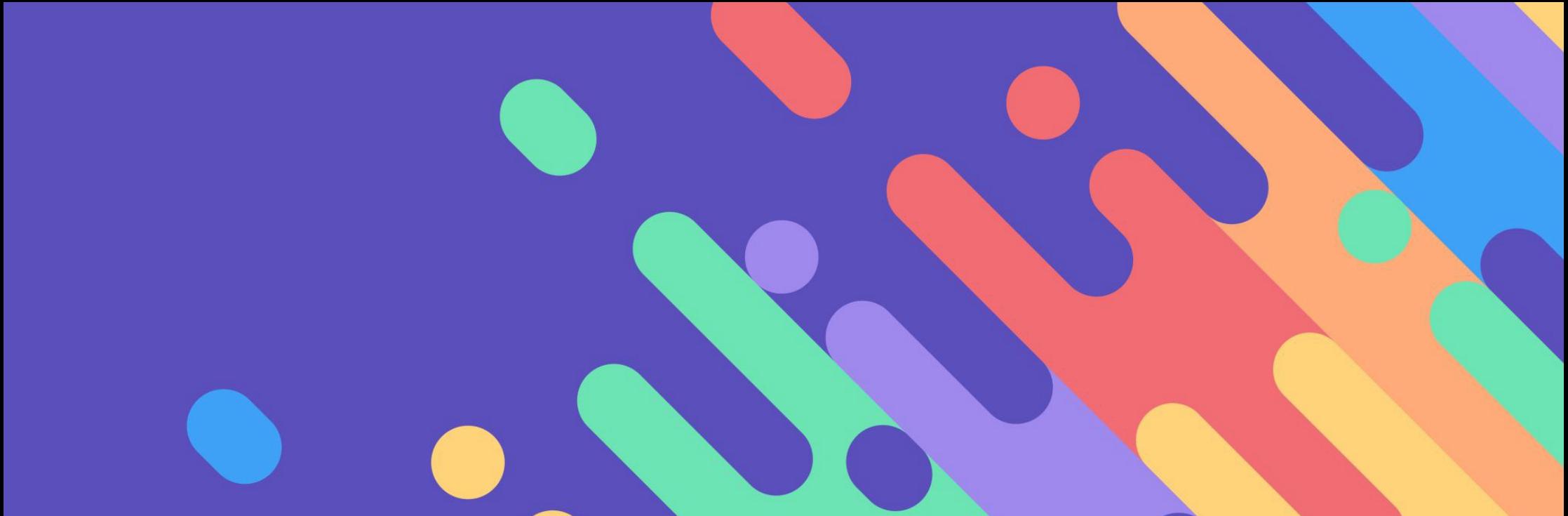


CLASIFICADORES



Inteligencia Artificial
CEIA - FIUBA
Dr. Ing. Facundo Adrián
Lucianna



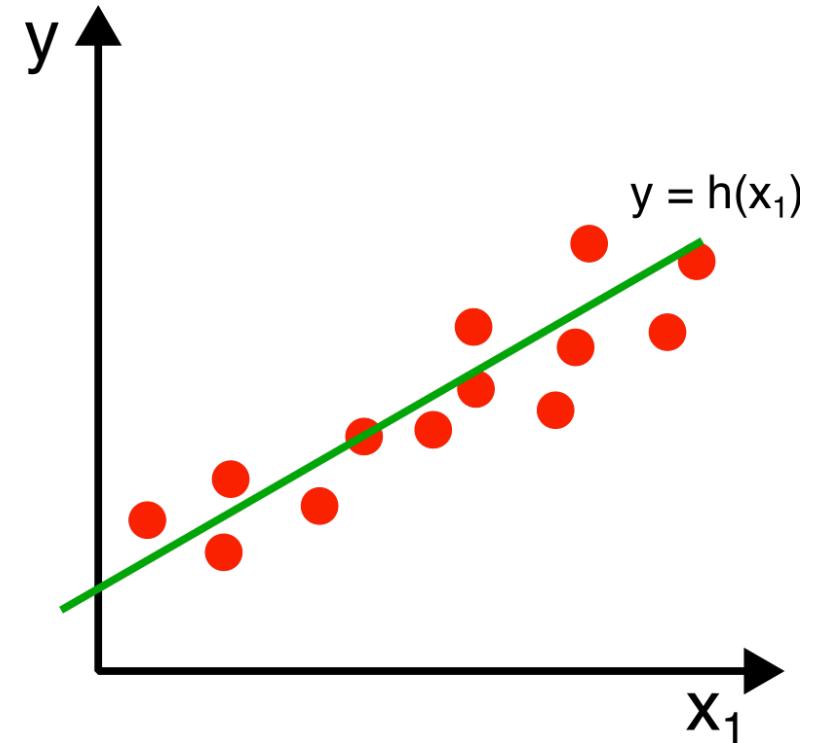
LO QUE VIMOS LA CLASE ANTERIOR...

REGRESIÓN

Si tenemos un problema donde el target y es una *variable numérica*, 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

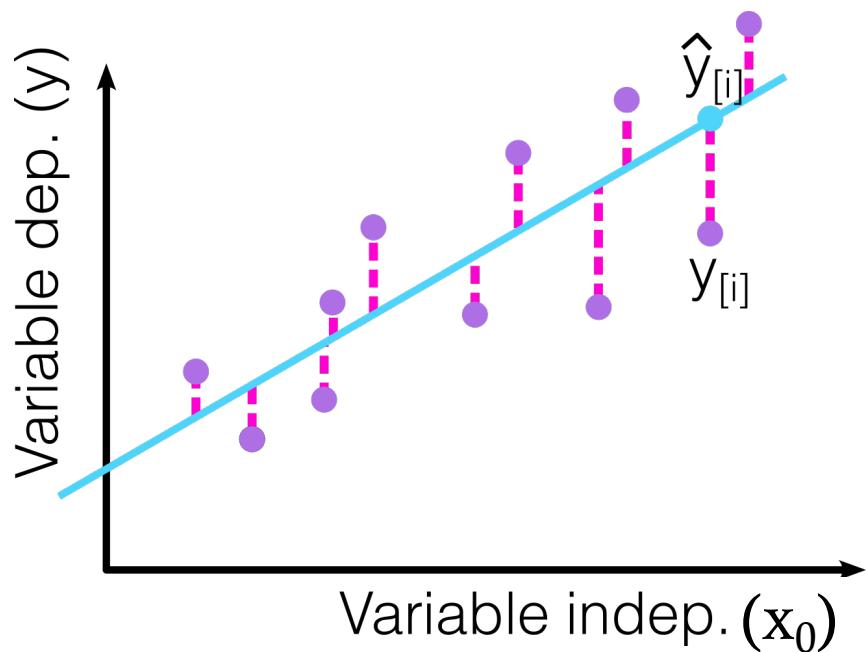
El modelo de regresión lineal más simple es el que involucra una combinación lineal de las variables de entradas:

$$\hat{y} = h(\mathbf{X}) = b + w_0x_0 + \cdots + w_dx_d$$

- $\mathbf{X} = (x_0, x_1, \dots, x_d)$ Son los *features* de nuestras observaciones. Son todas variables numéricas
- b, w_0, \dots, 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**.



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

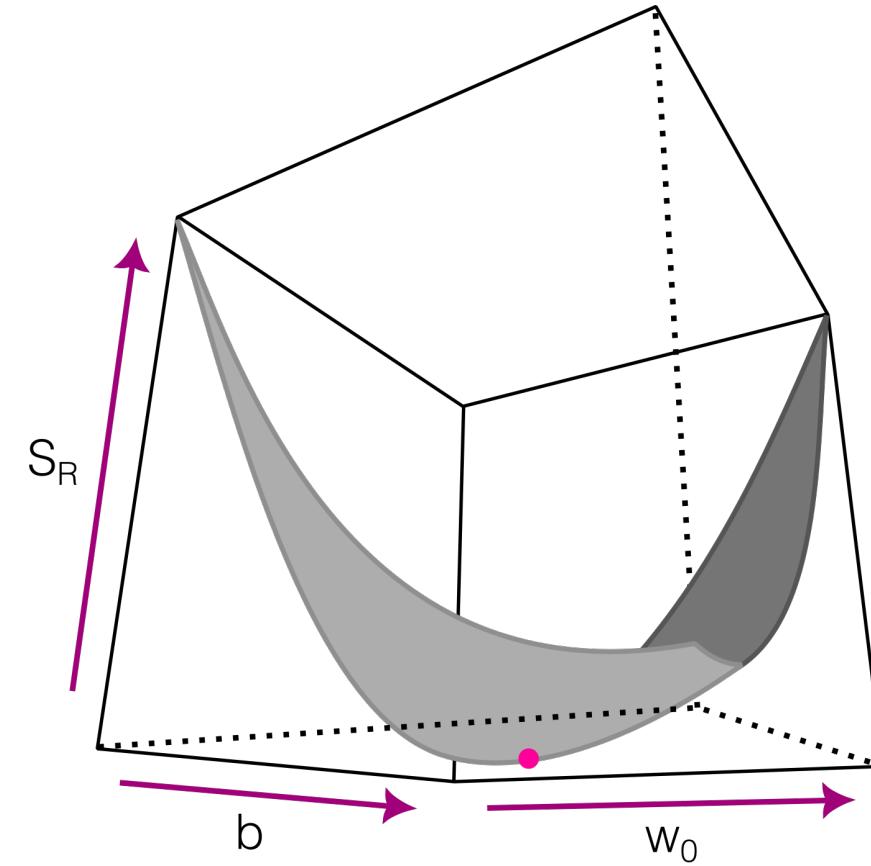
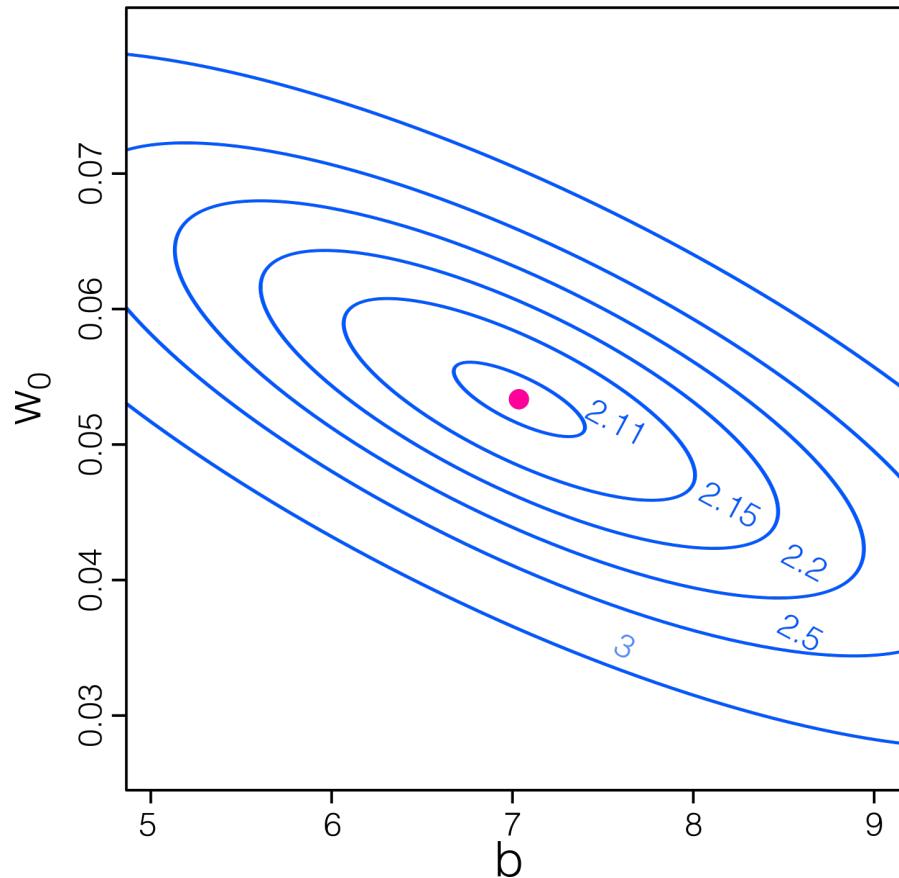
$$\min(S_R) = \min\left(\sum_{i=0}^{N-1} (e_{[i]})^2\right)$$

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

$$\frac{\partial S_R}{\partial b} = 0$$

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

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: $MSE = \frac{1}{N} \sum_{i=0}^{N-1} (y_{[i]} - \hat{y}_{[i]})^2$
- 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:

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

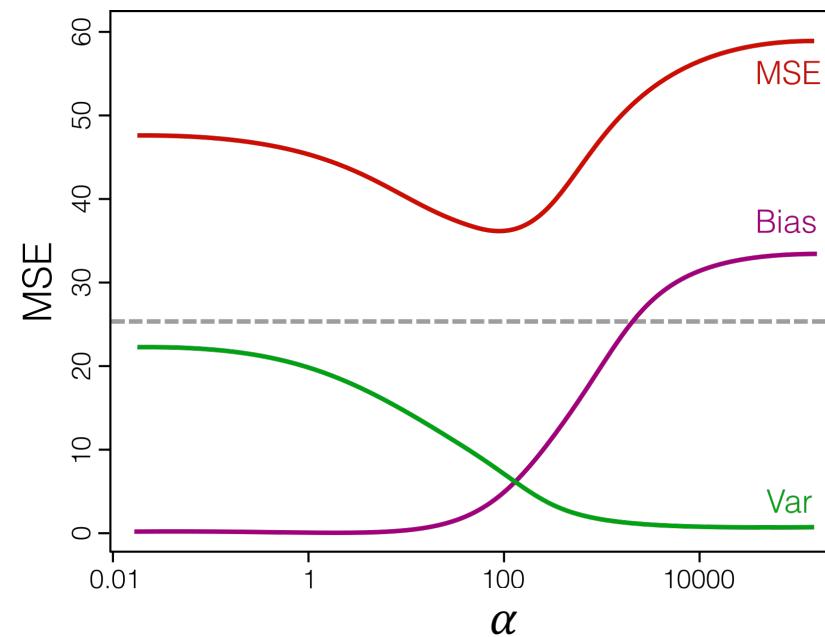
Regresión de Lasso:

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

REGRESIÓN DE RIDGE

¿Para qué nos sirve?

