CLASIFICACIÓN

APRENDIZAJE DE MAQUINA I - CEIA - FIUBA

Dr. Ing. Facundo Adrián Lucianna

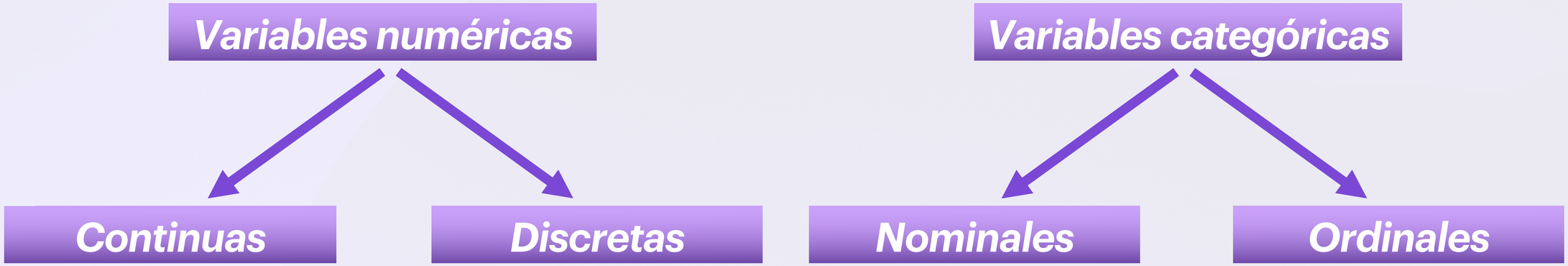
Dr. Ing. Álvaro Gabriel Pizá

REPASO CLASE ANTERIOR

- Tipos de variables y variables dummy
- Análisis de regresión
  - Regresión lineal simple y multiple
  - Regresión polinómica
  - Métricas de evaluación
- Construcción de un modelo

TIPOS DE VARIABLES

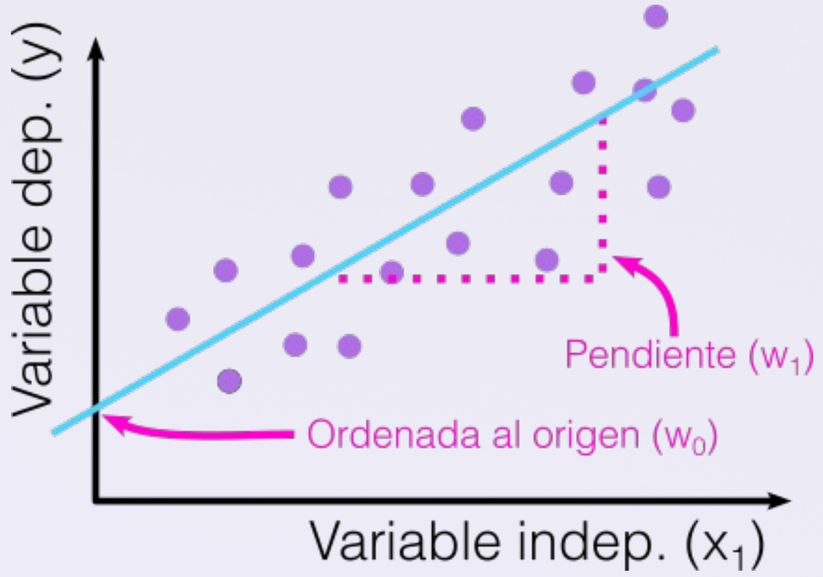
Una parte importante, es identificar el tipo de variable



