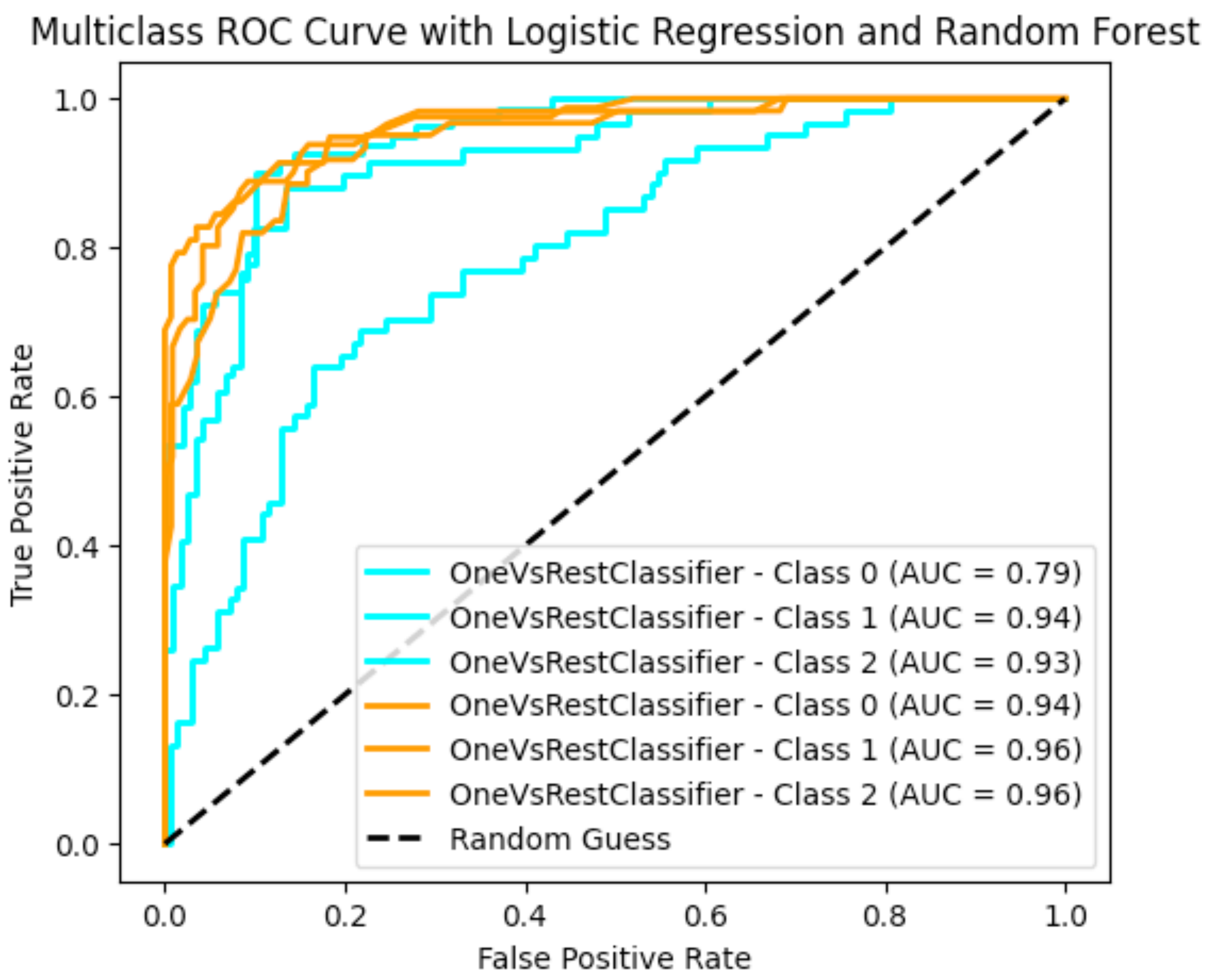
Curva de ROC



- Cada punto en la curva representa un par (FPR, TPR) para un determinado umbral.
- Umbral bajo -> casi todos clasificados como positivos -> TPR alto pero FPR también
- Umbral alto -> modelo muy estricto -> bajo FPR pero también bajo TPR