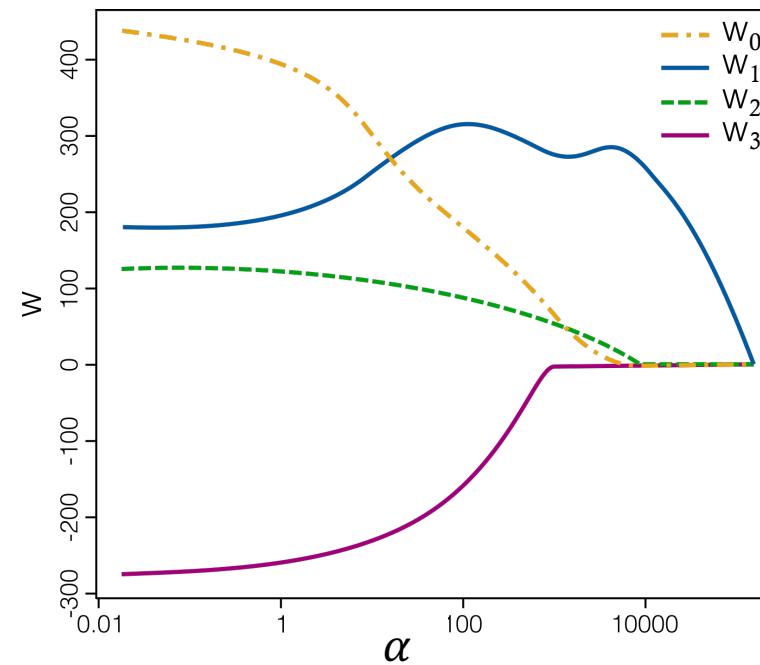
En general, cuando la verdadera relación es lineal, la regresión lineal tiene mucha varianza. Esto principalmente ocurre cuando el **número de observaciones es cercano al número de coeficientes**. En estos casos, la regresión de Ridge funciona mejor.

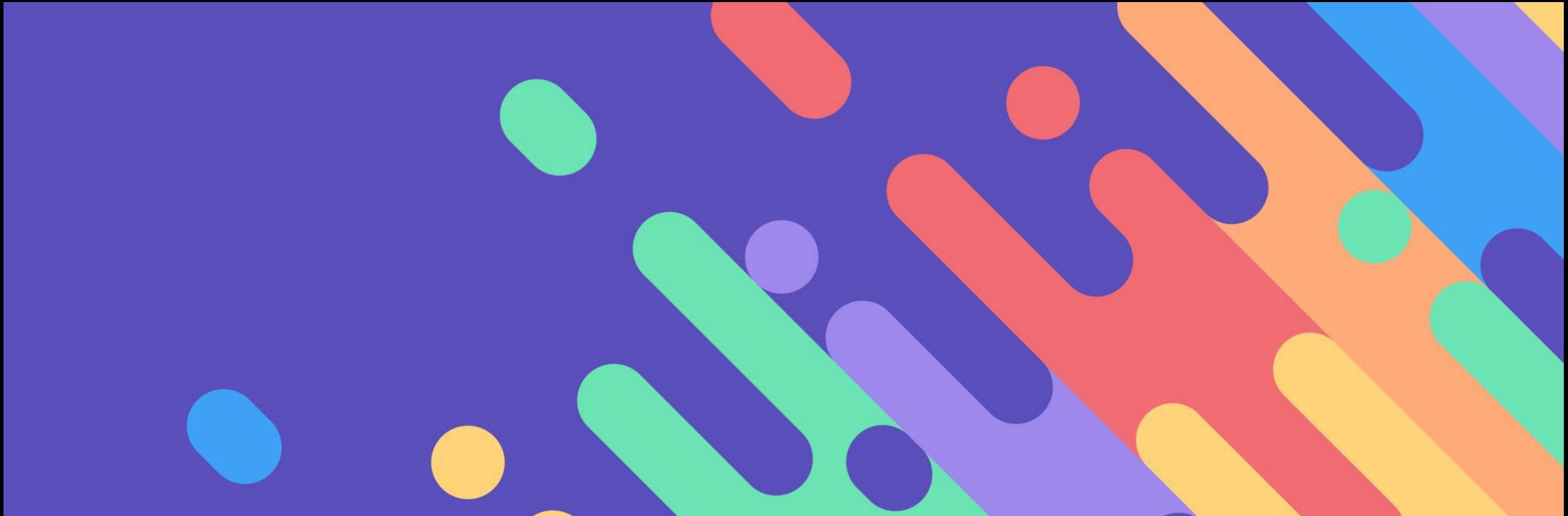


REGRESIÓN DE LASSO

¿Para qué nos sirve?

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



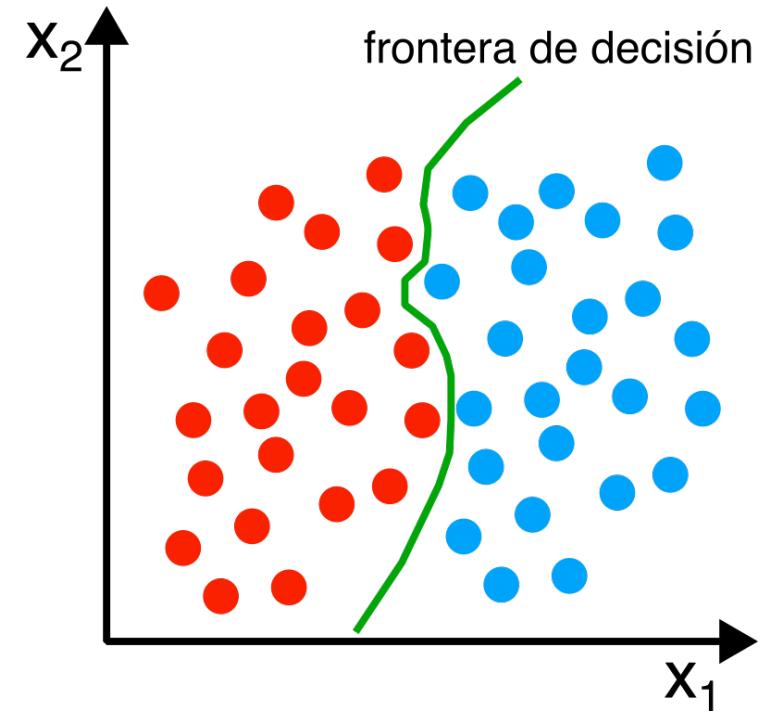


CLASIFICACIÓN

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.



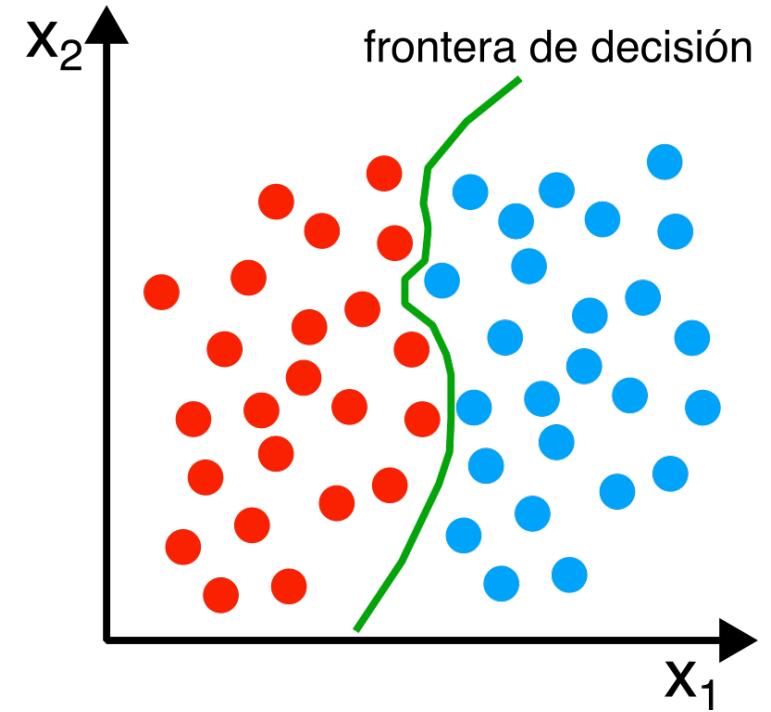
CLASIFICACIÓN

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 **categoría**.

¿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.

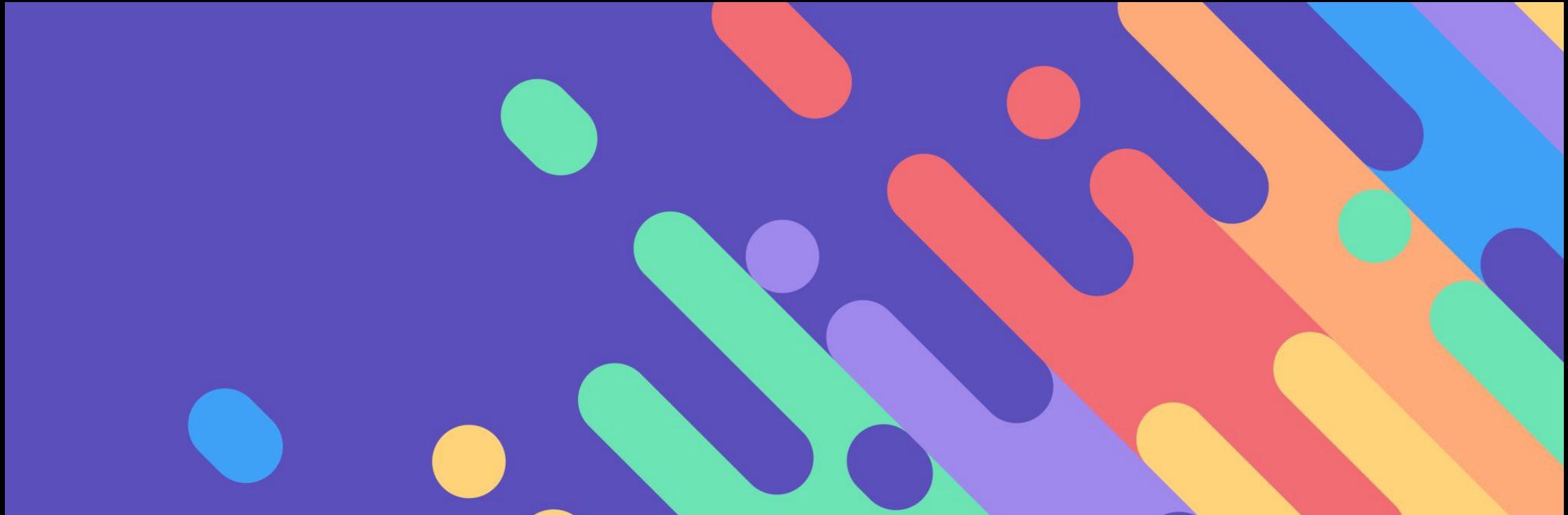
Cada una de estas codificaciones produciría modelos lineales diferentes que, en última instancia, conducirían a diferentes conjuntos de predicciones sobre observaciones de prueba.

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

$$P(ACV = 1|X) = b + W^T X$$



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.

$$P(y = k|X)$$

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:

$$P(y = 0|X)$$

$$P(y = 1|X)$$

REGRESIÓN LOGÍSTICA

Pero, además, en el caso de dos clases:

$$P(y = 1|X) = 1 - P(y = 0|X)$$

Por lo que podemos simplificar la notación...

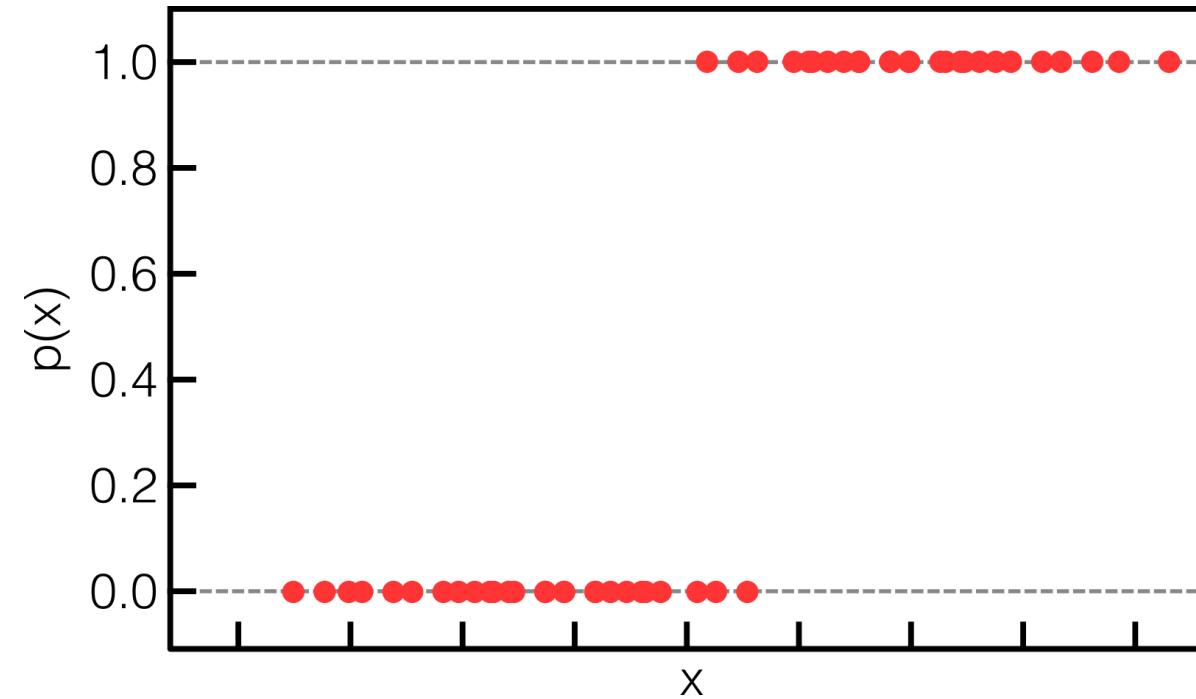
$$P(y = 1|X) = p(X)$$

Las probabilidades son valores que van entre 1 y 0.

