CLASIFICADORES

Inteligencia Artificial

CEIA - FIUBA

Dr. Ing. Facundo Adrián Lucianna

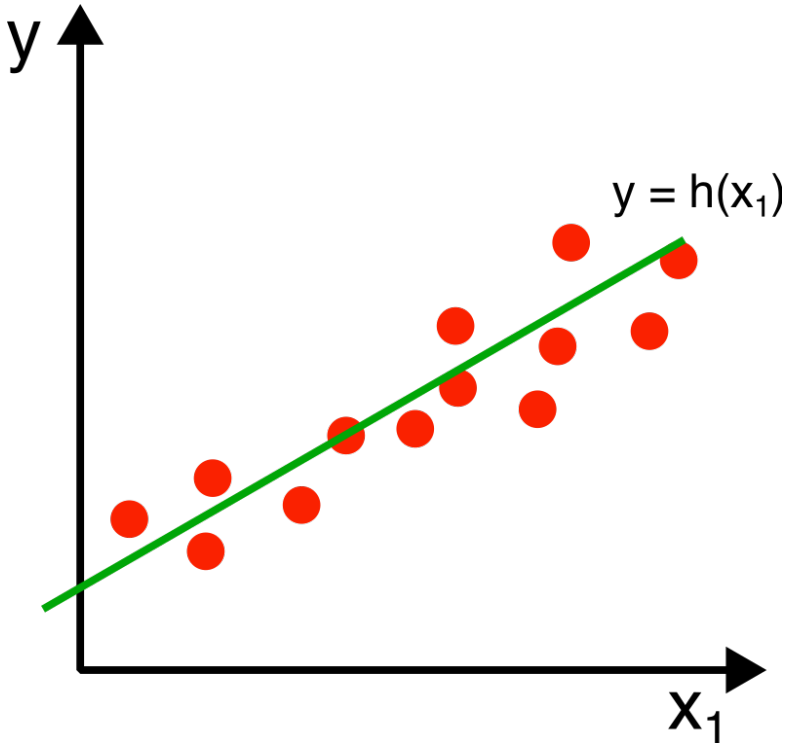
LO QUE VIMOS LA CLASE ANTERIOR…

REGRESIÓN

Si tenemos un problema donde el target **y** es una *variable numerica*, se llama un **problema de regresión**.

Se centra en estudiar las relaciones entre una variable dependiente de una o más variables independientes.

Es importante notar que, en Aprendizaje Automático, cuando buscamos una h(X) estamos armando un modelo puramente empírico. Es decir, nos basamos 100% en los datos medidos. En contraste con los modelos basados en propiedades fundamentales.



REGRESIÓN LINEAL

El modelo de regresión lineal más simple es el que involucra una combinación lineal de las variables de entradas:

$$\hat{y} = h(\boldsymbol{X}) = b + w_0 x_0 + \cdots + w_d x_d$$

- $X = (x_0, x_1, \ldots, x_d)$

  Son los *features* de nuestras observaciones. Son todas variables numéricas

- $b, w_0, \ldots, w_d$

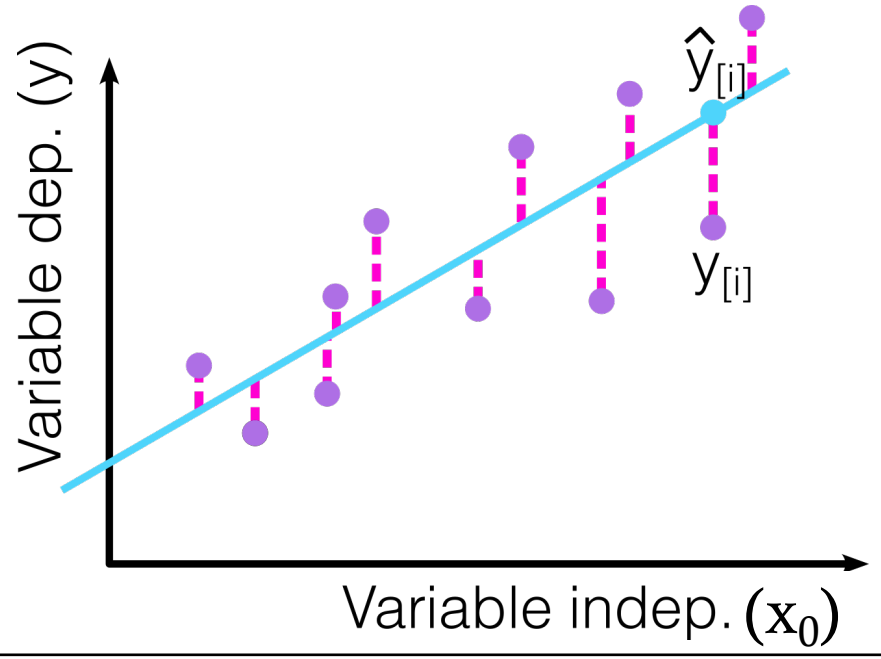
  Son los coeficientes del modelo. Son números reales. Cuanto más cerca de cero, la variable dependiente depende menos del *feature* que multiplica.

- $\hat{y}$

  Es la predicción del modelo. Es con quien comparamos con el *Label* de la observación

REGRESIÓN LINEAL

Buscamos minimizar el valor de los residuos. Para lograr esto, lo hacemos minimizando la suma de los cuadrados de los **residuos**.



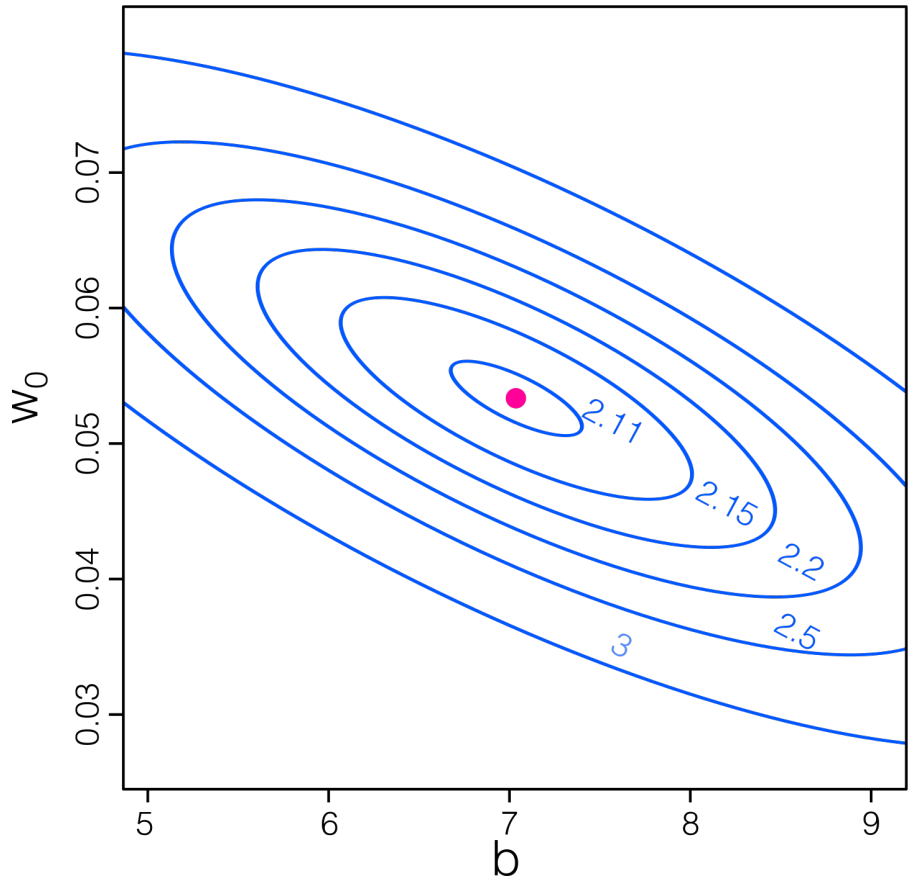
$$S_R = \sum_{i=0}^{N-1} \left(e_{[i]}\right)^2 = \sum_{i=0}^{N-1} \left(y_{[i]} - w_0 - w_1 x_{1[i]}\right)^2$$