REGRESIÓN

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

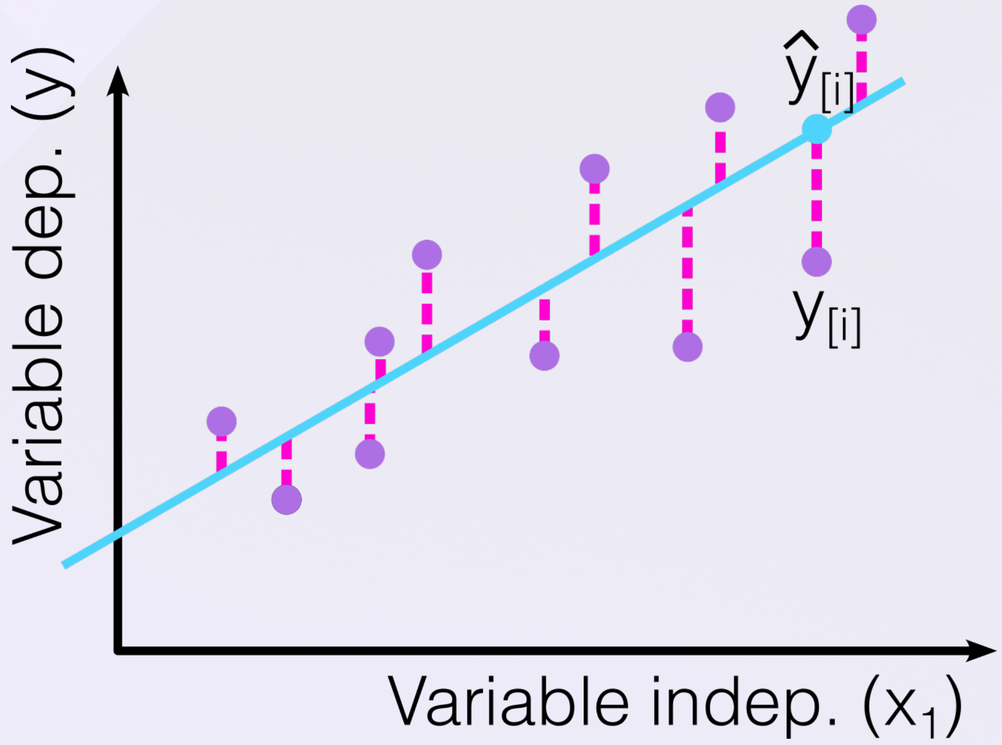
$$y(x_1) = w_0 + w_1 x_1$$



$$y(\widehat{x}) = w_0 + w_1 x_1 + \ldots + w_d x_d$$

REGRESIÓN LINEAL

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



MÉTRICAS DE EVALUACIÓN

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

CLASIFICACIÓN

CLASIFICACIÓN

Es mas 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 medicas.
- 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 numero 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 respuesta cualitativas?

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

- 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 ademas 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, hay 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

$$p(\mathrm{ACV} = 1|X) = w_0 + w_1x_1 + w_2x_2 + \ldots + w_nx_n$$

REGRESIÓN LOGISTICA

REGRESIÓN LOGISTICA

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

$$P(Y = k \mid \hat{X})$$

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:

$$P(Y = 0 \mid \hat{X})$$

$$P(Y = 1 \mid \hat{X})$$

REGRESIÓN LOGISTICA

Pero además, en el caso de dos clases:

$$P(Y = 1 \mid \hat{X}) = 1 - P(Y = 0 \mid \hat{X})$$

Por lo que podemos simplificar la notación:

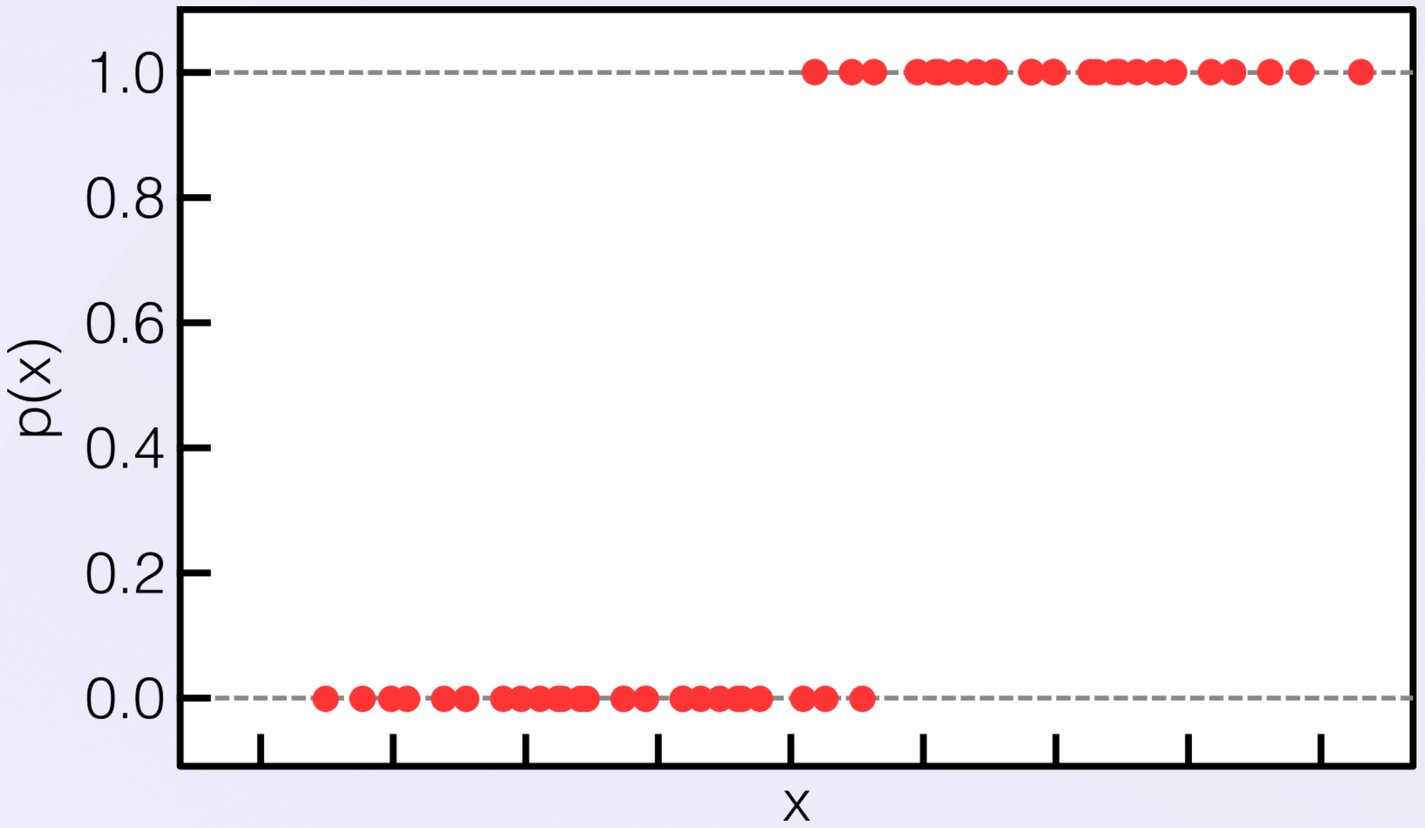
$$P(Y = 1 \mid \hat{X}) = p(\hat{X})$$

Las probabilidades son valores que van entre 1 y 0.