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

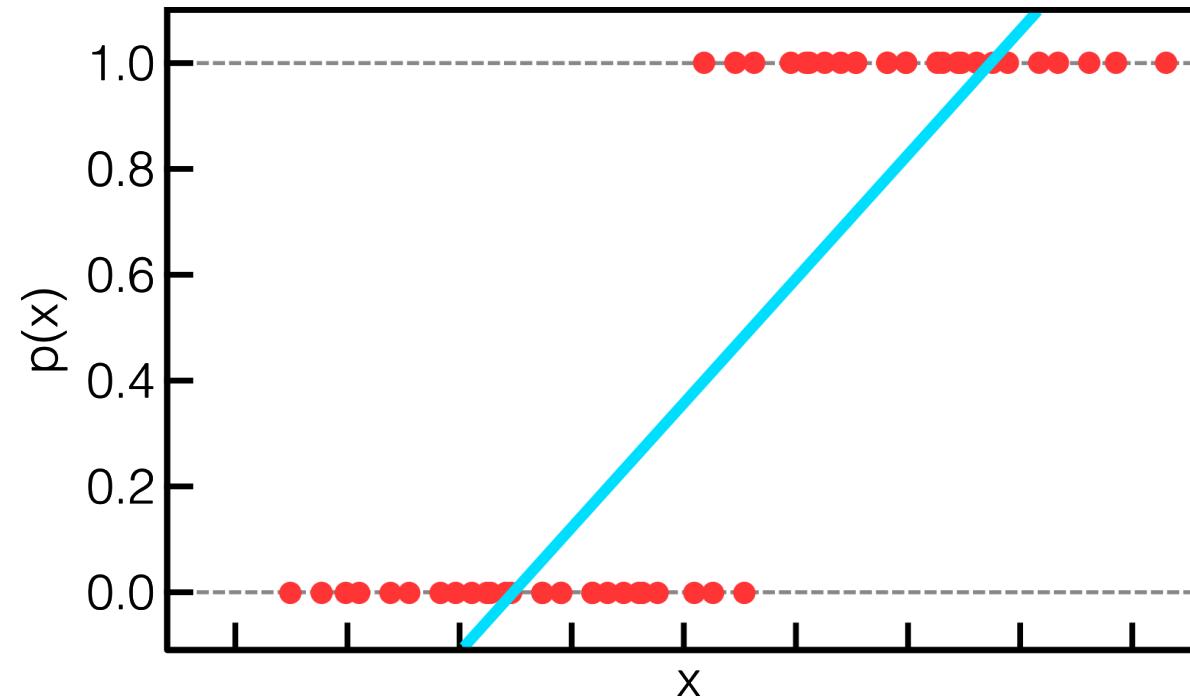
REGRESIÓN LOGÍSTICA

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



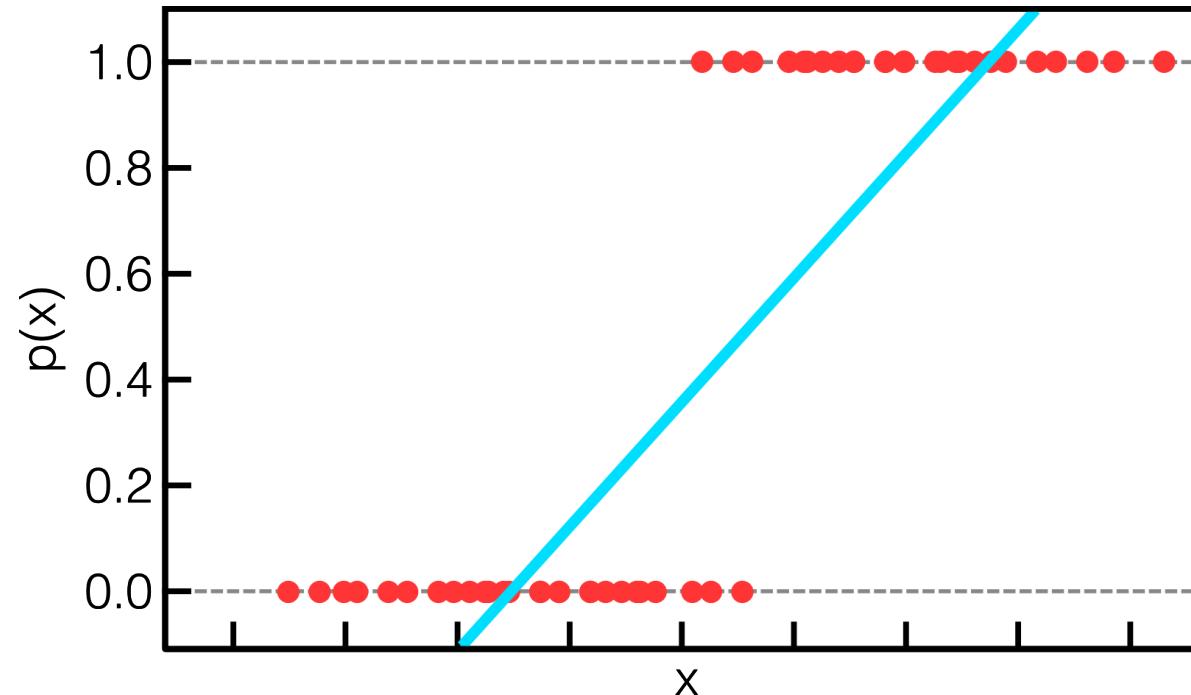
REGRESIÓN LOGÍSTICA

Podemos usar una regresión lineal para estimar la probabilidad $p(x) = b + w_0x$



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

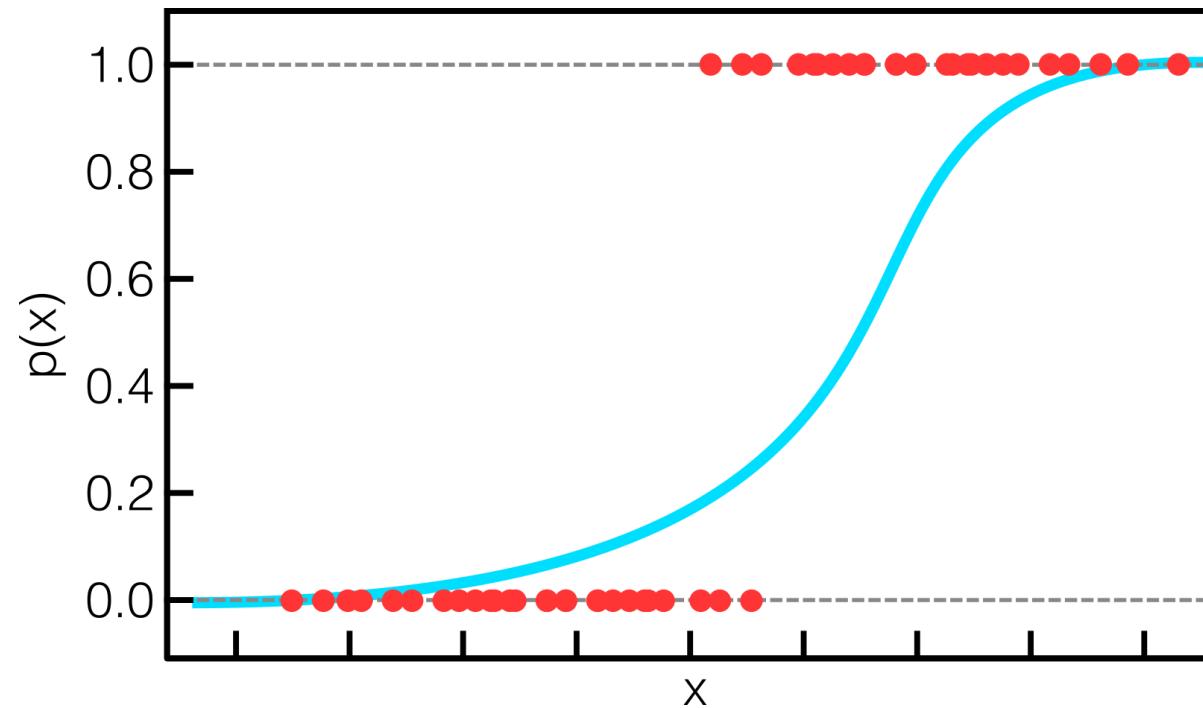
Para evitar esto, podemos modelar a la probabilidad usando una función que nos asegure que siempre tendremos valores entre 0 y 1.

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

$$p(x) = \frac{e^{b+w_0x}}{1 + e^{b+w_0x}} = \frac{1}{1 + e^{-(b+w_0x)}}$$

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 $p(x) = \frac{e^{b+w_0x}}{1+e^{b+w_0x}}$, llegamos a:

$$\frac{p(x)}{1 - p(x)} = e^{b+w_0x}$$

El cuál es la **chance** (o en inglés **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:

$$\text{logit}(p) = \ln\left(\frac{p(x)}{1 - p(x)}\right) = b + w_0x$$

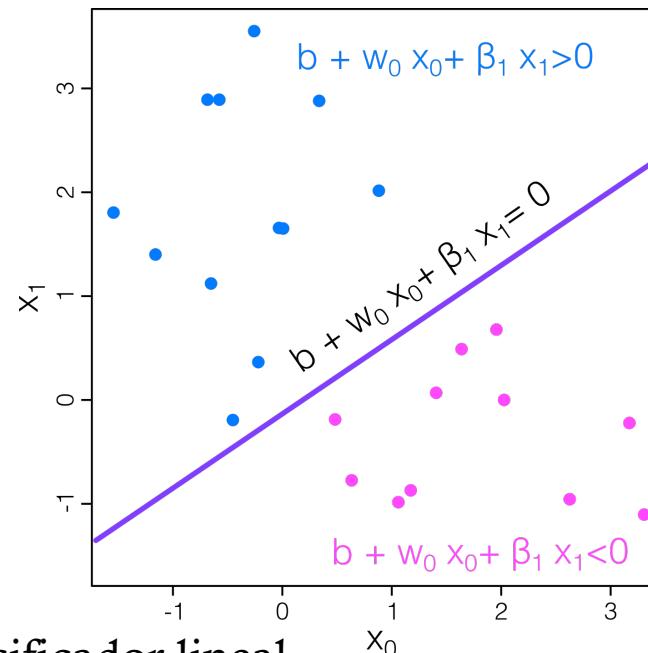
Obtenemos la función **logit**. *La función logit se utiliza para transformar variables de entrada en un rango que puede interpretarse como probabilidades. En la regresión logística es una relación lineal.*

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^{w_0} .

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 $p(x) = \frac{e^{b+w_0x}}{1+e^{b+w_0x}}$ 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

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

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

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

REGRESIÓN LOGÍSTICA - AJUSTE

$$J(b, w_0) = - \sum_{i=1}^N \ln \left(1 - \frac{e^{b+w_0x_i}}{1 + e^{b+w_0x_i}} \right) - \sum_{i=1}^N y_i(b + w_0x_i)$$

$$J(b, w_0) = \sum_{i=1}^N \ln(1 + e^{b+w_0x_i}) - \sum_{i=1}^N y_i(b + w_0x_i)$$

$$J(b, w_0) = \sum_{i=1}^N \ln(1 + e^{b+w_0x_i}) - y_i b - y_i w_0 x_i$$

REGRESIÓN LOGÍSTICA - AJUSTE

Para encontrar el mínimo, aplicamos el gradiente:

$$\frac{\partial J}{\partial b} = \sum_{i=1}^N \frac{e^{b+w_0x_i}}{1 + e^{b+w_0x_i}} - y_i = 0$$

$$\frac{\partial J}{\partial w_0} = \sum_{i=1}^N \left(\frac{e^{b+w_0x_i}}{1 + e^{b+w_0x_i}} - y_i \right) x_i = 0$$

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:

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

En este caso, el gradiente queda:

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

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



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

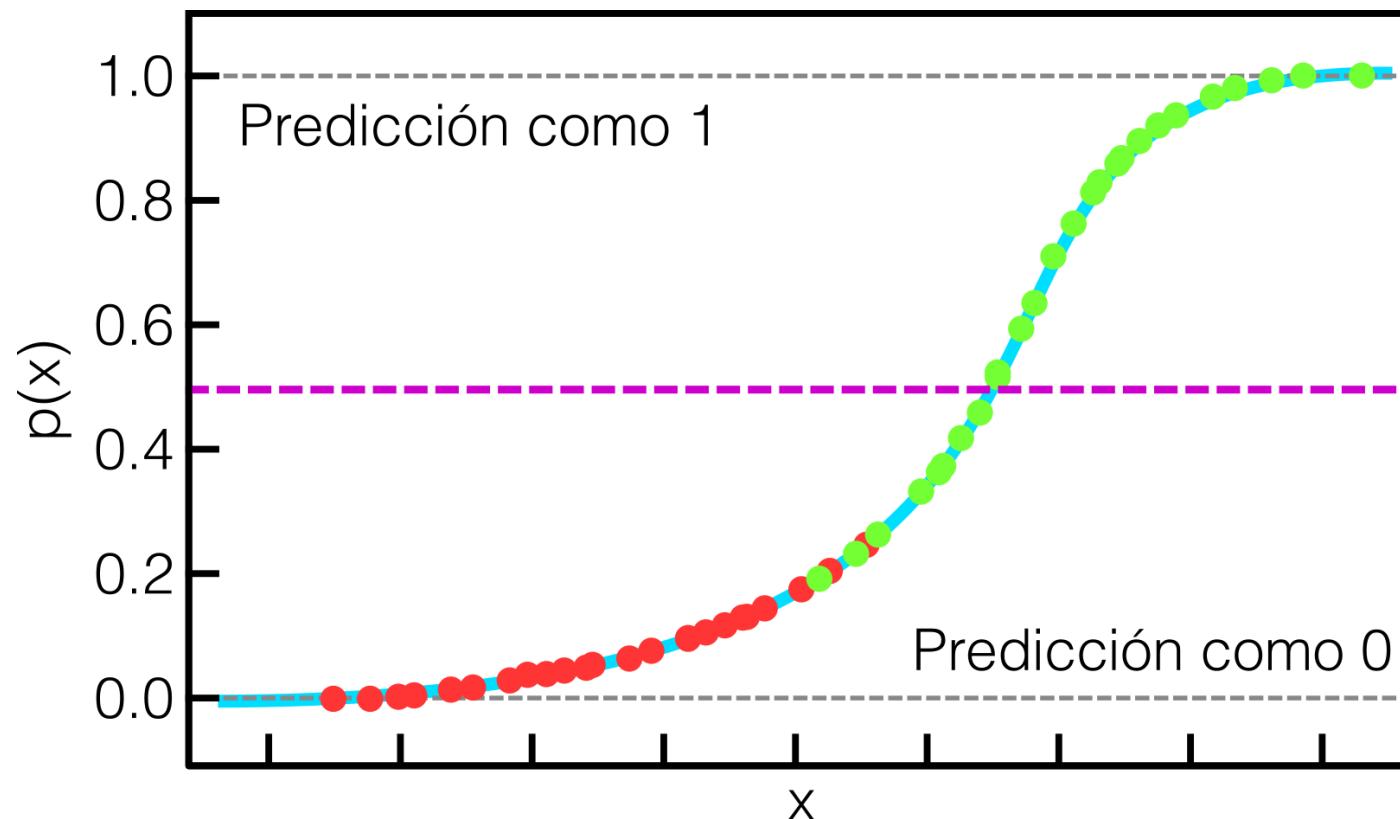
¿Pero por qué este valor?