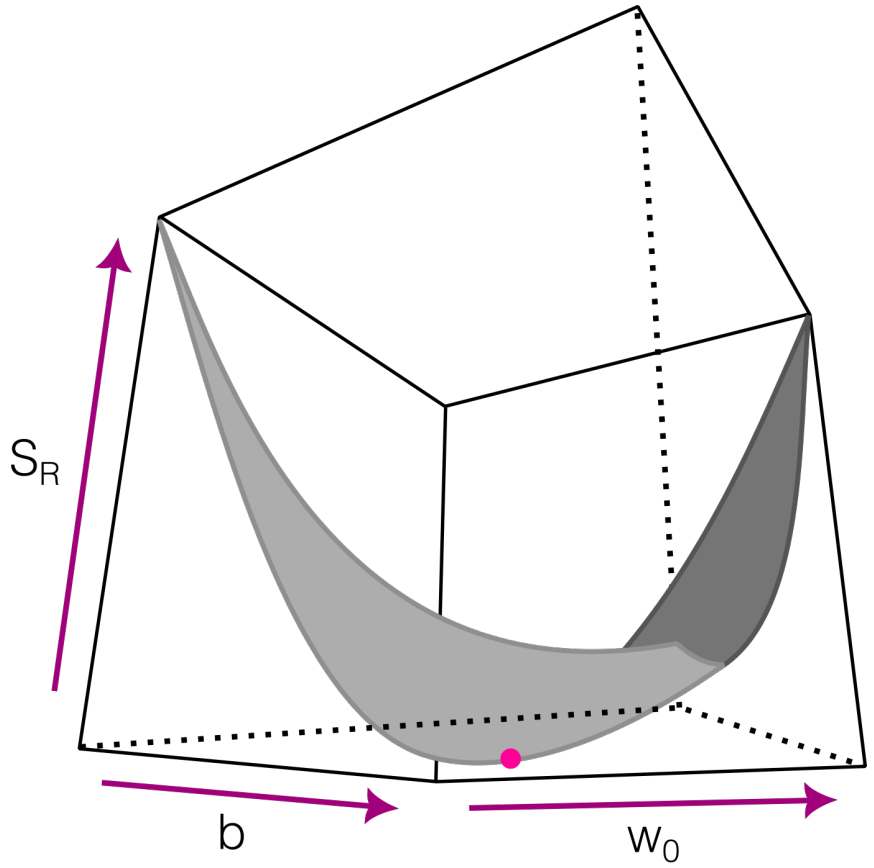
$$\min(S_R) = \min\left(\sum_{i=0}^{N-1} \left(e_{[i]}\right)^2\right)$$

Para minimizar, solo podemos tocar los coeficientes. Lo que hacemos es ir por el **gradiente**.

$$\frac{\partial S_R}{\partial w_0} = 0 \qquad \frac{\partial S_R}{\partial w_1} = 0$$

REGRESIÓN LINEAL





MÉTRICAS DE EVALUACIÓN

- El coeficiente de Pearson (R^2).
- Error absoluto medio: $MAE = \frac{1}{N}\sum_{i=0}^{N-1} |y_{[i]} - \hat{y}_{[i]}|$
- Error cuadrático medio: $MSE = \frac{1}{N}\sum_{i=0}^{N-1} \left(y_{[i]} - \hat{y}_{[i]}\right)^2$
- Error absoluto porcentual medio: $MAPE = \frac{100\%}{N}\sum_{i=0}^{N-1} \left|\frac{y_{[i]} - \hat{y}_{[i]}}{y_{[i]}}\right|$
- Error porcentual medio: $MPE = \frac{100\%}{N}\sum_{i=0}^{N-1} \frac{y_{[i]} - \hat{y}_{[i]}}{y_{[i]}}$

REGRESIÓN DE RIDGE Y LASSO

Regresión de Ridge:

$$\sum_{i=0}^{N-1} \left(y_{[i]} - b - W^T X_{[i]}\right)^2 + \alpha \sum_{j=0}^{d-1} w_j^2$$

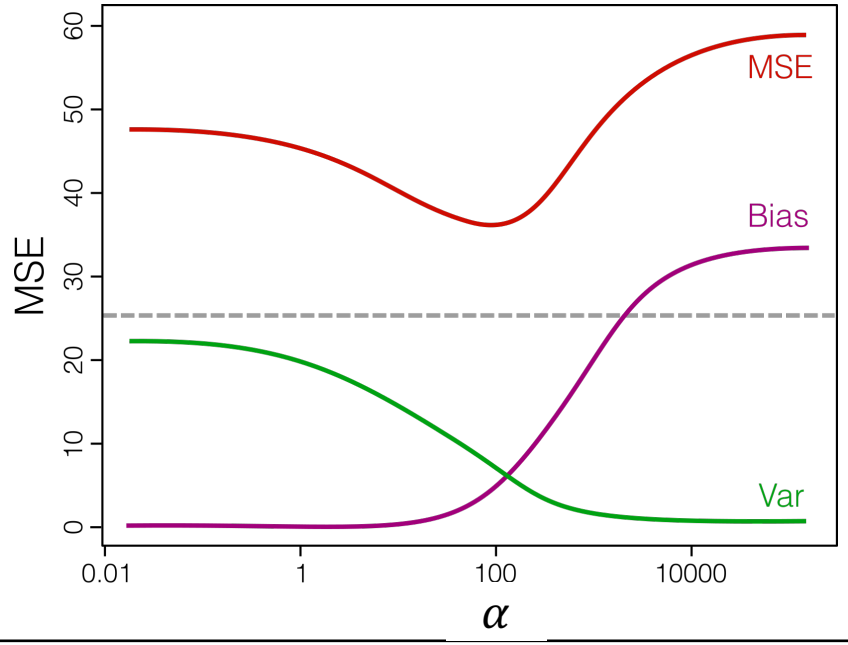
Regresión de Lasso:

$$\sum_{i=0}^{N-1} \left(y_{[i]} - b - W^T X_{[i]}\right)^2 + \alpha \sum_{j=0}^{d-1} |w_j|$$

REGRESIÓN DE RIDGE

¿Para qué nos sirve?

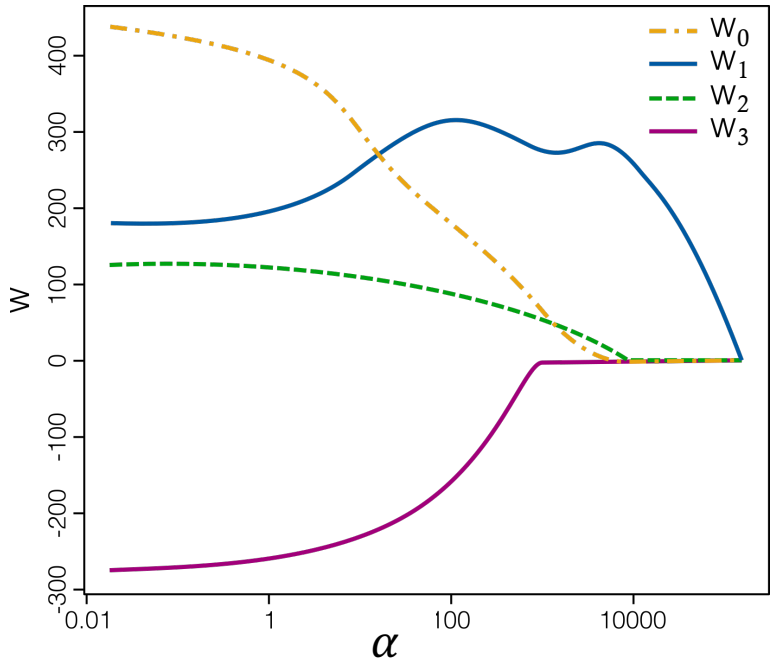
En general, cuando la verdadera relación es lineal, la regresión lineal tiene mucha varianza. Esto principalmente ocurre cuando el **número de observaciones es cercano al número de coeficientes**. En estos casos, la regresión de Ridge funciona mejor.



REGRESIÓN DE LASSO

¿Para qué nos sirve?

Esta regresión cuando α crece, algunos coeficientes se hacen exactamente cero. Por lo que Lasso realiza una selección de atributos.

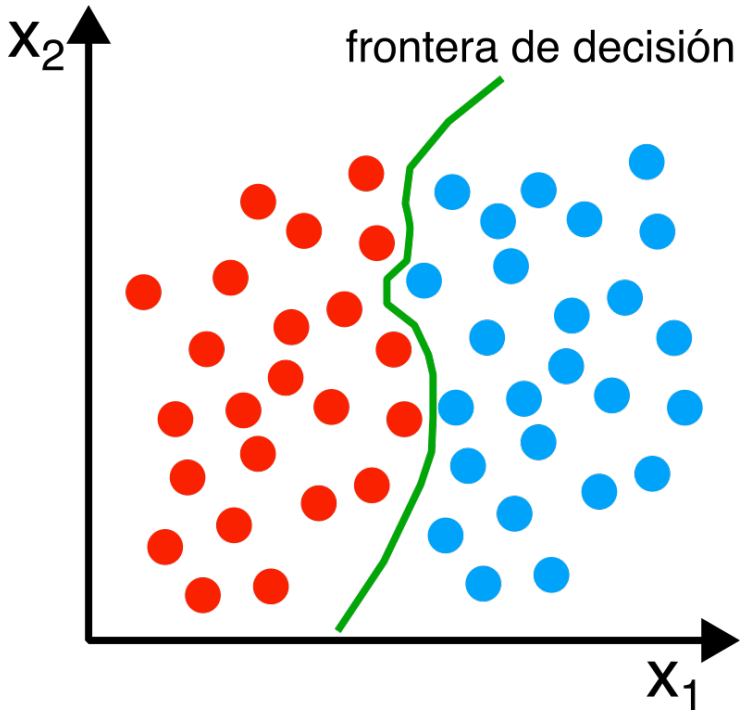


CLASIFICACIÓN

CLASIFICACIÓN

Es más común encontrarnos con problema de clasificación que de regresión:

- Una persona llega a una guardia con un set de síntomas atribuidos a una de tres condiciones médicas.
- Un servicio de banca online debe determinar si una transacción en el sitio es fraudulenta o no, usando como base la dirección IP, historia de transacciones, etc.
- En base a la secuencia de ADN de un número de pacientes con y sin una enfermedad dada, un genetista debe determinar que mutaciones de ADN genera un efecto nocivo relacionado a la enfermedad o no.