Además, la hagamos mas simple, el caso de una sola feature: $p(x)$

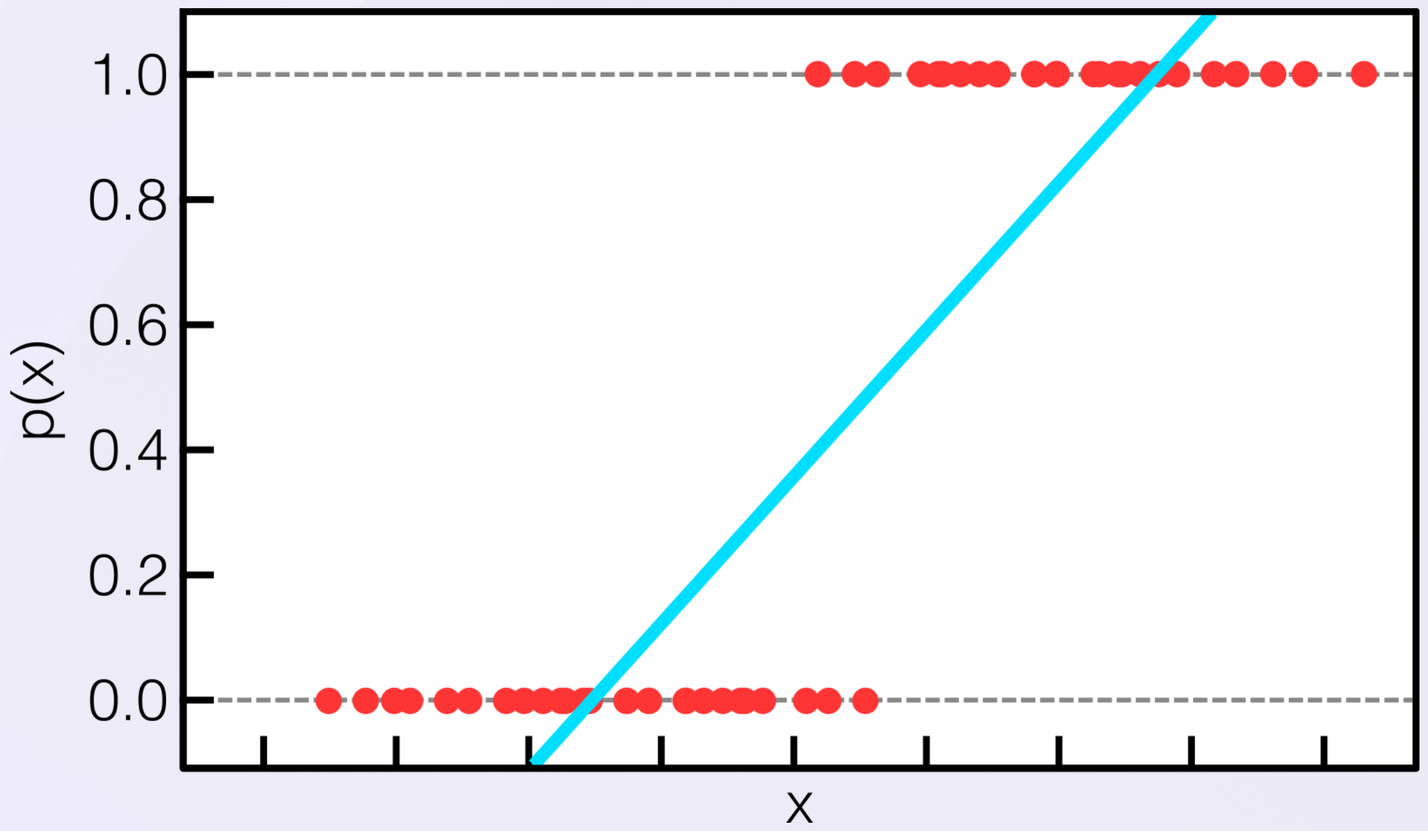
REGRESIÓN LOGISTICA

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



REGRESIÓN LOGISTICA

Podemos usar una regresión lineal para estimar la probabilidad $p(x) = w_0 + w_1 x$



REGRESIÓN LOGISTICA

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.

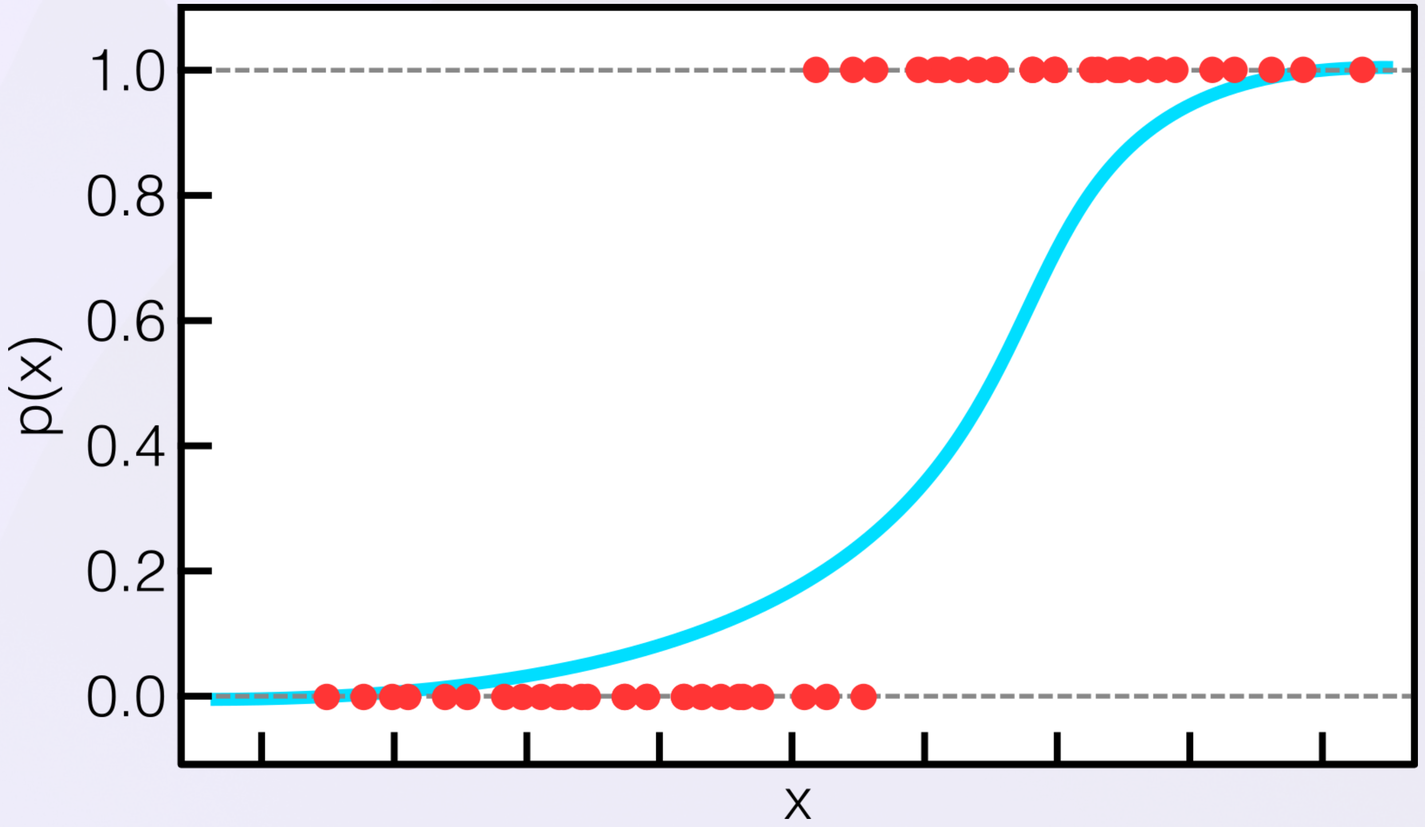
Para evitar este problema, se debe modelar a p(x) usando una función que nos asegure que siempre tendremos valores entre 0 y 1.