CURVA ROC

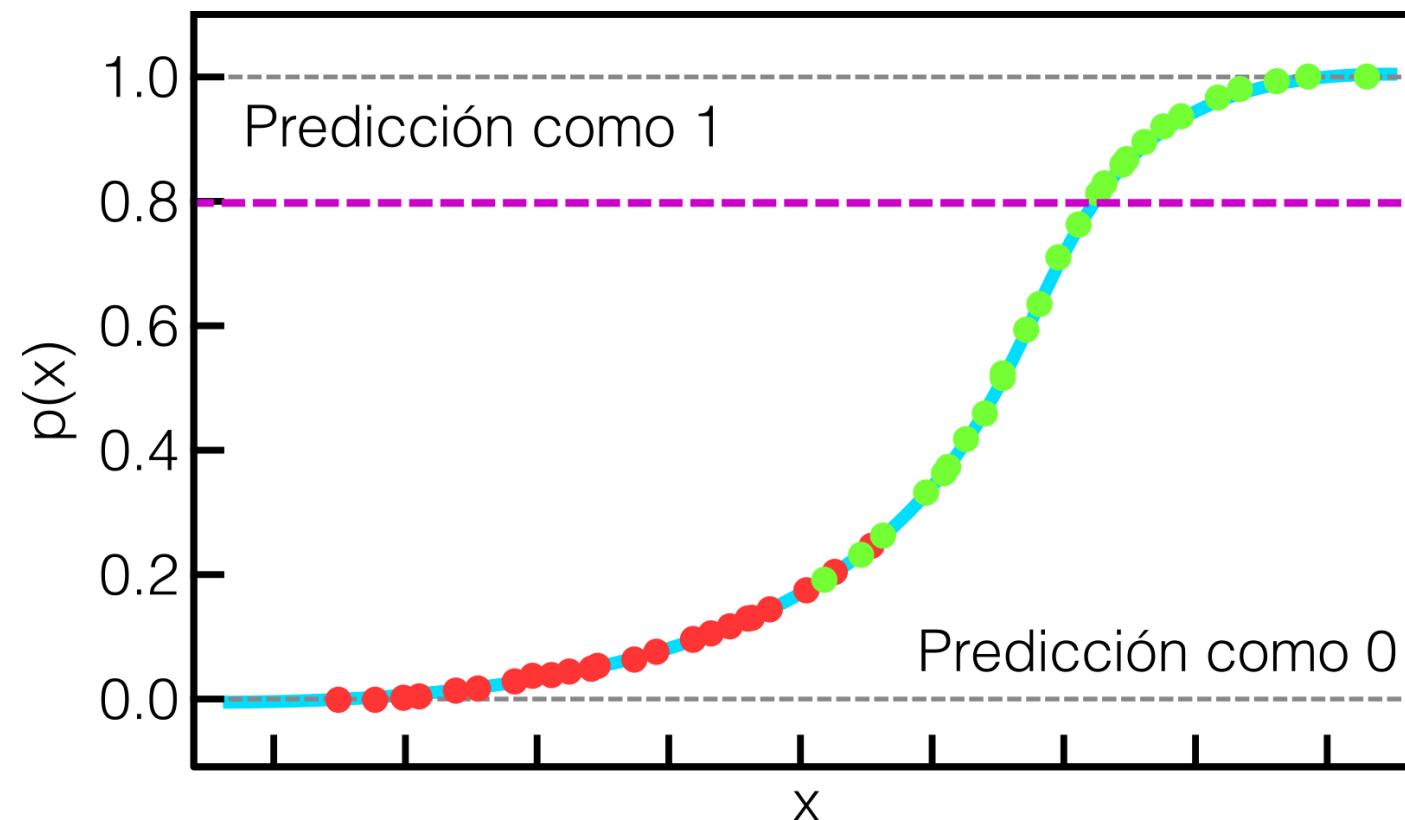
¿Pero por qué este valor?

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.