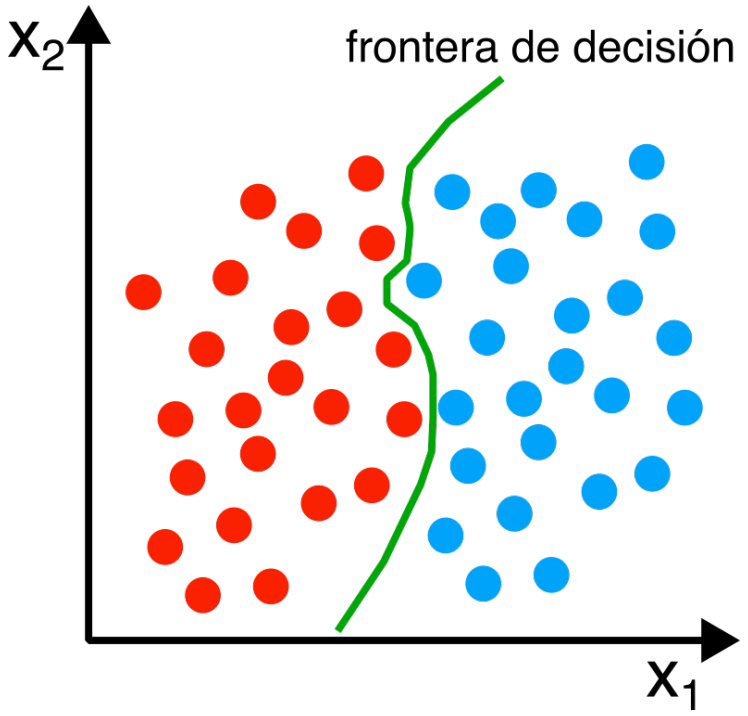
CLASIFICACIÓN

Regresión y clasificación son problemas muy similares entre sí. En ambos buscamos predecir una variable, la diferencia radica en que **regresión** predice una variable **numérica** y **clasificación** una **categórica**.

¿Por qué no usar regresión para predecir respuestas cualitativas?

Si usamos el ejemplo de los pacientes que llegan a la guardia, supongamos que hay tres diagnósticos:

- **ACV**
- **Sobredosis**
- **Ataques epilépticos**



CLASIFICACIÓN

Realizamos la siguiente codificación

- **ACV**: 1
- **Sobredosis**: 2
- **Ataques epilépticos**: 3

Aplicamos un modelo de regresión lineal para predecir en base a los predicadores del paciente.

El problema con esto es que la codificación implica un orden en los resultados, poniendo a **sobredosis** entre **ACV** y **ataques epilépticos**, y además que la distancia entre **ACV** y **sobredosis** es la misma que **sobredosis** y **ataques epilépticos**.

CLASIFICACIÓN

Pero tranquilamente podríamos haber elegido:

- **Ataques epilépticos**: 1
- **ACV**: 2
- **Sobredosis**: 3

Esto nos da una relación totalmente diferente.

Cada una de estas codificaciones produciría modelos lineales diferentes que, en última instancia, conducirían a diferentes conjuntos de predicciones sobre observaciones de prueba.

CLASIFICACIÓN

Si el target es una **variable categórica ordinal**, ahí el orden tiene sentido y está en un gris la elección de valores posibles modelos de clasificación y regresión.

Es más, un caso de respuesta booleanas, por ejemplo, si una persona tiene **ACV** (igual a 1) o no (igual a 0), podemos lograr mostrar que un modelo de regresión lineal es de hecho una estimación de la probabilidad de tener **ACV** dado un conjunto de entradas

$$P(ACV = 1|\boldsymbol{X}) = b + \boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{X}$$

REGRESIÓN LOGÍSTICA

REGRESIÓN LOGÍSTICA

Lo que buscamos modelar en regresión logística no es el label y, sino la probabilidad de que y pertenezca a una clase en particular.

$$P(y = k|\boldsymbol{X})$$

En una clasificación multiclase k puede ser 0, 1, 2, … (También podría ser cualquier cosa, “perro”, “gato”, “cebra”).

En el caso de clasificación de dos clases:

$$P(y = 0|\boldsymbol{X})$$
$$P(y = 1|\boldsymbol{X})$$

REGRESIÓN LOGÍSTICA

Pero, además, en el caso de dos clases:

$$P(y = 1|\boldsymbol{X}) = 1 - P(y = 0|\boldsymbol{X})$$

Por lo que podemos simplificar la notación…

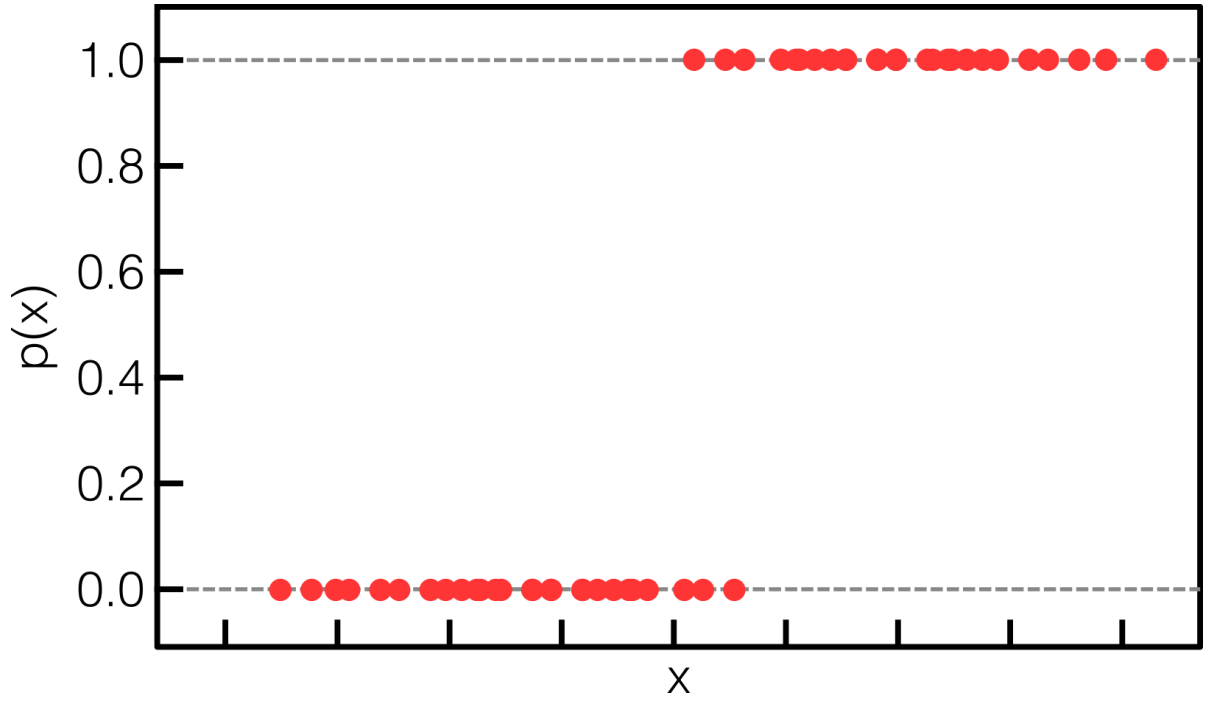
$$P(y = 1|\boldsymbol{X}) = p(\boldsymbol{X})$$

Las probabilidades son valores que van entre 1 y 0.