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

$$p(x) = \frac{e^{w_0+w_1x_1}}{1+e^{w_0+w_1x_1}}$$

REGRESIÓN LOGISTICA

Lo que visualmente se observa:



REGRESIÓN LOGISTICA

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 $p(x) = \frac{e^{w_0+w_1x_1}}{1+e^{w_0+w_1x_1}}$, llegamos a:

$$\frac{p(x)}{1-p(x)} = e^{w_0+w_1x_1}$$

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 LOGISTICA

Si aplicamos el logaritmo de ambos lados:

$$ln\left(\frac{p(x)}{1-p(x)}\right) = w_0 + w_1 x_1$$

Que es la función **logit**. En la regresión logística, es una relación lineal.

Esto nos permite que ver qué incremento de una unidad de x_1, cambia el logit en w_1 unidades. Equivalentemente multiplica la chance en e^{w1}.

Pero, la cantidad de p(x) al aumentar x_1 en una unidad, al no ser lineal, depende del valor actual de x_1.

REGRESIÓN LOGISTICA - AJUSTE

Para buscar los coeficientes (w_0 y w_1), es decir entrenar, lo hacemos con un método llamado estimación por **máxima verosimilitud**.

La básica intuición detrás de la máxima verosimilitud es que buscamos estimaciones para w_0 y w_1 tales que la probabilidad prevista $p(x_i)$ de cada caso, utilizando $p(x) = \frac{e^{w_0+w_1x_1}}{1+e^{w_0+w_1x_1}}$, corresponda lo más cerca posible al estado observado.

En otras palabras, tratamos de encontrar w_0 y w_1 tales que al conectar estas estimaciones se obtenga un número cercano a uno para la clase positiva, y lo más cercano a 0 para la clase negativa.

REGRESIÓN LOGISTICA - AJUSTE

Matemáticamente la **función de verosimilitud** es:

$$\ell\left(w_0, w_1\right) = \prod_{i:y_i=1} p(x_i) \prod_{j:y_j=0} (1 - p(x_j))$$

Y buscamos encontrar los valores de w_0 y w_1 que hacen el valor máximo de esta función.

REGRESIÓN LOGISTICA MULTIPLE

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

$$ln\left(\frac{p(\widehat{X})}{1-p(\widehat{X})}\right) = w_0 + w_1x_1 + \ldots + w_nx_n \qquad p(\widehat{X}) = \frac{e^{w_0+w_1x_1+\ldots+w_nx_n}}{1+e^{w_0+w_1x_1+\ldots+w_nx_n}}$$

Y el ajuste es exactamente igual, usando **máxima verosimilitud.**


VAMOS A PRÁCTICAR UN POCO...