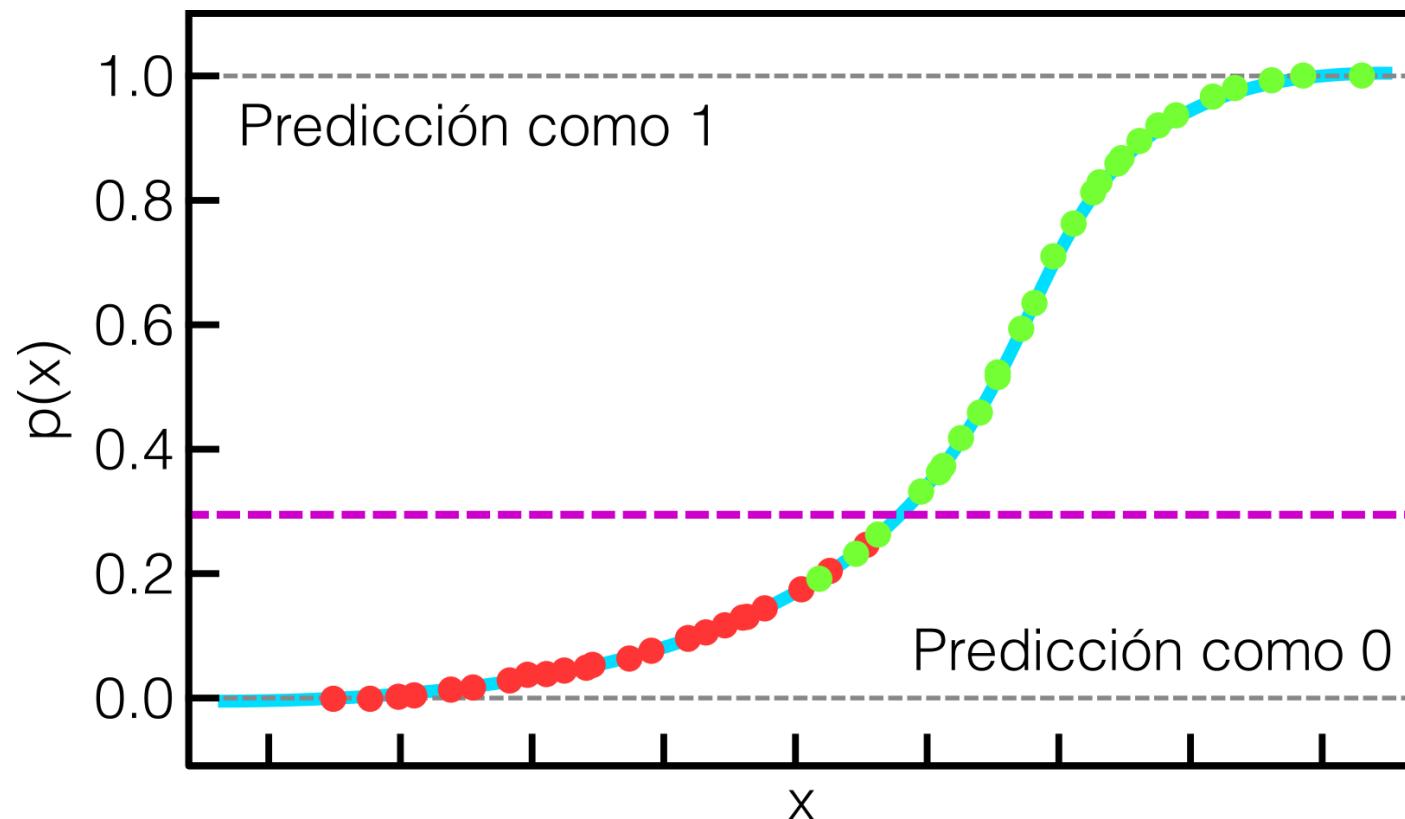
CURVA ROC



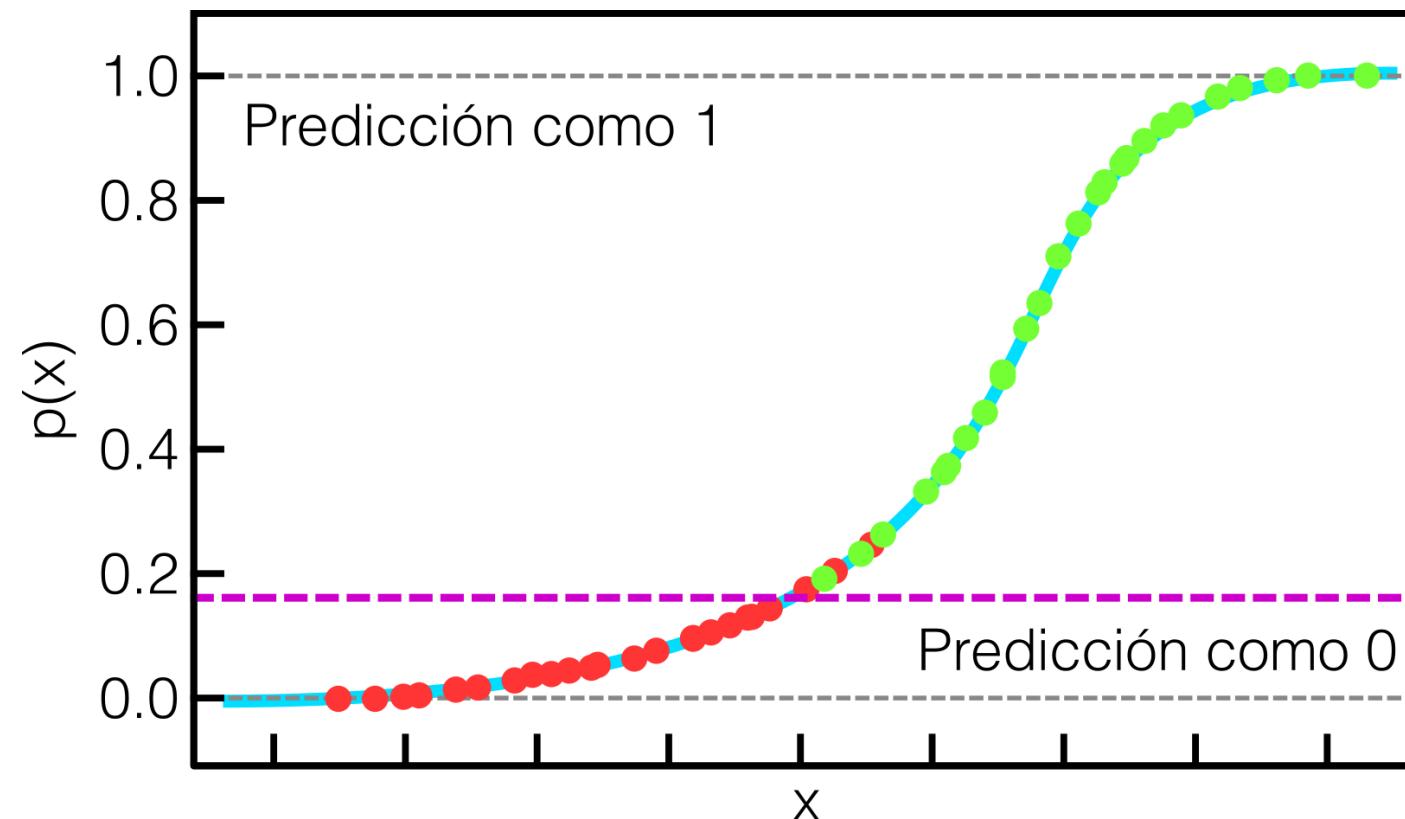
CURVA ROC



CURVA ROC



CURVA ROC

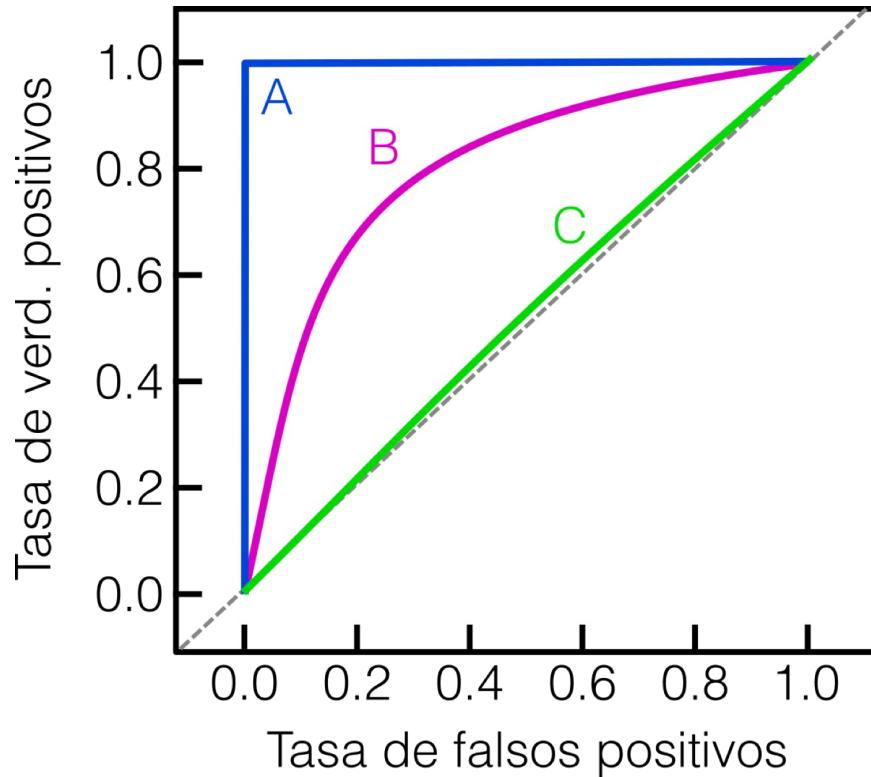


CURVA ROC

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



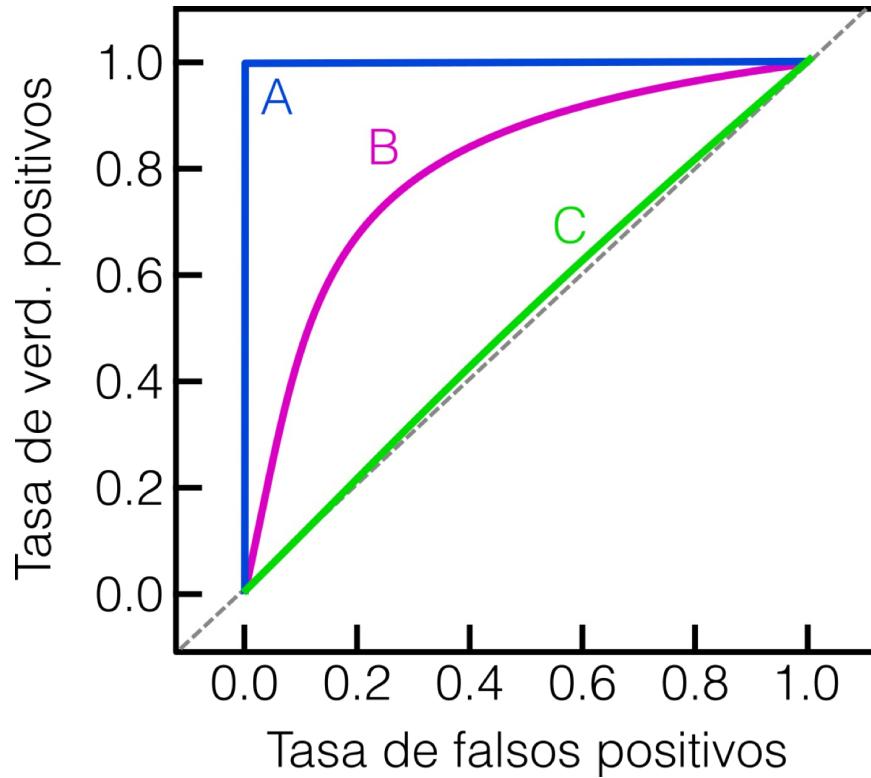
Siempre se arranca de umbral 1, donde la TPR es 0 y TFP es 0 y termina en 0 donde TVP es 1 y TFP es 1.

- **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

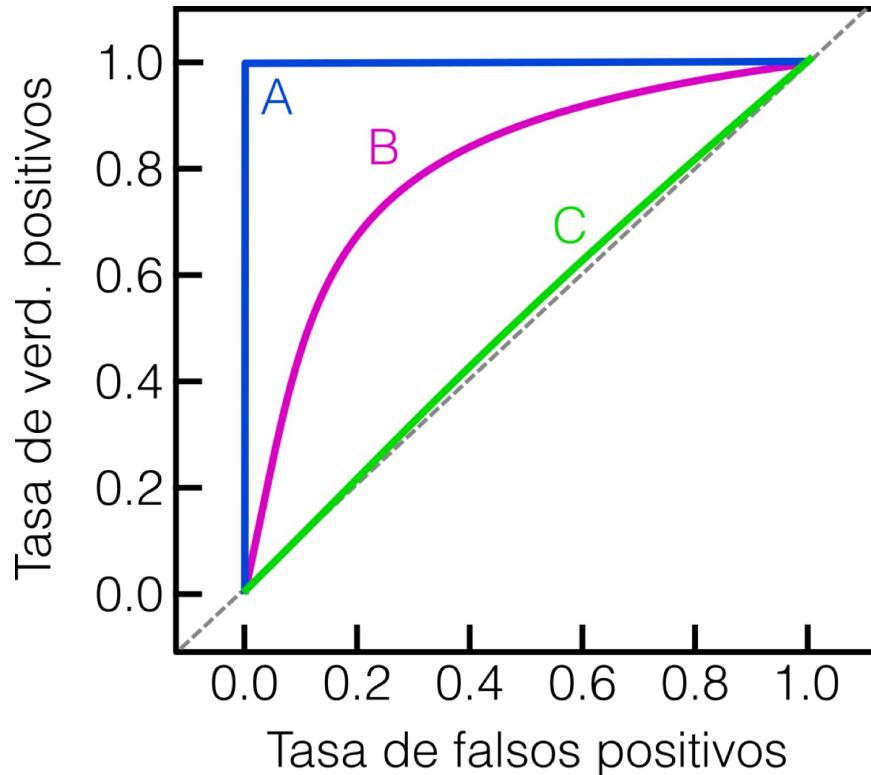


Si quiero bajar a una métrica a esta curva, podemos calcular el área bajo la curva (AUC).

- **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

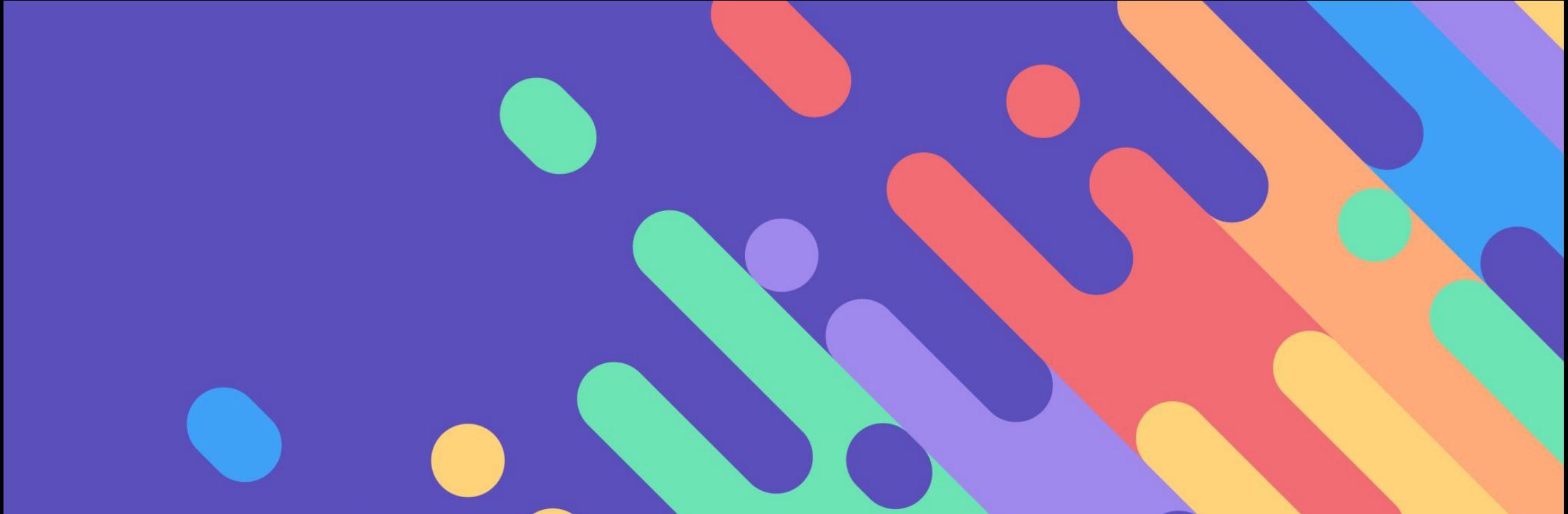


Si quiero bajar a una métrica a esta curva, podemos calcular el área bajo la curva (AUC).

- **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:

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

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

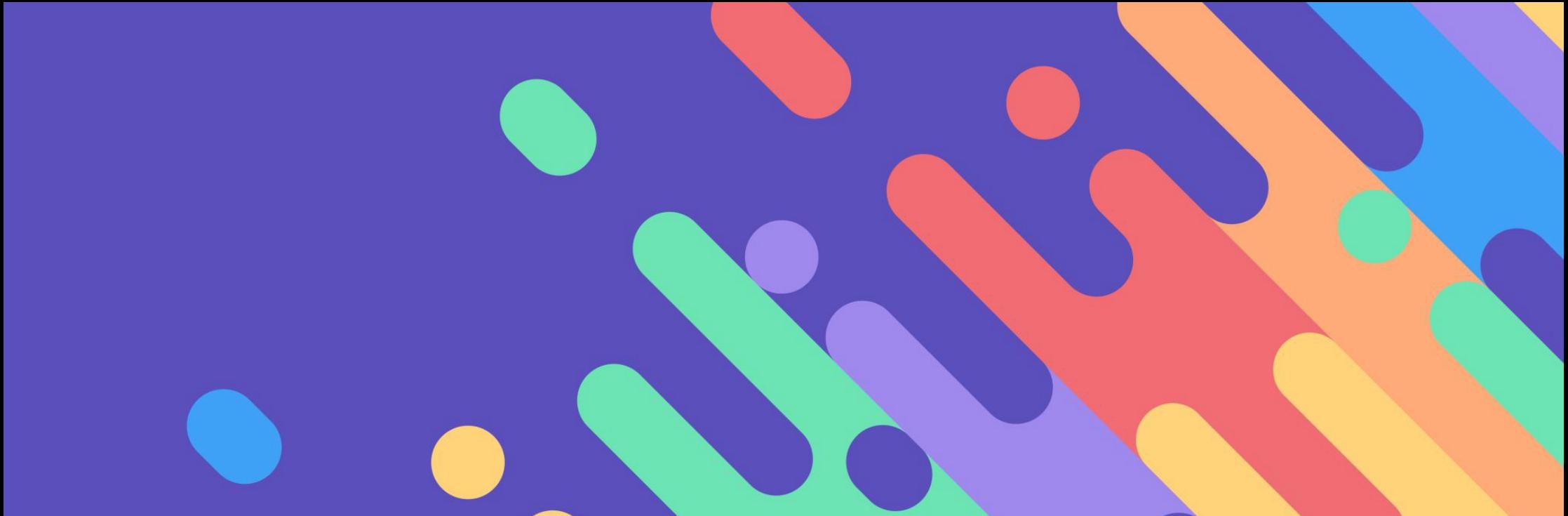
- $b = b_1 - b_0$
- $\mathbf{W} = \mathbf{W}_1 - \mathbf{W}_0$

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:

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



CLASIFICADOR BAYESIANO INGENUO

CLASIFICADOR BAYESIANO INGENUO

Teorema de Bayes

Este teorema es uno de los teoremas más importantes de probabilidad, y uno que hasta el día de hoy genera divisiones en el plano filosófico por su implicancia

Este describe la probabilidad de un evento, basado en conocimiento previo de condiciones que pueden estar relacionados con el evento.

Por ejemplo, si se sabe que el riesgo de desarrollar problemas de salud aumenta con la edad, el teorema de Bayes permite evaluar con mayor precisión el riesgo para un individuo de una edad conocida condicionándolo en relación con su edad, en lugar de asumir que el individuo es típico de la población en su conjunto.

CLASIFICADOR BAYESIANO INGENUO

Teorema de Bayes

Para tener un entendimiento más intuitivo de este teorema, sigamos el camino que nos guía [3Blue1Brown](#), veamos a Esteban:

Esteban es muy tímido y retraído, invariablemente servicial, pero con muy poco interés en las personas o en el mundo de la realidad. Tranquilo y ordenada, necesita orden y estructura, y pasión por los detalles.

¿Cuál de las siguientes opciones te parece más probable (Esteban es de EEUU)?

- Esteban es bibliotecario
- Esteban es agricultor

CLASIFICADOR BAYESIANO INGENUO

Teorema de Bayes

Este ejercicio es de un trabajo de Daniel Kahneman y Emos Tversky sobre juicios humanos el cual valió un premio Nobel en Economía.

Los autores concluyen que la elección de las personas es **irracional**, y no es dado por nuestro sesgo sobre personalidades de agricultores o bibliotecarios, sino que casi nadie piensa en incorporar información de proporción entre agricultores y bibliotecarios.

CLASIFICADOR BAYESIANO INGENUO

Teorema de Bayes

Según Kahneman y Tversky, hay 20 agricultores por cada bibliotecario.