Además, la hagamos más simple, el caso de un solo atributo: $p(x)$

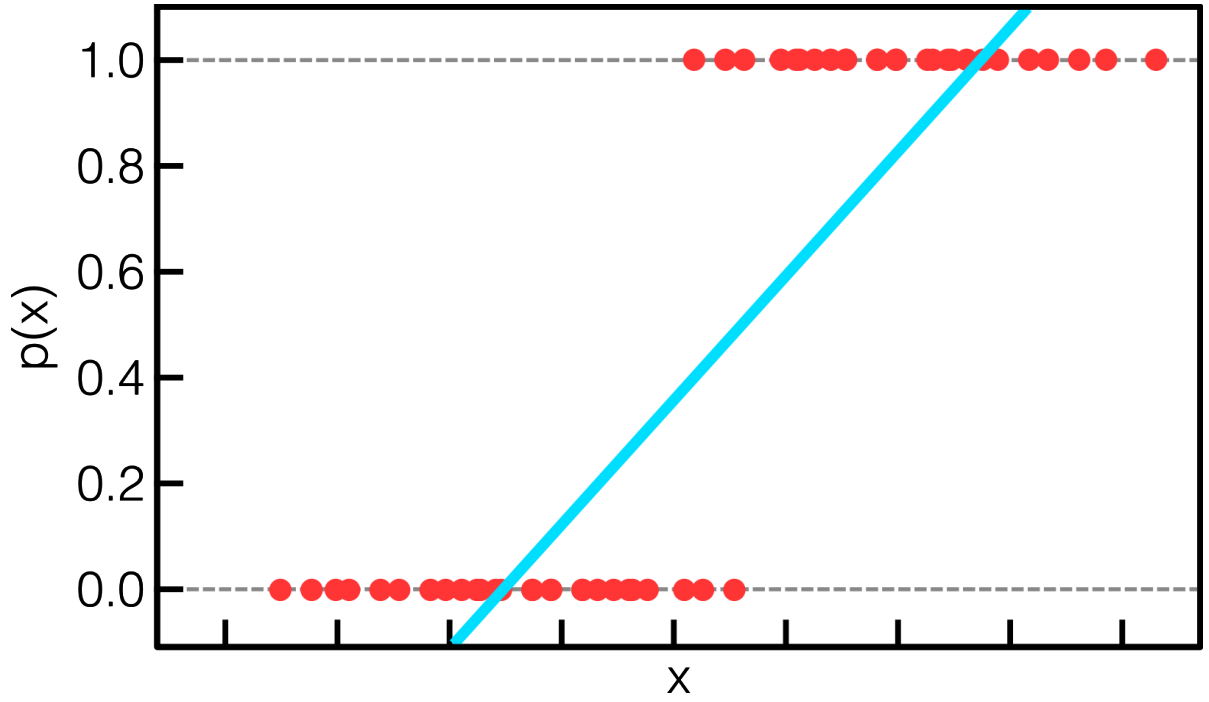
REGRESIÓN LOGÍSTICA

En un dataset, ya tenemos la probabilidad a la que pertenece. Es 1 en la clase que le pertenece.



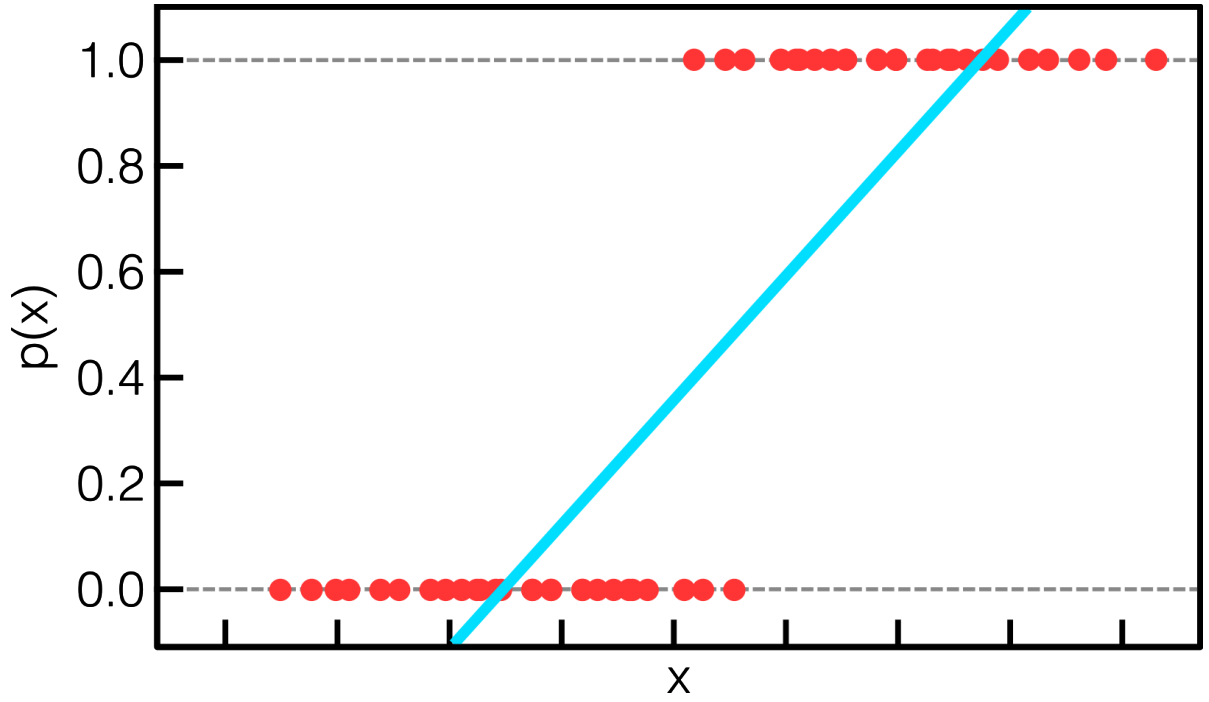
REGRESIÓN LOGÍSTICA

Podemos usar una regresión lineal para estimar la probabilidad $p(x) = b + w_0 x$



REGRESIÓN LOGÍSTICA

En la gráfica se observa el problema de predecir usando **regresión lineal**. Dada la naturaleza de la función, hay valores en donde se obtienen $p(x) < 0$, o $p(x) > 1$. Esto va a ocurrir con cualquier regresión que de valores por fuera a 0 y 1.



REGRESIÓN LOGÍSTICA

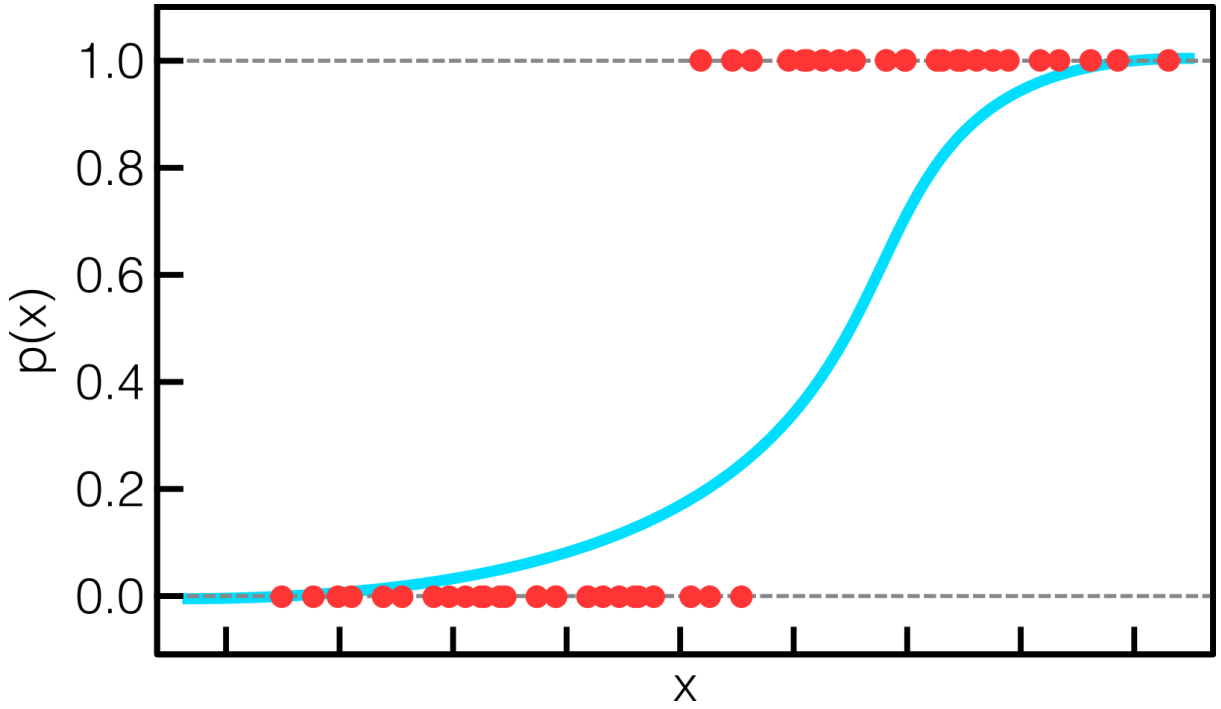
Para evitar esto, podemos modelar a la probabilidad usando una función que nos asegure que siempre tendremos valores entre 0 y 1.