MÉTRICAS DE EVALUACIÓN

MÉTRICAS DE EVALUACIÓN

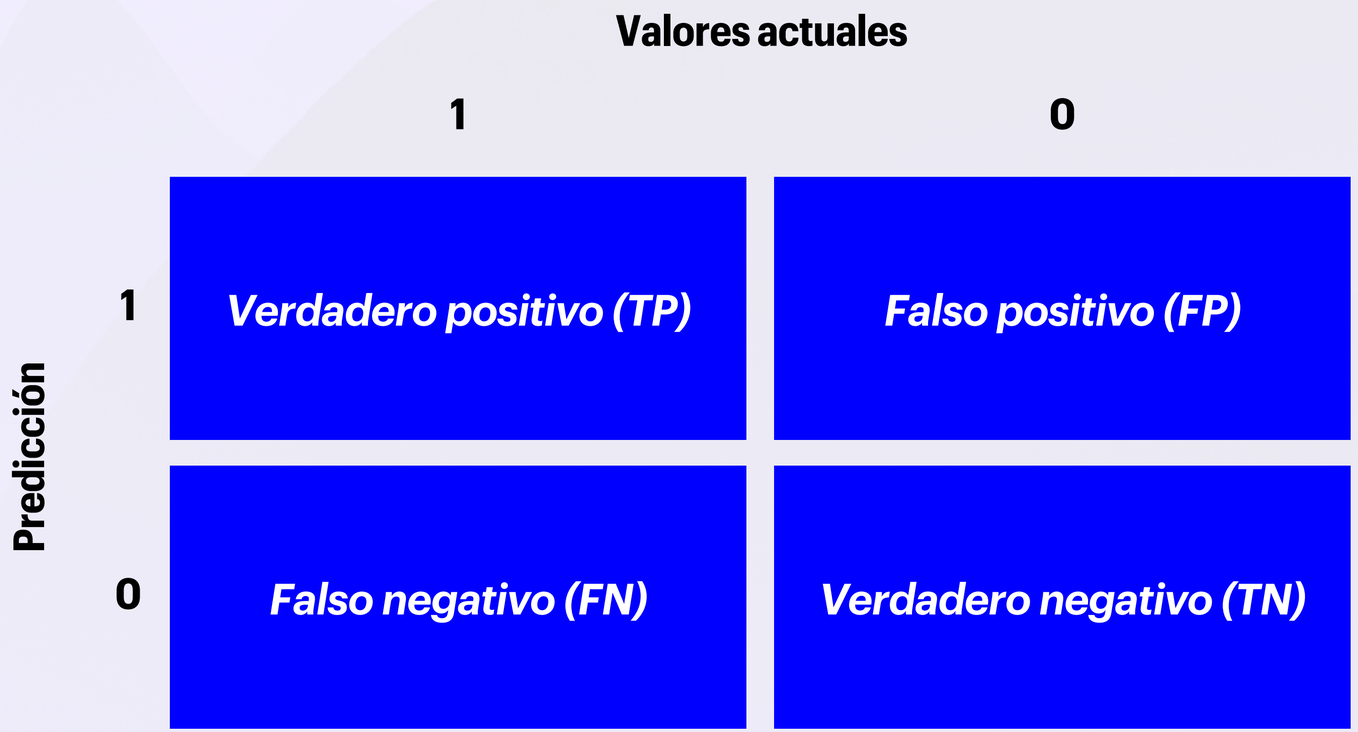
¿Ahora como clasificamos? Dado a que tenemos una probabilidad, por lo que por el momento elijamos el valor **naive**. Es decir, es 1 si P(x) > 0.5, 0 en caso contrario.

¿Y como medimos al clasificador?

Una forma es obteniendo el **porcentaje de clasificaciones correctas**. Aunque es buena información, deja afuera algunas consideraciones, tales como que clase se equivoco más.

Esto lo podemos ver con la **matriz de confusión**.

MATRIZ DE CONFUSIÓN



MATRIZ DE CONFUSIÓN

- **Verdadero positivo:** Es aquellas observaciones que clasificamos como 1 y que realmente eran 1.
- **Verdadero negativo:** Es aquellas observaciones que clasificamos como 0 y que realmente eran 0.
- **Falso positivo:** Es aquellas observaciones que clasificamos como 1 y que realmente eran 0. Este tipo de error se llaman de **tipo I**.
- **Falso negativo:** Es aquellas observaciones que clasificamos como 0 y que realmente eran 1. Este tipo de error se llaman de **tipo II**.

MATRIZ DE CONFUSIÓN

El desempeño específico de clase también es importante en medicina y biología, donde los términos **sensibilidad** y **especificidad** caracterizan el desempeño de un clasificador o prueba de detección.

- **Sensibilidad:** $TPR = \frac{TP}{P} = \frac{TP}{TP + FN} = 1 - FNR$
- **Especificidad:** $TNR = \frac{TN}{N} = \frac{TN}{TN + FP} = 1 - FPR$

La métrica del rendimiento general se llama **exactitud:**

$$ACC = \frac{TP + TN}{P + N}$$

MATRIZ DE CONFUSIÓN

Un problema de las métricas que principalmente la exactitud es sensible es el desbalance de clases:

$$ACC = \frac{TP + TN}{P + N}$$

Si tenemos mucho más de una que la otra, una clase tendrá más peso que la otra, dominando las métricas. Por lo que hay métricas que buscan compensar esto:

- **Exactitud balanceada**: $BA = \frac{TPR + TNR}{2}$

MATRIZ DE CONFUSIÓN

Otras dos métricas importantes son precisión y recuperación, y que tiene más importancia cuando la clase positiva es la que nos importa (clasificación booleana)

- **Precisión:** Es la fracción de clases positivas correctamente predichas de las que el modelo dijo que eran positivas

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recuperación:** Es la fracción de clases positivas que el modelo predijo que eran verdadero sobre el total de las verdaderas.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Es decir, la recuperación nos dice cuanta clases positivas, el modelo pudo obtener. Y precisión es de los elementos traídos por el modelo son relevantes.