CLASIFICADOR BAYESIANO INGENUO

Teorema de Bayes

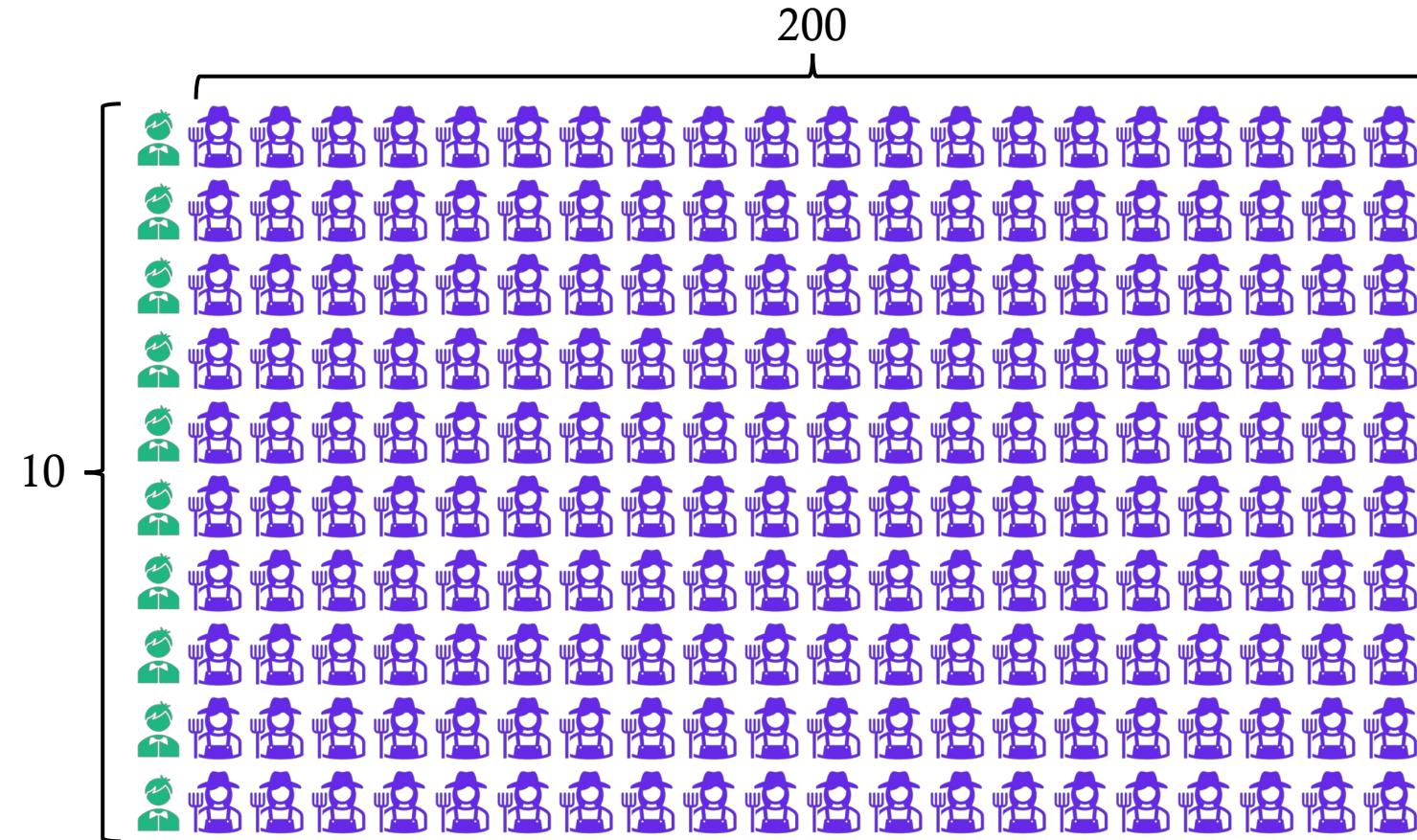
Si se piensa hacer la estimación, hay una forma sencilla de pensar esto, que implica el razonamiento esencial detrás del teorema de Bayes.

Podemos pensar en una muestra representativa de agricultores y bibliotecarios (200 y 10).

Luego, cuando leemos la descripción de *tranquilo y ordenado*, suponemos intuitivamente que el 40% de los bibliotecarios tendrían esa descripción y que el 10% de los agricultores también.

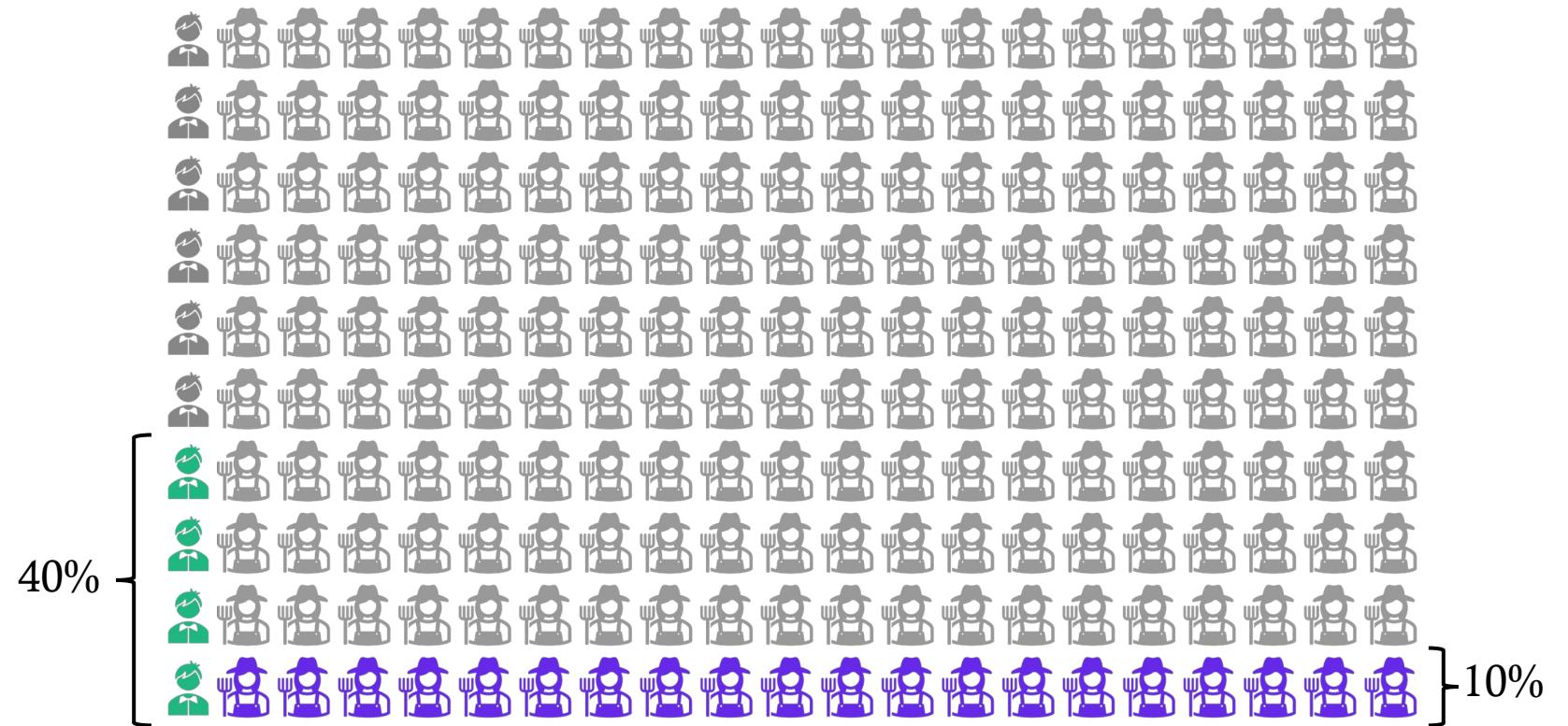
CLASIFICADOR BAYESIANO INGENUO

Teorema de Bayes



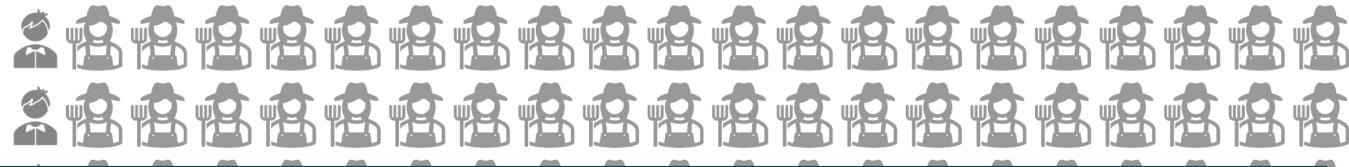
CLASIFICADOR BAYESIANO INGENUO

Teorema de Bayes



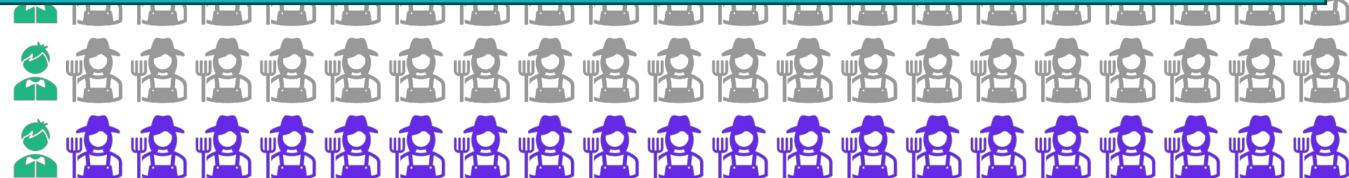
CLASIFICADOR BAYESIANO INGENUO

Teorema de Bayes



Alrededor del 40% de los bibliotecarios encajan en la descripción, pero sólo el 10% de los agricultores lo hacen. ¡Todavía hay más agricultores que bibliotecarios!

La probabilidad de que una persona aleatoria que encaja en la descripción es un bibliotecario es $\frac{4}{24}$, 16.7%



CLASIFICADOR BAYESIANO INGENUO

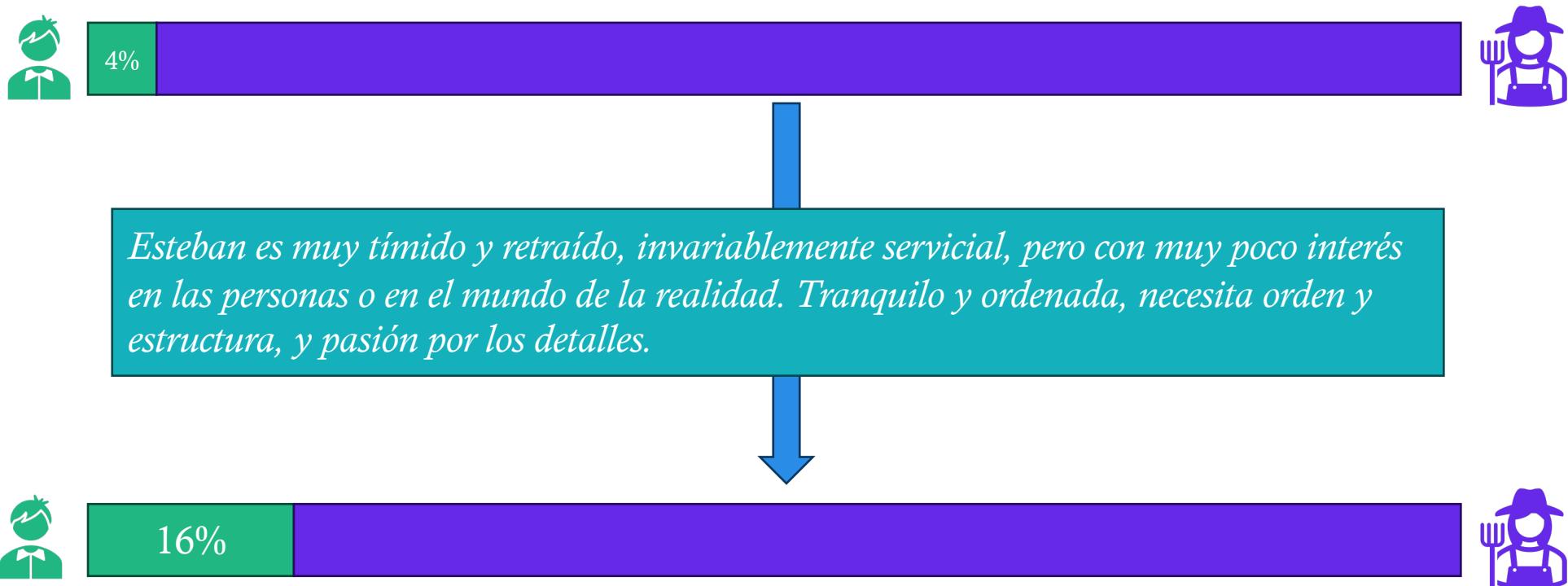
Teorema de Bayes

Entonces, incluso si se piensa que un bibliotecario tiene 4 veces más probabilidades que un agricultor de encajar en esta descripción, eso no es suficiente para superar el hecho de que hay muchos más agricultores.

El resultado es que la nueva evidencia no debería determinar completamente nuestras creencias en el vacío, **debería actualizar creencias anteriores**. Esta es la idea del teorema de Bayes.