En regresión logística, esto lo resolvemos usando una función sigmoide:

$$p(x) = \frac{e^{b+w_0 x}}{1 + e^{b+w_0 x}} = \frac{1}{1 + e^{-(b+w_0 x)}}$$

REGRESIÓN LOGÍSTICA

Lo que visualmente se observa:



REGRESIÓN LOGÍSTICA

Esta regresión siempre va a formar una curva con forma sigmoidea. E independientemente del valor de x, siempre estará contenido entre 0 y 1.

Si manipulamos a $p(x) = \frac{e^{b+w_0 x}}{1+e^{b+w_0 x}}$, llegamos a:

$$\frac{p(x)}{1-p(x)} = e^{b+w_0 x}$$

El cuál es la **chance** (o en ingles **odds**), es la proporción entre dos probabilidades complementarias. Estos valores pueden tomar desde 0 a infinito.

Para entender, en una semana la probabilidad de ser sábado es $1/7$, pero la chance es $1/6$, es decir 6 a 1 de que no sea sábado.

REGRESIÓN LOGÍSTICA

Si aplicamos el logaritmo de ambos lados:

$$logit(p) = \ ln\left(\frac{p(x)}{1-p(x)}\right) = b + w_0 x$$

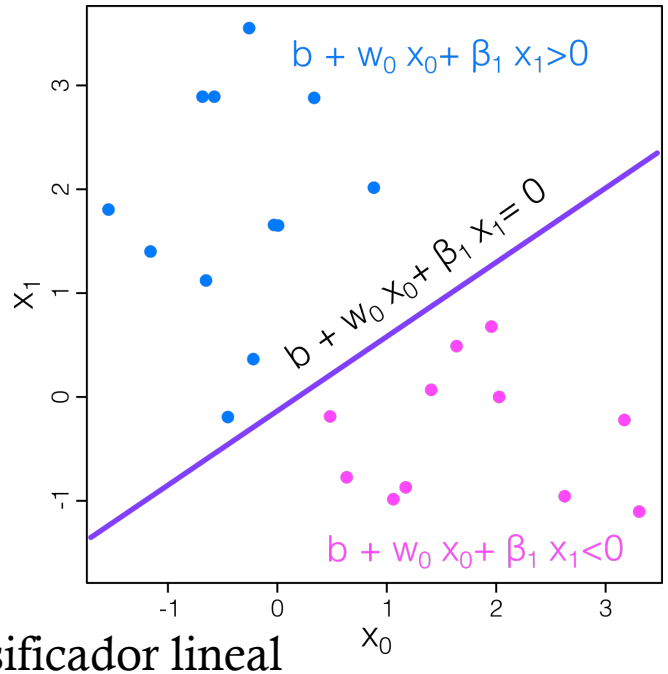
Obtenemos la función **logit**. *La función logit se utiliza para transformar variables de entrada en un rango que puede interpretarse como probabilidades. En la regresión logística es una relación lineal.*

Esto nos permite que ver qué incremento de una unidad de x, cambia el **logit** en w_0 unidades. Equivalentemente multiplica la chance en e^{w0}.

Pero, la cantidad de p(x) al aumentar x en una unidad, al no ser lineal, depende del valor actual de x.

REGRESIÓN LOGÍSTICA

Dado que el **logit** es una función lineal, si observamos la frontera de clasificación para un caso con dos atributos de entrada:



Que es lo que se conoce como un clasificador lineal

REGRESIÓN LOGÍSTICA - AJUSTE

Para buscar los coeficientes (b y w_0), es decir entrenar, lo hacemos realizándolo por **máxima verosimilitud**.

La intuición básica detrás de la máxima verosimilitud es que buscamos estimaciones para b y w_0 tales que la probabilidad prevista $p(x_i)$ de todos los valores del dataset, utilizando $p(x) = \frac{e^{b+w_0 x}}{1+e^{b+w_0 x}}$ corresponda lo más cerca posible al estado observado.

En otras palabras, tratamos de encontrar b y w_0 tales que al encontrar estas estimaciones se obtenga un número cercano a uno para la clase positiva, y lo más cercano a 0 para la clase negativa

REGRESIÓN LOGÍSTICA - AJUSTE

Matemáticamente la función de verosimilitud es:

$$l(b, w_0) = \prod_{i=1}^{N} p(x_i)^{y_i} \ (1 - p(x_i))^{1-y_i}$$

Similar a la regresión lineal, es mejor minimizar la función log-verosimilitud multiplicada por -1:

$$J(b, w_0) = -\sum_{i=1}^{N} y_i \ln(p(x_i)) + (1 - y_i)\ln(1 - \ p(x_i))$$

REGRESIÓN LOGÍSTICA - AJUSTE

$$J(b, w_0) = -\sum_{i=1}^{N} \ln\big(1 - p(x_i)\big) - \sum_{i=1}^{N} y_i \, ln\left(\frac{p(x_i)}{1 - p(x_i)}\right)$$

$$J(b, w_0) = -\sum_{i=1}^{N} \ln\big(1 - p(x_i)\big) - \sum_{i=1}^{N} y_i (b + w_0 x_i)$$

$$J(b, w_0) = -\sum_{i=1}^{N} \ln\left(1 - \frac{e^{b+w_0 x_i}}{1 + e^{b+w_0 x_i}}\right) - \sum_{i=1}^{N} y_i (b + w_0 x_i)$$

REGRESIÓN LOGÍSTICA - AJUSTE

$$J(b, w_0) = -\sum_{i=1}^{N} \ln\left(1 - \frac{e^{b+w_0 x_i}}{1 + e^{b+w_0 x_i}}\right) - \sum_{i=1}^{N} y_i (b + w_0 x_i)$$

$$J(b, w_0) = \sum_{i=1}^{N} \ln\left(1 + e^{b+w_0 x_i}\right) - \sum_{i=1}^{N} y_i (b + w_0 x_i)$$

$$J(b, w_0) = \sum_{i=1}^{N} \ln\left(1 + e^{b+w_0 x_i}\right) - y_i b - y_i w_0 x_i$$

REGRESIÓN LOGÍSTICA - AJUSTE

Para encontrar el mínimo, aplicamos el gradiente:

$$\frac{\partial J}{\partial b} = \sum_{i=1}^{N} \frac{e^{b+w_0 x_i}}{1+e^{b+w_0 x_i}} - y_i = 0$$

$$\frac{\partial J}{\partial w_0} = \sum_{i=1}^{N} \left(\frac{e^{b+w_0 x_i}}{1+e^{b+w_0 x_i}} - y_i \right) x_i = 0$$

Resolver esto se necesita de aplicar métodos numéricos o usar gradiente descendiente. *Scikit-learn implementa varios métodos.*

REGRESIÓN LOGÍSTICA MÚLTIPLE

Igual que la regresión lineal, podemos tener más de una variable:

$$logit(p) = ln\left(\frac{p(\boldsymbol{X})}{1-p(\boldsymbol{X})}\right) = b + \boldsymbol{W}^T\boldsymbol{X} \qquad p(\boldsymbol{X}) = \frac{e^{b+\boldsymbol{W}^T\boldsymbol{X}}}{1+e^{b+\boldsymbol{W}^T\boldsymbol{X}}}$$

En este caso, el gradiente queda:

$$\frac{\partial J}{\partial b} = \sum_{i=1}^{N} \frac{e^{b+\boldsymbol{W}^T\boldsymbol{X}}}{1+e^{b+\boldsymbol{W}^T\boldsymbol{X}}} - y_i = 0$$

$$\frac{\partial J}{\partial w_j} = \sum_{i=1}^{N} \left(\frac{e^{b+\boldsymbol{W}^T\boldsymbol{X}}}{1+e^{b+\boldsymbol{W}^T\boldsymbol{X}}} - y_i\right) x_{ij} = 0$$

CURVA ROC

CURVA ROC

En la clase que introducimos a Aprendizaje Automático, vimos varias métricas de clasificadores binarios.

Ahora, siempre supusimos que nuestro clasificador nos da la salida 1 si es la clase positiva, 0 si es negativa, pero ahora tenemos un clasificador que nos da una probabilidad de que tan probable es que sea de la clase positiva.

De forma intuitiva, podemos definir que si la regresión logística nos devuelve un valor a mayor a 0.5, definimos como clase positiva, sino la negativa. De ahí podemos calcular **exactitud**, **precisión**, etc.