MATRIZ DE CONFUSIÓN

¿Qué es mejor tener precisión o recuperación?

Podemos usar una sola métrica que pese uno u otro en función de que queremos. Si queremos medir un score que nos encuentre modelos que balanceen ambos tenemos el F_1-score:

$$F_1 = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{2TP}{2TP + FP + FN}$$

Pero podemos elegir un score que le dé prioridad uno u otro:

$$F_\beta = (1 + \beta^2)\frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$

Si $0 < \beta < 1$, recuperación es más prioritario, $\beta > 1$, precisión es más pesado.

CURVA ROC

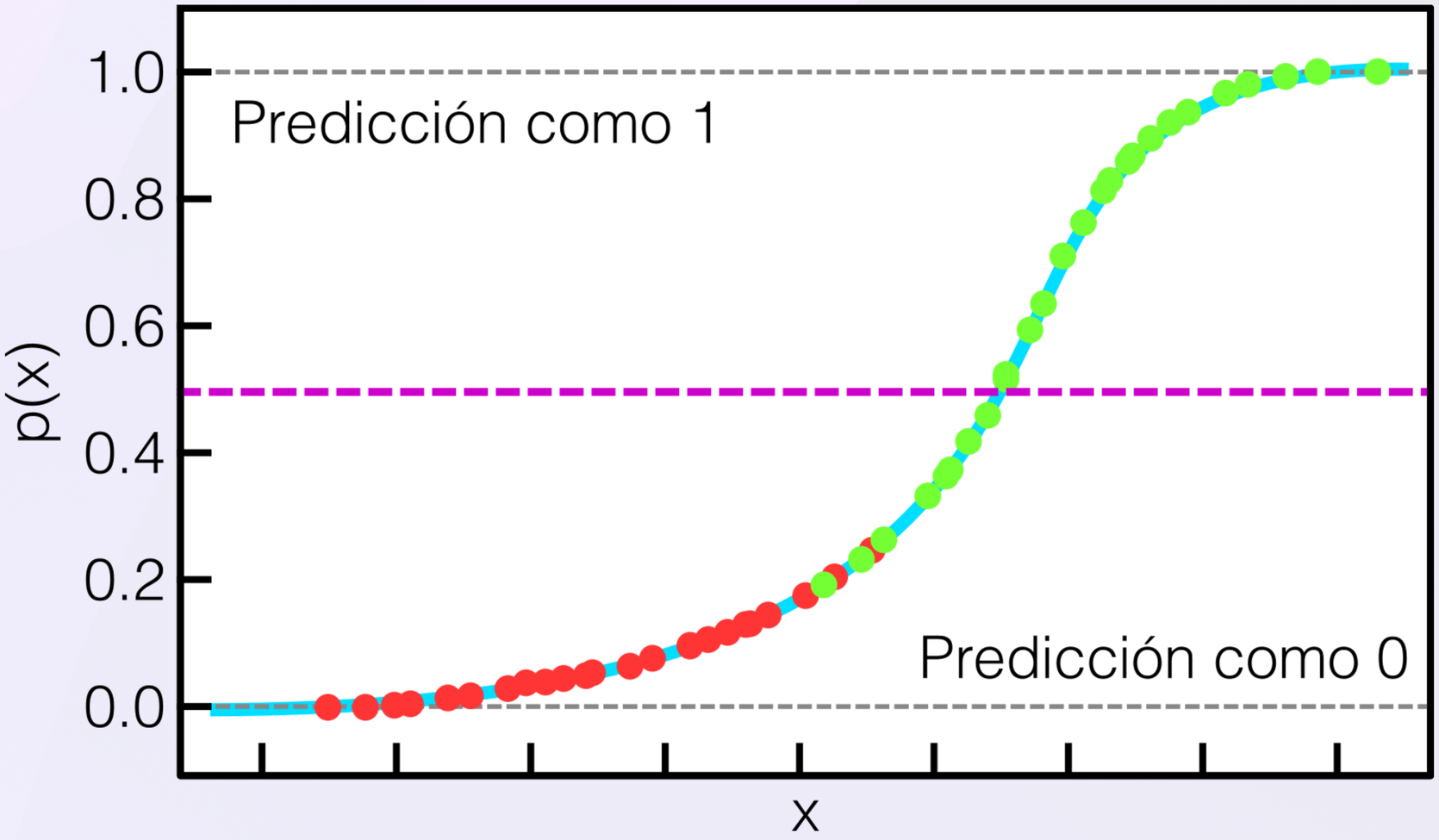
CURVA ROC

Todas estas métricas son decididas en función de la elección de umbral elegido en 0.5.