CLASIFICADOR BAYESIANO INGENUO

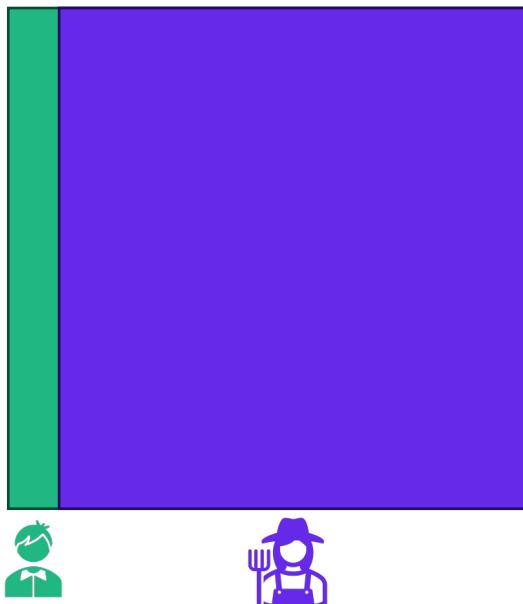
Teorema de Bayes



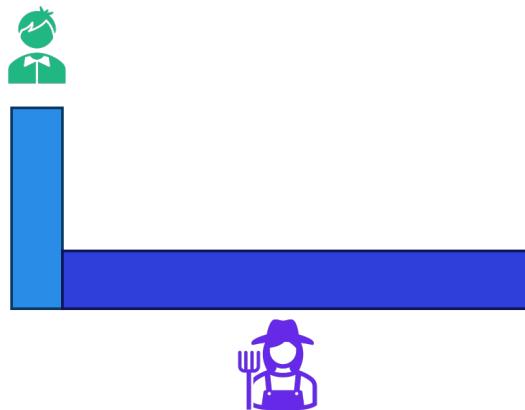
CLASIFICADOR BAYESIANO INGENUO

Teorema de Bayes

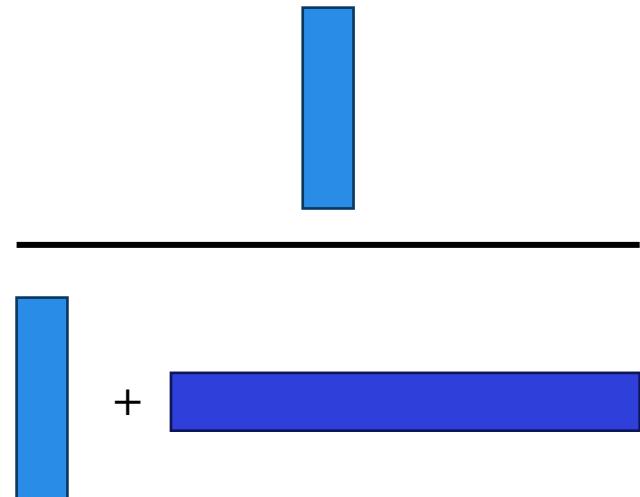
Todas las posibilidades



Todas las posibilidades
que encajan la evidencia



$P(\text{Bibliotecarios dado})$
la evidencia



CLASIFICADOR BAYESIANO INGENUO

Teorema de Bayes

El teorema de Bayes es relevante en situaciones en las que se tiene alguna hipótesis (Esteban es bibliotecario) y se observa alguna evidencia (Esteban es "tranquilo y ordenado") y se quiere saber la probabilidad de que la hipótesis se cumpla dado que la evidencia es cierta.

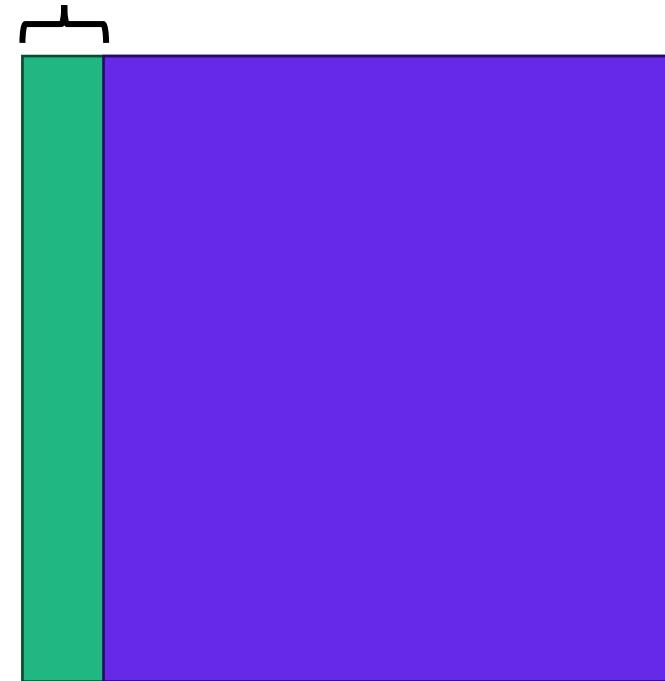
$$P(H | E) = P\begin{pmatrix} \textit{Hipótesis} \\ \textit{dada} \\ \textit{la evidencia} \end{pmatrix}$$

CLASIFICADOR BAYESIANO INGENUO

Teorema de Bayes

El primer número relevante es la probabilidad de que la hipótesis se cumpla antes de considerar la nueva evidencia. En el ejemplo, eso fue $1/21$, que surgió de considerar la proporción de agricultores y bibliotecarios en la población general. Esto se conoce como el *a priori*.

$$P(H) = 1/21$$

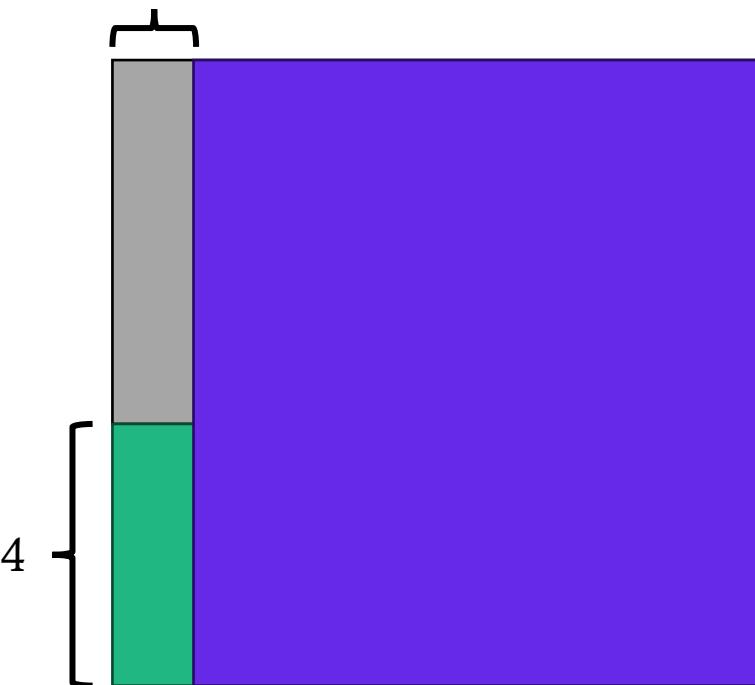


CLASIFICADOR BAYESIANO INGENUO

Teorema de Bayes

Después de eso, necesitábamos considerar la proporción de bibliotecarios que encajaban en esta descripción. Esa proporción es la probabilidad de que veamos la evidencia, dado que la hipótesis es cierta. Es decir, $P(E|H)$.

$$P(H) = 1/21$$



CLASIFICADOR BAYESIANO INGENUO

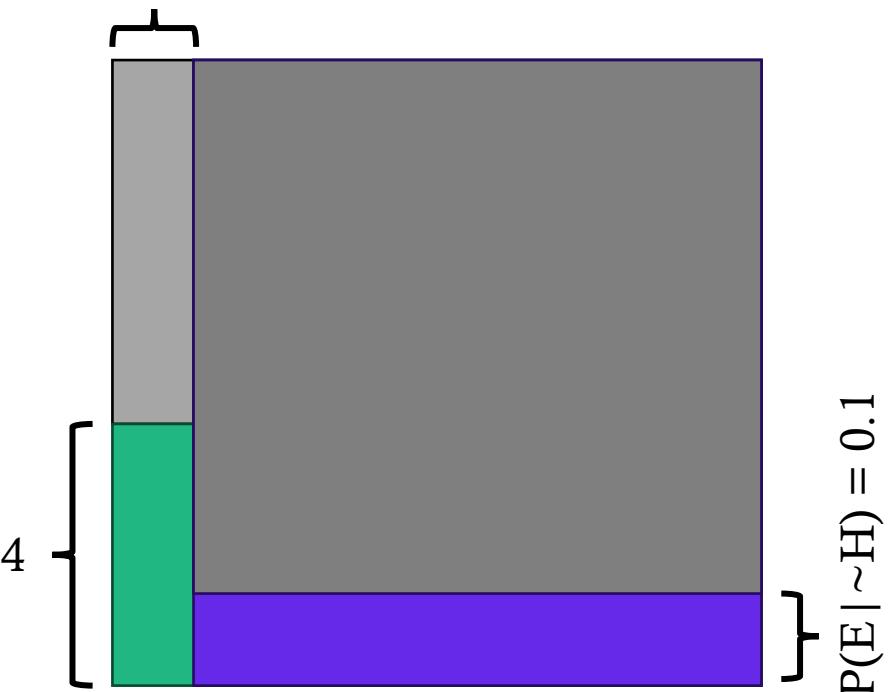
Teorema de Bayes

De manera similar, necesitamos saber qué parte del otro lado de nuestro espacio incluye evidencia.

Esa es la probabilidad de ver la evidencia dado que nuestra hipótesis no es cierta.

En el ejemplo, la probabilidad de que alguien que no sea bibliotecario coincida con la descripción de Esteban.

$$P(H) = 1/21$$



CLASIFICADOR BAYESIANO INGENUO

Teorema de Bayes

La probabilidad de que nuestra hipótesis del bibliotecario sea cierta dada la evidencia es el número total de bibliotecarios que se ajustan a la evidencia, 4, dividido por el número total de personas que se ajustan a la evidencia, 24.

$$P(H | E) = \frac{\text{[green bar]}}{\text{[green bar]} + \text{[purple bar]}}$$

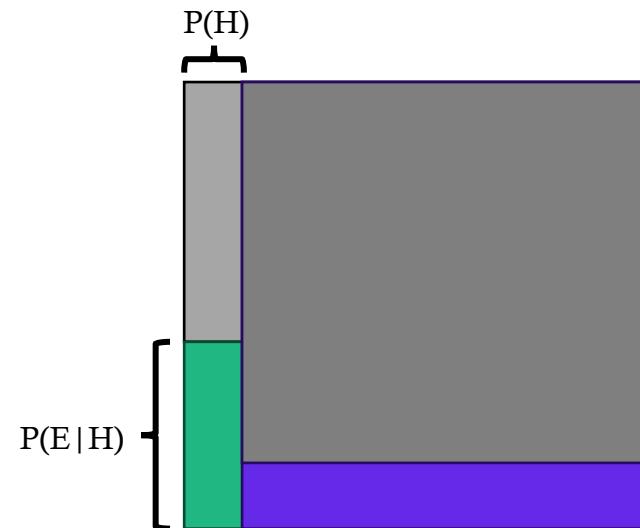
CLASIFICADOR BAYESIANO INGENUO

Teorema de Bayes

¿De dónde viene ese 4?

Es el número total de personas, multiplicado por la **probabilidad a priori de ser bibliotecario**, lo que es un total de **10 bibliotecarios**, multiplicado **por la probabilidad de que uno de ellos se ajuste a la evidencia**.

$$\text{[green bar]} = (\# \text{personas}) \mathbf{P(H)} \mathbf{P(E|H)}$$



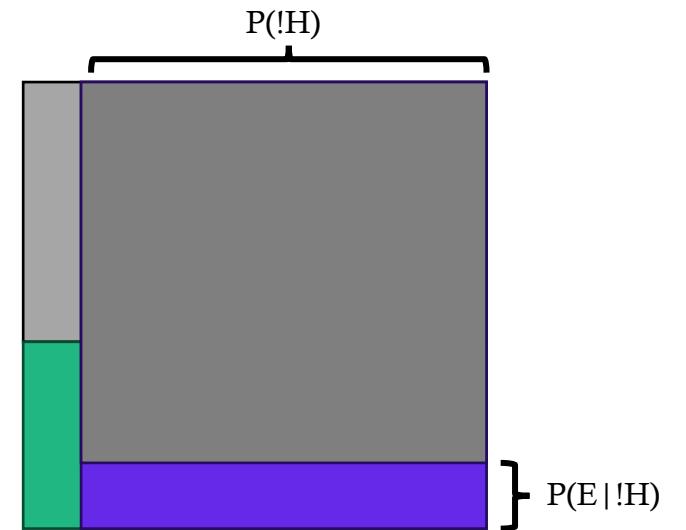
CLASIFICADOR BAYESIANO INGENUO

Teorema de Bayes

Ese mismo número aparece en el denominador. Pero también se necesita agregar otro número, que en nuestro ejemplo fue 20.

Eso surgió del número total de personas multiplicado por la **proporción que no son bibliotecarios**, multiplicado por **la proporción de aquellos que se ajustan a la evidencia**.

$$\text{[purple box]} = (\# \text{personas}) \mathbf{P}(\text{!H}) \mathbf{P}(\mathbf{E} | \text{!H})$$



CLASIFICADOR BAYESIANO INGENUO

Teorema de Bayes

Una vez que tenemos estos números, podemos ensamblarlos en la relación del teorema de Bayes:

$$\frac{(\# \text{Personas})P(H)P(E|H)}{(\# \text{Personas})P(H)P(E|H) + (\# \text{Personas})P(\neg H)P(E|\neg H)}$$

$$\frac{P(H)P(E|H)}{P(H)P(E|H) + P(\neg H)P(E|\neg H)}$$

A menudo se ve ese denominador escrito de manera más simple como $P(E)$, la probabilidad total de ver la evidencia. En la práctica, para calcularlo, casi siempre hay que descomponerlo en el caso en el que la hipótesis es cierta y en el que no lo es.

CLASIFICADOR BAYESIANO INGENUO

Teorema de Bayes

$$P(H|E) = \frac{P(H)P(E|H)}{P(E)} = \frac{P(H)P(E|H)}{P(H)P(E|H) + P(!H)P(E|!H)}$$

Posteriori



Este nuevo valor es lo que se conoce a posteriori, el cual es la creencia sobre la hipótesis luego de ver la evidencia.

CLASIFICADOR BAYESIANO INGENUO

Una de las aplicaciones de este teorema es el denominado clasificador bayesiano ingenuo.

Este clasificador utiliza la probabilidad de observar atributos, dado un resultado, para estimar la probabilidad de observar el resultado y_j , dado un conjunto de atributos.

CLASIFICADOR BAYESIANO INGENUO

Para entender su funcionamiento, usemos un caso de ejemplo...