Evaluación de un problema de clasificación

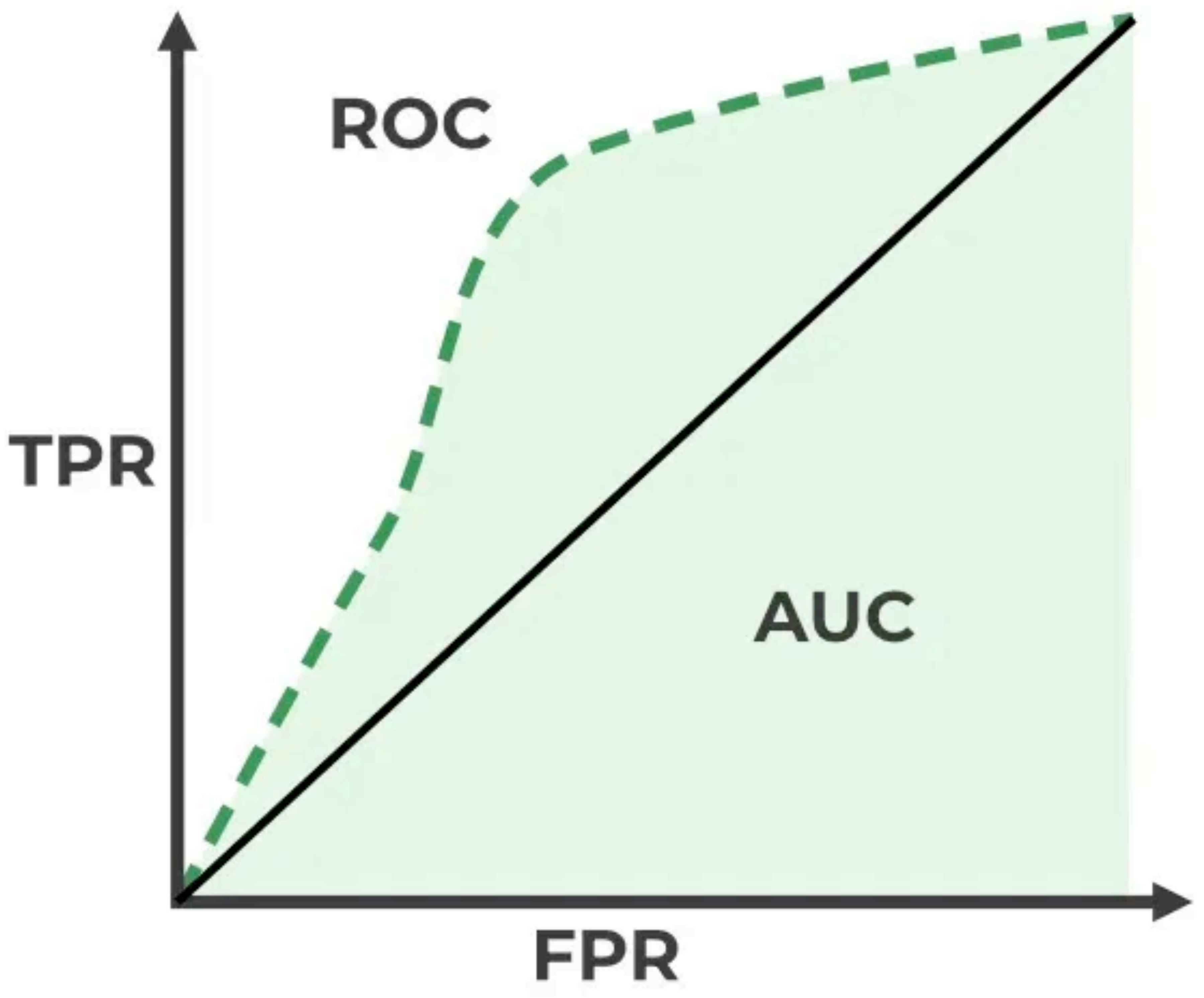
Curva de ROC



- Un modelo perfecto se acerca a la esquina superior izquierda.
- Una línea diagonal representa un modelo que adivina al azar.

Evaluación de un problema de clasificación

Curva de ROC



- AUC score - Area under the Curve. Es un número entre 0 y 1 representando el área bajo la curva de ROC.
    - 1.0 - Modelo perfecto
    - 0.5 - el modelo no tiene poder de clasificación (es como tirar una moneda básicamente)
    - <0.5 - peor que el azar (algo se hizo mal)