CURVA ROC

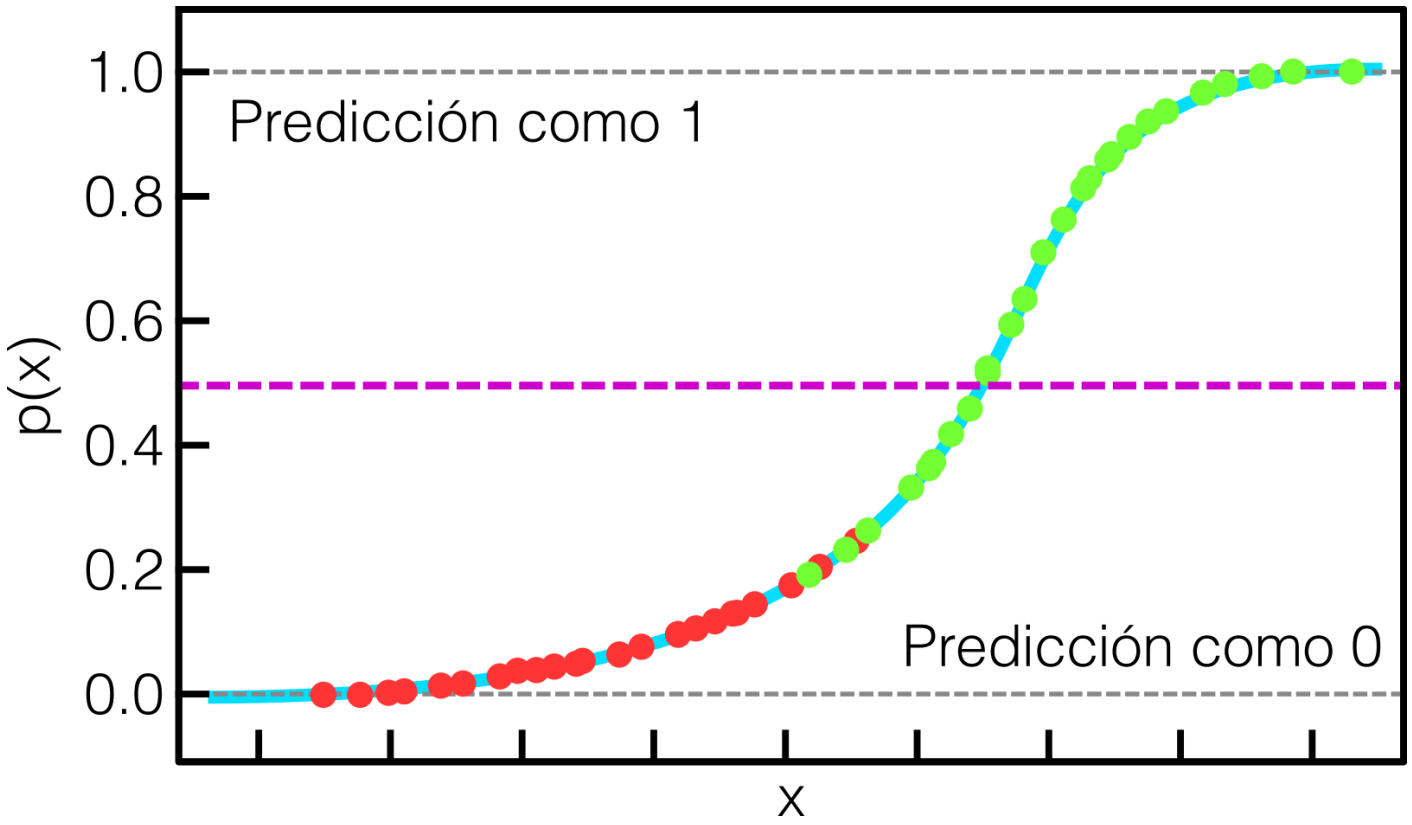
¿Pero por qué este valor?

CURVA ROC

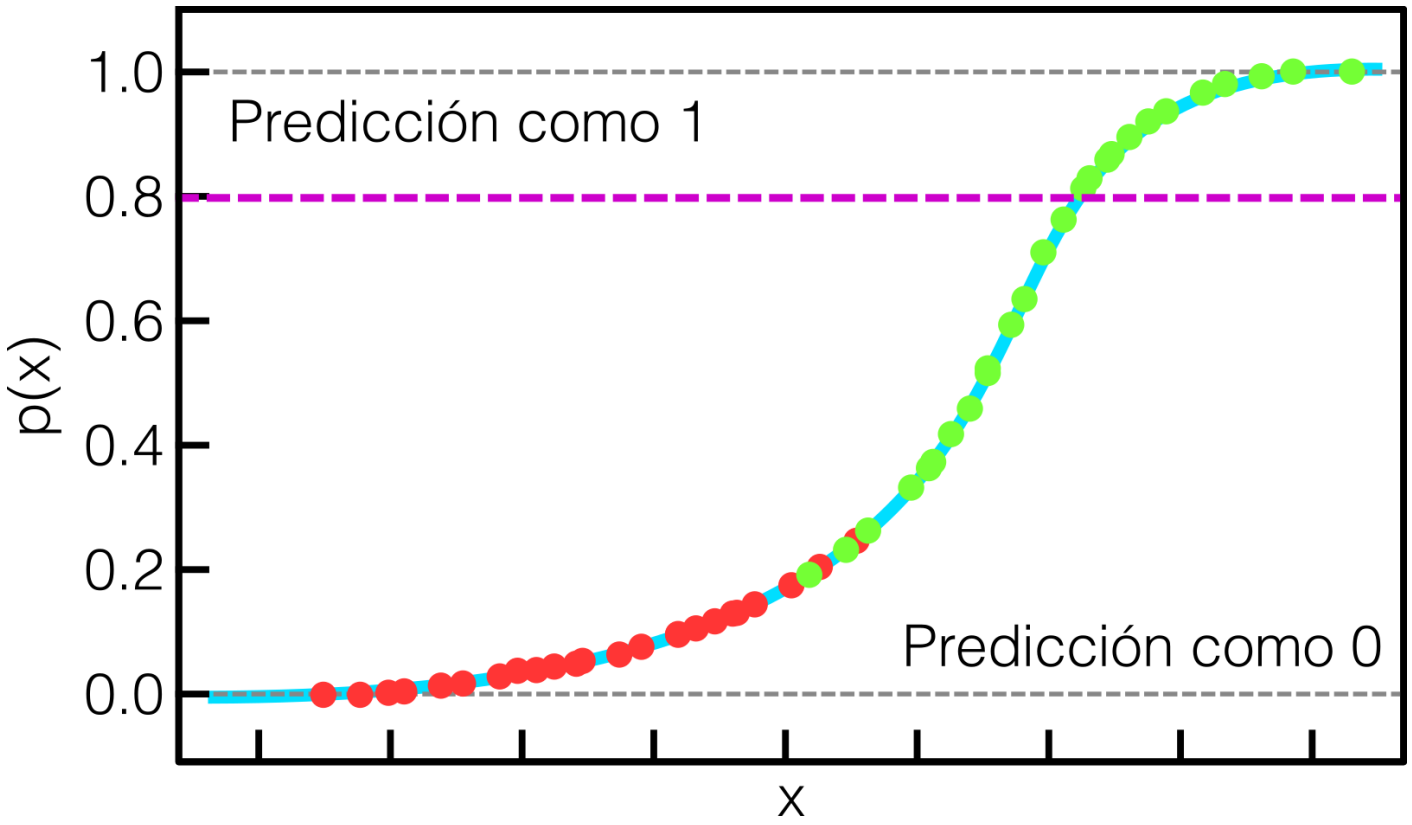
¿Pero por qué este valor?

Nos basamos en la idea de que el modelo nos da un valor de probabilidad. Pero nada impide de que el umbral pueda ser definido en diferentes valores, sobre todo si las clases están desbalanceadas.