Queremos armar un clasificador de **que prediga si un estudiante va a aprobar un examen**, dado:

- Tiempo de estudio: Este atributo representa la cantidad de tiempo que un estudiante pasa estudiando.
 - Bajo tiempo de estudio
 - Moderado tiempo de estudio
 - Alto tiempo de estudio
- Método de estudio: Este atributo representa el método utilizado por el estudiante para estudiar.
 - Solo lectura
 - Solo problemas prácticos
 - Lectura y problemas prácticos
- Puntación de exámenes anteriores: Este atributo representa el rendimiento del estudiante en exámenes anteriores
 - Baja puntuaciones
 - Promedio
 - Alta puntuaciones

CLASIFICADOR BAYESIANO INGENUO

Tenemos un dataset pequeño:

Tiempo de estudio	Método de estudio	Puntuación	Resultado
Bajo	Lectura	Bajo	Desaprobó
Bajo	Prob. Pract.	Alta	Aprobó
Moderado	Lectura y Prac.	Promedio	Aprobó
Alto	Lectura y Prac.	Alta	Aprobó
Alto	Lectura	Alta	Desaprobó
Bajo	Lectura y Prac.	Baja	Desaprobó
Alto	Prob. Pract.	Alta	Aprobó
Moderado	Lectura	Alta	Aprobó
Moderado	Lectura y Prac.	Promedio	Aprobó
Moderado	Prob. Pract.	Bajo	Desaprobó

CLASIFICADOR BAYESIANO INGENUO

Para cada uno de los atributos, construyamos las tablas de frecuencia:

Tabla de frecuencia		Resultado	
		Aprobó	Desaprobó
Tiempo de estudio	Bajo	1	2
	Moderado	3	1
	Alto	2	1

Tabla de frecuencia		Resultado	
		Aprobó	Desaprobó
Método de estudio	Lectura	1	2
	Practica	2	1
	L y P	3	1

Tabla de frecuencia		Resultado	
		Aprobó	Desaprobó
Puntuación	Bajo	0	3
	Promedio	2	0
	Alto	4	1

CLASIFICADOR BAYESIANO INGENUO

Para cada uno de los atributos, construyamos las tablas de frecuencia:

$P(E)$

Tabla de frecuencia		Resultado	
Tiempo de estudio	Aprobó	Desaprobó	
	1	2	
	3	1	
Alto	2	1	

$P(E)$

Tabla de frecuencia		Resultado	
Puntuación	Aprobó	Desaprobó	
	0	3	
	2	0	
Alto	4	1	

$P(H)$

Tabla de frecuencia		Resultado	
Método de estudio	Aprobó	Desaprobó	
	1	2	
	2	1	
L y P	3	1	

$P(E)$

$$P(H|E) = \frac{P(H)P(E|H)}{P(E)}$$

CLASIFICADOR BAYESIANO INGENUO

Para cada uno de los atributos, construyamos las tablas de frecuencia:

$$P(H)$$

Vamos a aplicar el teorema, para cada atributo como si fueran atributos independientes (por lo que multiplicamos a las probabilidades). Esto lo asumimos de forma *ingenua*

$P(E)$	estudio	Moderado	3	1
		Alto	2	1

$P(E)$	estudio	Practica	2	1
		L y P	3	1

$P(E)$	Tabla de frecuencia		Resultado		
	Puntuación	Bajo	Aprobó	Desaprobó	
		0	3		
		2	0		
		4	1		

$$P(H|E) = \frac{P(H)P(E|H)}{P(E)}$$

CLASIFICADOR BAYESIANO INGENUO

Ahora construyamos la tabla de probabilidad:

Tabla de frecuencia		Resultado		
		Aprobó	Desaprobó	
Tiempo de estudio	Bajo	1	2	3
	Moderado	3	1	4
	Alto	2	1	3
		6	4	10

Tabla de frecuencia		Resultado		
		Aprobó	Desaprobó	
Método de estudio	Lectura	1	2	3
	Practica	2	1	3
	L y P	3	1	4
		6	4	10

Tabla de frecuencia		Resultado		
		Aprobó	Desaprobó	
Puntuación	Bajo	0	3	3
	Promedio	2	0	2
	Alto	4	1	5
		6	4	10

CLASIFICADOR BAYESIANO INGENUO

Ahora construyamos la tabla de probabilidad:

Tabla de frecuencia		Resultado		
		Aprobó	Desaprobó	
Tiempo de estudio	Bajo	1/6	1/2	3/10
	Moderado	1/2	1/4	2/5
	Alto	1/3	1/4	3/10
		3/5	2/5	10

Tabla de frecuencia		Resultado		
		Aprobó	Desaprobó	
Método de estudio	Lectura	1/6	1/2	3/10
	Practica	1/3	1/4	3/10
	L y P	1/2	1/4	2/5
		3/5	2/5	10

Tabla de frecuencia		Resultado		
		Aprobó	Desaprobó	
Puntuación	Bajo	0	3/4	3/10
	Promedio	1/3	0	1/5
	Alto	2/3	1/4	1/2
		3/5	2/5	10

CLASIFICADOR BAYESIANO INGENUO

Ahora construyamos la tabla de probabilidad:

Tabla de frecuencia		Resultado		
		Aprobó	Desaprobó	
Tiempo de estudio	Bajo	1/6	1/2	3/10
	Moderado	1/2	1/4	2/5
	Alto	1/3	1/4	3/10
		3/5	2/5	10

$$P(H) = P(Aprobo) = 3/5 = 0.6$$

$$P(!H) = P(Desaprobo) = 2/5 = 0.4$$

$$P(E) = P(Bajo \ tiempo \ de \ estudio) = 3/10 = 0.3$$

$$P(E|H) = P(Bajo|Aprobo) = 1/6 = 0.17$$

$$P(E|!H) = P(Bajo|Desaprobo) = 1/2 = 0.5$$

Aplicaremos el teorema de Bayes:

$$P(H|E) = P(Aprobo|Bajo) = \frac{P(E|H)P(H)}{P(E)} = \frac{P(Bajo|Aprobo)P(Aprobo)}{P(Bajo \ tiempo \ de \ estudio)} = \frac{0.17 * 0.6}{0.3} = 0.34$$

$$P(!H|E) = P(Desaprobo|Bajo) = \frac{P(E|!H)P(!H)}{P(E)} = \frac{P(Bajo|Desaprobo)P(Desaprobo)}{P(Bajo \ tiempo \ de \ estudio)} = \frac{0.5 * 0.4}{0.3} = 0.67$$

CLASIFICADOR BAYESIANO INGENUO

Ahora construyamos la tabla de probabilidad:

Tabla de frecuencia		Resultado		
		Aprobó	Desaprobó	
Tiempo de estudio	Bajo	1/6	1/2	3/10
	Moderado	1/2	1/4	2/5
	Alto	1/3	1/4	3/10

$$P(H) = P(Aprobo) = 3/5 = 0.6$$

$$P(!H) = P(Desaprobo) = 2/5 = 0.4$$

$$P(E) = P(Bajo \ tiempo \ de \ estudio) = 3/10 = 0.3$$

$$P(E|H) = P(Bajo|Aprobo) = 1/6 = 0.17$$

$$P(E|!H) = P(Bajo|Desaprobo) = 1/2 = 0.5$$

Basado en esto, podemos concluir que es más probable que un alumno desaproveche si estudia poco

Apliquemos el teorema

$$P(H|E) = P(Aprobo|Bajo) = \frac{P(E|H)P(H)}{P(E)} = \frac{P(Bajo|Aprobo)P(Aprobo)}{P(Bajo \ tiempo \ de \ estudio)} = \frac{0.17 * 0.6}{0.3} = 0.34$$

$$P(!H|E) = P(Desaprobo|Bajo) = \frac{P(E|!H)P(!H)}{P(E)} = \frac{P(Bajo|Desaprobo)P(Desaprobo)}{P(Bajo \ tiempo \ de \ estudio)} = \frac{0.5 * 0.4}{0.3} = 0.67$$

CLASIFICADOR BAYESIANO INGENUO

Se asume que los tres atributos son independientes entre sí, por lo que podemos encontrar la probabilidad de que un alumno apruebe si dedico muchas horas de estudio, realizo lecturas y trabajos prácticos, y tenía puntuaciones altas en exámenes anteriores.

$P(E)$ equivale a *Estudio = Alto, Metodo = L y P, Puntuacion = Alta*

$P(H)$ equivale a $P(\text{Aprueba})$

$$P(H|E) = \frac{P(\text{Alto}|\text{Aprueba})P(LyP|\text{Aprueba})P(\text{Alta}|\text{Aprueba})P(\text{Aprueba})}{P(\text{Alto})P(LyP)P(\text{Alto})}$$

$$P(H|E) = \frac{0.333 * 0.5 * 0.667 * 0.6}{0.3 * 0.4 * 0.5} = 1.11$$

CLASIFICADOR BAYESIANO INGENUO

Calculemos ahora quienes desaprueban con las mismas condiciones

$P(E)$ equivale a *Estudio = Alto, Metodo = L y P, Puntuacion = Alta*

$P(\neg H)$ equivale a *P(Desaprueba)*

$$P(\neg H|E) = \frac{P(\text{Alto}|\text{Desaprueba})P(\text{LyP}|\text{Desaprueba})P(\text{Alta}|\text{Desaprueba})P(\text{Desaprueba})}{P(\text{Alto})P(\text{LyP})P(\text{Alto})}$$

$$P(\neg H|E) = \frac{0.25 * 0.25 * 0.25 * 0.4}{0.3 * 0.4 * 0.5} = 0.10$$

CLASIFICADOR BAYESIANO INGENUO

Normalizamos:

$$\text{Suma} = P(H|E) + P(!H|E) = 1.11 + 0.10 = 1.21$$

$$P(H|E) = \frac{1.11}{1.21} = 0.92$$

$$P(!H|E) = \frac{0.1}{1.21} = 0.08$$

Dado que, si un alumno le dedica muchas horas, estudia practicando y leyendo, y tiene buenas notas, va a aprobar el examen.

CLASIFICADOR BAYESIANO INGENUO

Normalizamos:

$$\text{Sum} = P(\text{HUE}) + P(\text{LUCE}) = 1.11 + 0.10 = 1.21$$

En el contexto del clasificador Naive Bayes, las probabilidades condicionales pueden sumar más de uno debido a la independencia asumida entre las características, lo que puede resultar en una sobreestimación.

Aunque una probabilidad no puede ser mayor que uno, en este contexto, este valor no invalida el resultado, pero para interpretarlas como una distribución de probabilidad válida, normalizamos.

Dado que $P(\text{HUE}) / \text{Sum} = 1.11 / 1.21 = 0.91$ y

tiene buenas notas, va a aprobar el examen.

CLASIFICADOR BAYESIANO INGENUO

En general, dado que el denominador es siempre el mismo, lo que se calcula es solo el numerador, y el clasificador clasifica con la clase que es mayor:

- $P(H|E) = 0.333 * 0.5 * 0.667 * 0.6 = 0.06663$
- $P(!H|E) = 0.25 * 0.25 * 0.25 * 0.4 = 0.00625$

CLASIFICADOR BAYESIANO INGENUO

A veces si no tenemos valores en alguna combinación ya sea porque nuestro dataset es chico o es falta de valores, nos puede afectar el resultado, por ejemplo,

Si queremos ver si el alumno aprueba si estudio mucho tiempo, realizo practica y lecturas y venia con puntuación baja en exámenes anteriores:

- $P(H|E) = 0.333 * 0.5 * \mathbf{0} * 0.6 = \mathbf{0}$

Vemos que aquel que es cero, tiene demasiada fuerza para clasificar. Podemos mitigar este cambio, sumando uno a cada uno de los valores de la tabla de frecuencias.

CLASIFICADOR BAYESIANO INGENUO

Tabla de frecuencia		Resultado	
		Aprobó	Desaprobó
Tiempo de estudio	Bajo	1+1=2	2+1=3
	Moderado	3+1=4	1+1=2
	Alto	2+1=3	1+1=2

Tabla de frecuencia		Resultado	
		Aprobó	Desaprobó
Método de estudio	Lectura	1+1=2	2+1=3
	Practica	2+1=3	1+1=2
	L y P	3+1=4	1+1=2

Tabla de frecuencia		Resultado	
		Aprobó	Desaprobó
Puntuación	Bajo	0+1=1	3+1=4
	Promedio	2+1=3	0+1=1
	Alto	4+1=5	1+1=2

CLASIFICADOR BAYESIANO INGENUO

Tabla de frecuencia		Resultado		
		Aprobó	Desaprobó	
Tiempo de estudio	Bajo	2/9	3/7	5/16
	Moderado	4/9	2/7	6/16
	Alto	3/9	2/7	5/16
		9/16	7/16	16

Tabla de frecuencia		Resultado		
		Aprobó	Desaprobó	
Método de estudio	Lectura	2/9	3/7	5/16
	Practica	3/9	2/7	5/16
	L y P	4/9	2/7	6/16
		9/16	7/16	16

Tabla de frecuencia		Resultado		
		Aprobó	Desaprobó	
Puntuación	Bajo	1/9	4/7	5/16
	Promedio	3/9	1/7	4/16
	Alto	5/9	2/7	6/16
		9/16	7/16	16

CLASIFICADOR BAYESIANO INGENUO

Entonces ahora podemos calcular,

Si queremos ver si el alumno aprueba si estudio mucho tiempo, realizo practica y lecturas y venia con puntuación baja en exámenes anteriores:

- $P(H|E) = 0.333 * 0.444 * 0.111 * 0.5625 = 0.0092$
- $P(!H|E) = 0.286 * 0.286 * 0.286 * 0.4375 = 0.0102$

CLASIFICADOR BAYESIANO INGENUO

Entonces ahora podemos calcular,

Si queremos ver si el alumno aprueba si estudio mucho tiempo, realizo practica y lecturas y venia con puntuación baja en exámenes anteriores:

- $P(H|E) = 0.333 * 0.444 * 0.111 * 0.5625 = 0.0092$
- $P(!H|E) = 0.286 * 0.286 * 0.286 * 0.4375 = 0.0102$

Este número que agregamos se llama hiperparámetro ν (alfa).

CLASIFICADOR BAYESIANO INGENUO

El clasificador bayesiano ingenuo funciona para **variables categóricas**. Para **variables numéricas** podemos tratarlas de dos formas:

- Discretizarlas en contenedores o rangos.
- Asumir que poseen una distribución, y usar esa distribución para calcular la probabilidad.