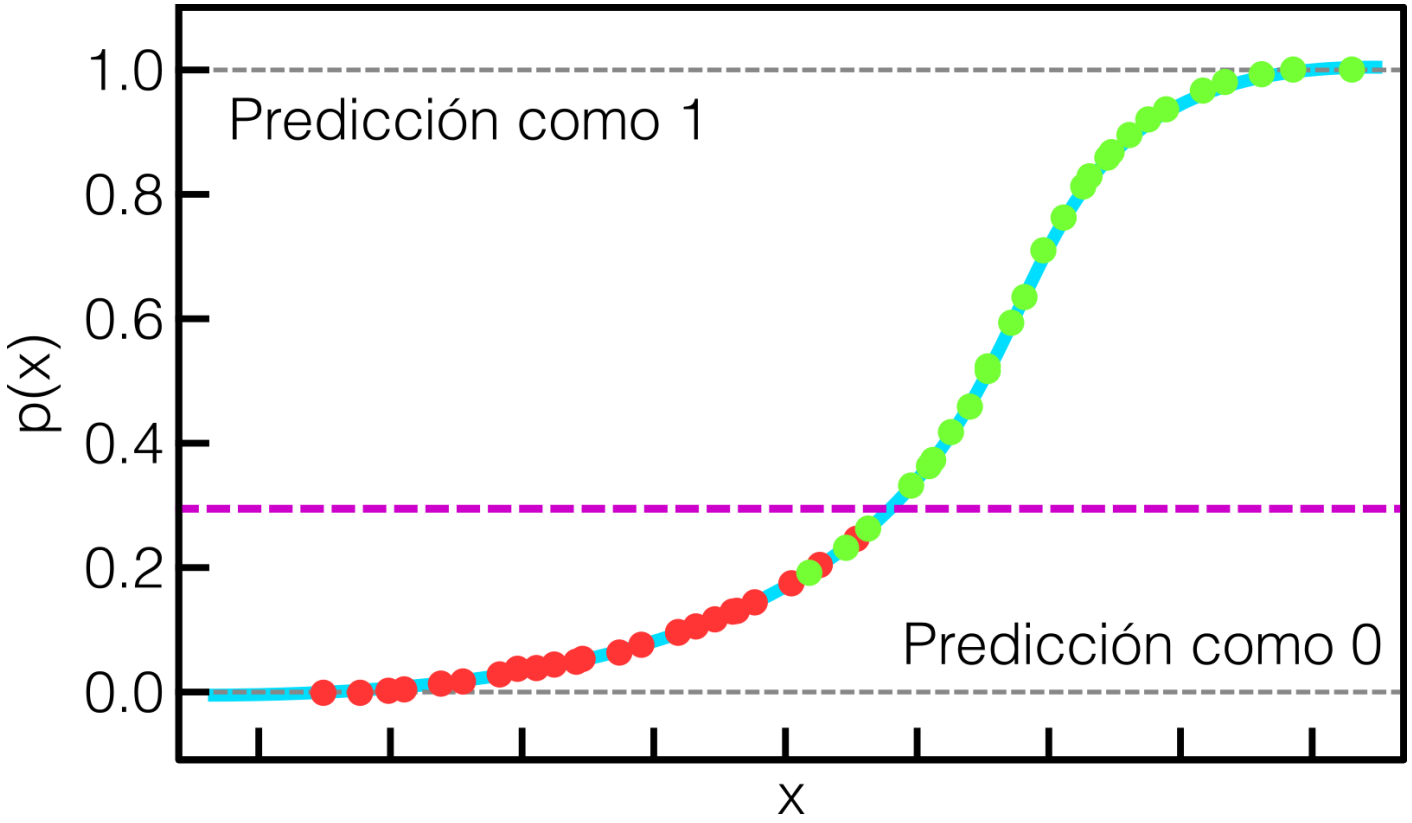
CURVA ROC



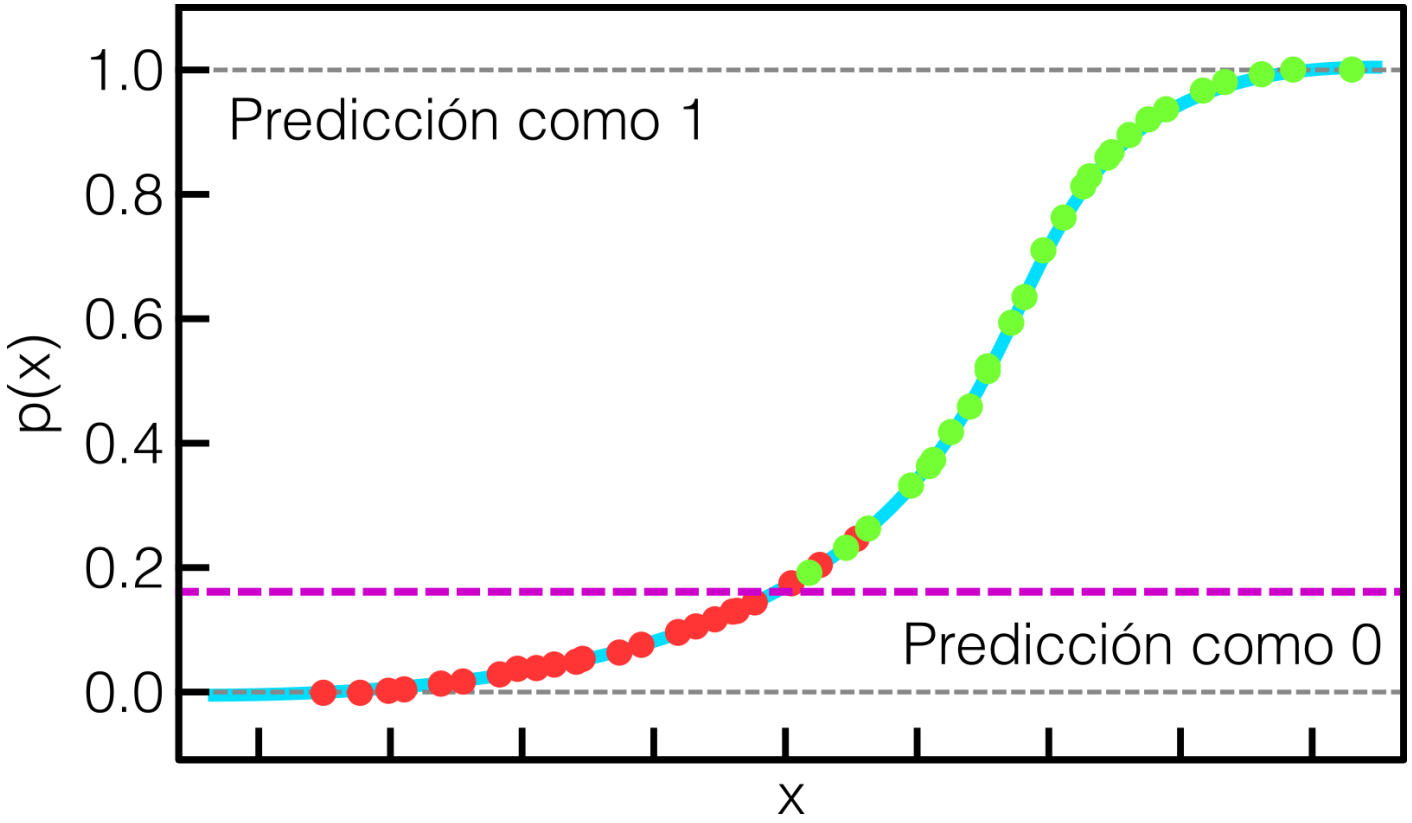
CURVA ROC



CURVA ROC



CURVA ROC

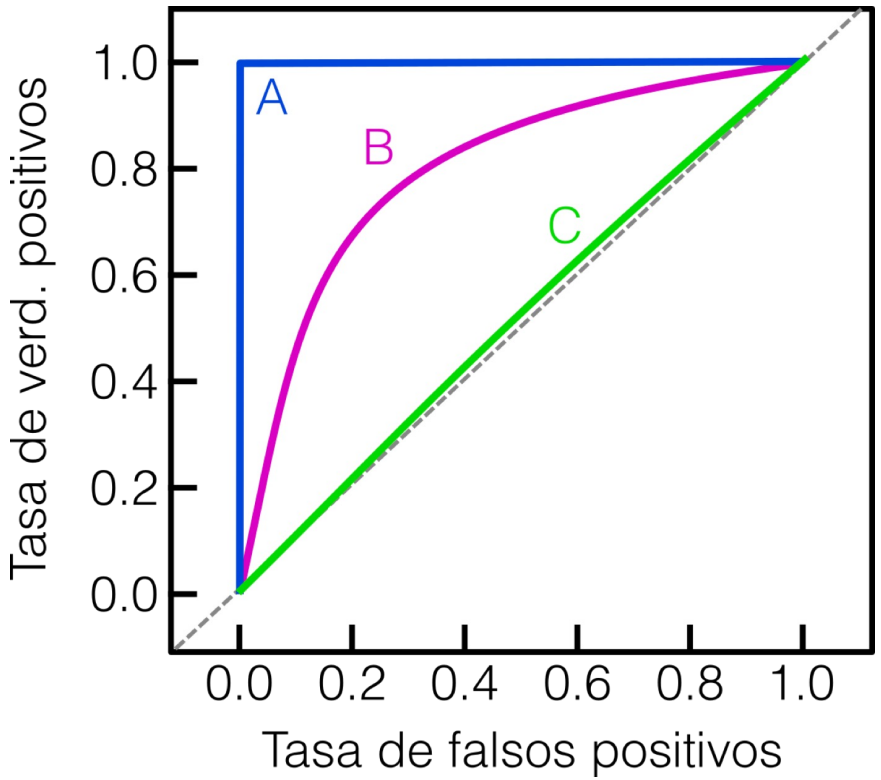


CURVA ROC

La curva ROC nos permite ver para todo valor de umbral, los dos tipos de errores. En el eje de las abscisas se utiliza la **tasa de falsos positivos** (o 1-especificidad) y en la ordenada **la tasa de verdadero positivos** (sensibilidad).

La curva se obtiene midiendo la sensibilidad y la especificad para todos los valores de umbrales de 0 a 1.

CURVA ROC



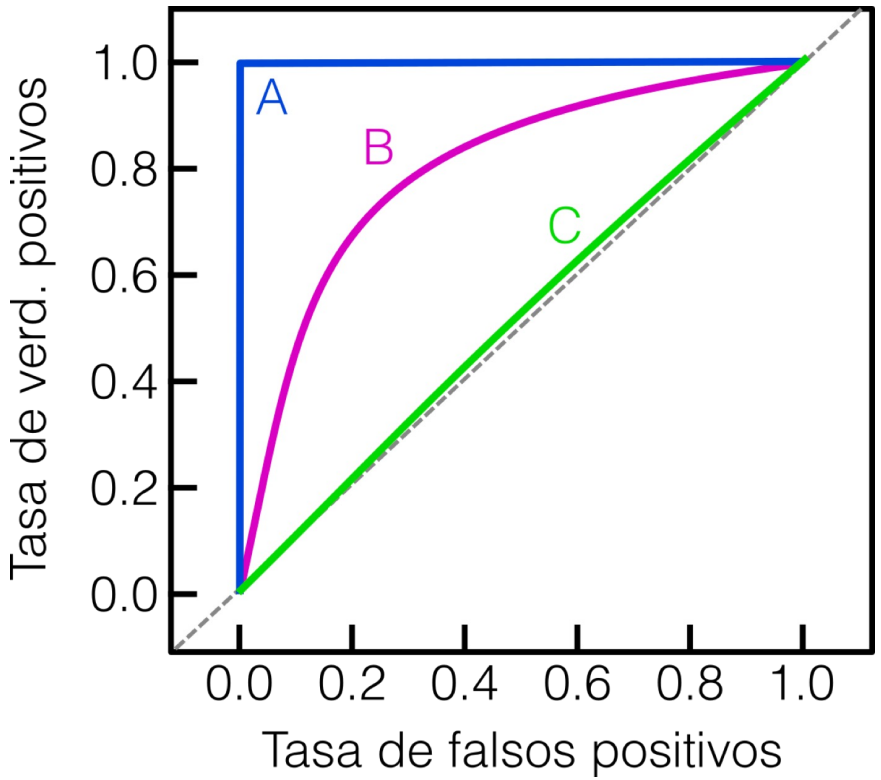
Siempre se arranca de umbral 1, donde la TPR es 0 y TFP es 0 y termina en 0 donde TVP es 1 y TFP es 1.

- **A** es la curva de un clasificador perfecto
- **B** es la curva de un clasificador estándar.
- **C** es la curva de un clasificador que adivina (el peor caso).

La curva ROC permite encontrar el valor umbral que mejor resultado dé.

Además, permite comparar clasificadores sin preocuparme del valor umbral elegido.

CURVA ROC

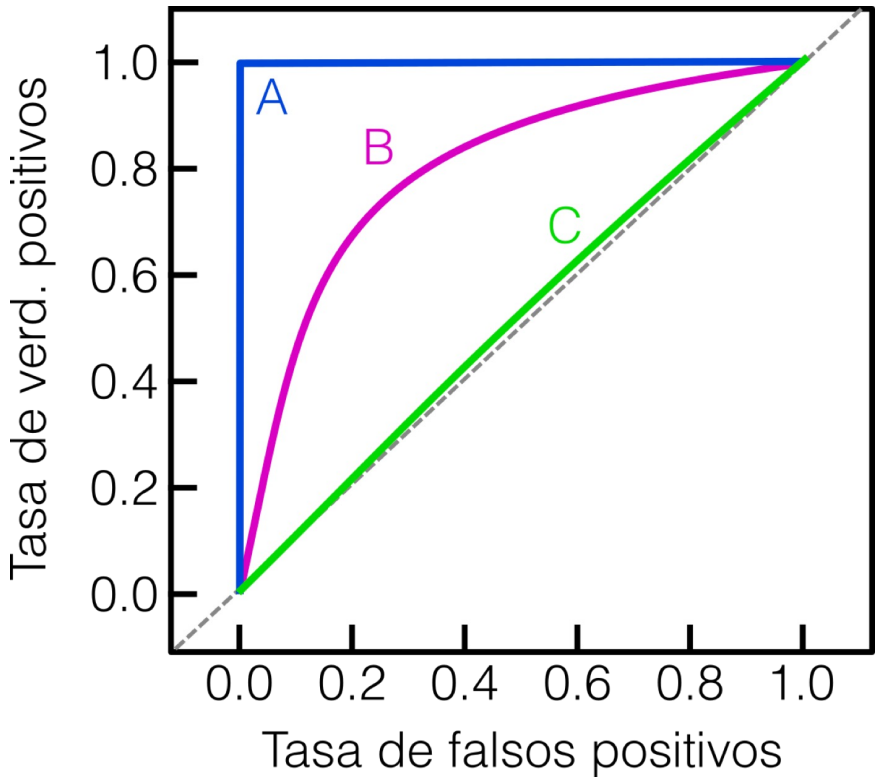


Si quiero bajar a una métrica a esta curva, podemos calcular el área bajo la curva (AUC).

- **A** tendrá un AUC = 1
- **B** tendrá un 0.5 < AUC < 1
- **C** tendrá un AUC = 0.5

Si un clasificador tiene AUC menor a 0.5, ¿qué significa?

CURVA ROC



Si quiero bajar a una métrica a esta curva, podemos calcular el área bajo la curva (AUC).

- **A** tendrá un AUC = 1
- **B** tendrá un 0.5 < AUC < 1
- **C** tendrá un AUC = 0.5

Si un clasificador tiene AUC menor a 0.5, ¿qué significa?

Significa que las clases están **invertidas**. Con solo cambiar las positivas por negativas, se soluciona.

REGRESIÓN LOGÍSTICA MULTI-CLASE

REGRESIÓN LOGÍSTICA MULTI-CLASE

Hasta ahora hemos visto clasificadores binarios, es decir, pueden predecir dos clases. Pero es posible extender a la regresión logística para que pueda predecir 3 o más clases (K).

Por ejemplo, si queremos clasificar entre 3 clases: perro, gato y tero.

Creamos tres regresiones logísticas individuales, y para una observación particular tenemos:

$$[0.73, 0.55, 0.2]$$

Vemos que si sumamos a los tres nos dan mayor a uno ($0.73 + 0.55 + 0.2 = 1.48$), y por lo tanto perdemos lo que buscamos, que se mantenga una probabilidad

REGRESIÓN LOGÍSTICA MULTI-CLASE

Si normalizamos los tres valores con respecto a la suma, recuperamos esta habilidad:

$$\left[\frac{0.73}{1.48}, \frac{0.55}{1.48}, \frac{0.2}{1.48}\right]$$

$$[0.49, 0.38, 0.13]$$

Y nuestro clasificador combinado nos dice que, para esta observación, la observación es más probable que sea un perro. Cuando tenemos multi-clase, se elige la salida más grande.

Obsérvese además que esta salida tiene una forma de **one-hot encoding**.

$$[1, 0, 0]$$

REGRESIÓN LOGÍSTICA MULTI-CLASE

Este proceso es el que llamamos regresión logística multi-clase:

$$P(y = k|X) = \frac{e^{b_k + \boldsymbol{W}_k^T \boldsymbol{X}}}{\sum_k e^{b_{(k)} + \boldsymbol{W}_{(k)}^T \boldsymbol{X}}}$$

Se puede chequear que esta fórmula vuelve a la formula original si tenemos clases, y se hace:

- $b = b_1 - b_0$
- $\mathbf{W} = \boldsymbol{W}_1 - \boldsymbol{W}_0$

REGRESIÓN LOGÍSTICA MULTI-CLASE

De hecho, no importa cuantas clases hay, siempre podemos elegir una clase y hacer que todos sus parámetros sean cero, sin perder generalidad.

Esto es porque la probabilidad de una clase está formada el complemento de las otras. En general, por convención se elige a la clase 0:

$$P(y=0|\boldsymbol{X}) = \frac{e^{0}}{e^{0} + \sum_{k=1}^{K-1} e^{b_{(k)} + \boldsymbol{W}_{(k)}^{T}\boldsymbol{X}}} = \frac{1}{1 + \sum_{k=1}^{K-1} e^{b_{(k)} + \boldsymbol{W}_{(k)}^{T}\boldsymbol{X}}}$$

$$P(y=k|\boldsymbol{X}) = \frac{e^{b_k + \boldsymbol{W}_k^{T}\boldsymbol{X}}}{1 + \sum_{k=1}^{K-1} e^{b_{(k)} + \boldsymbol{W}_{(k)}^{T}\boldsymbol{X}}}$$