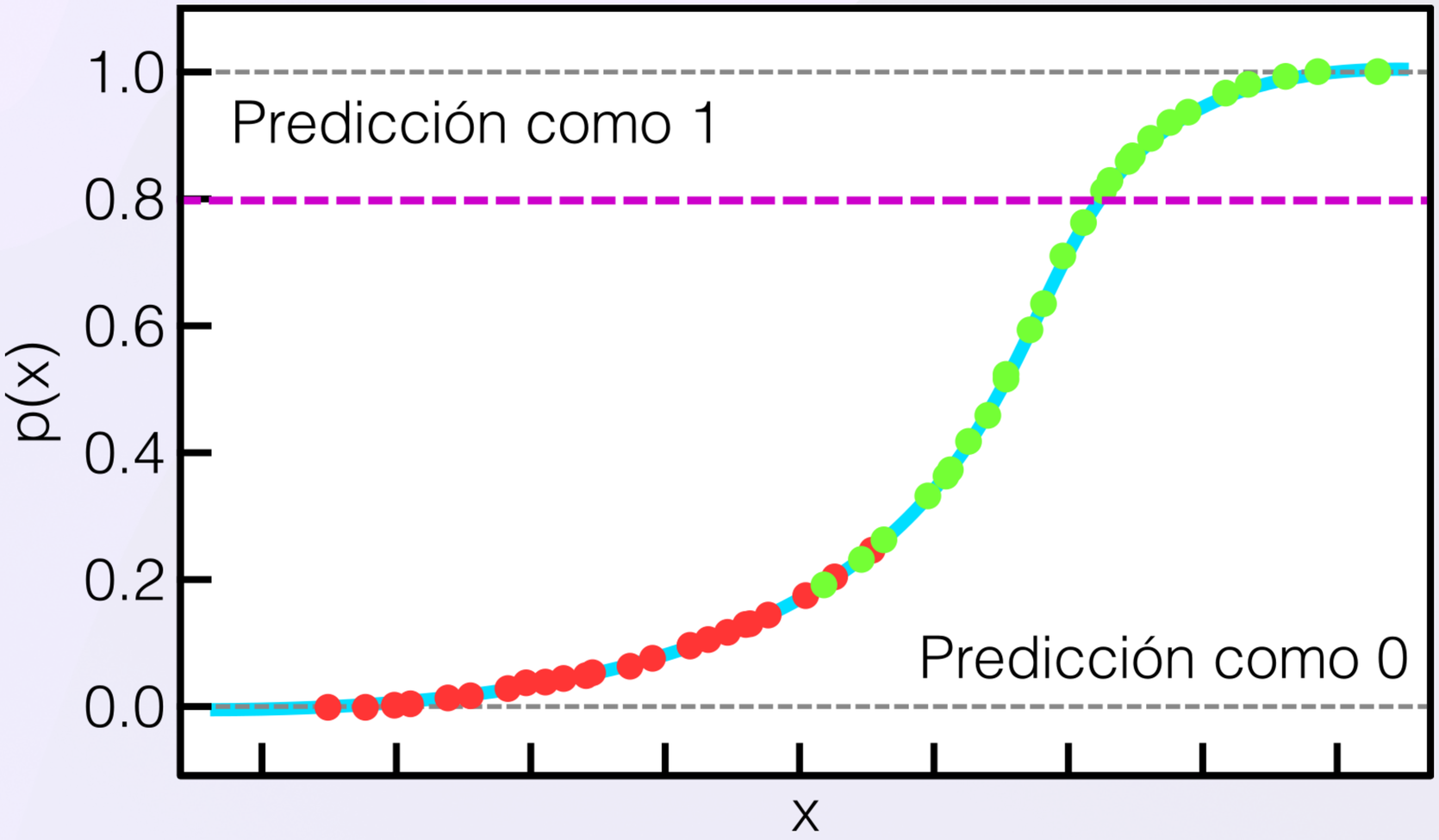
¿Pero una pregunta es, porque elegimos ese valor?

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.

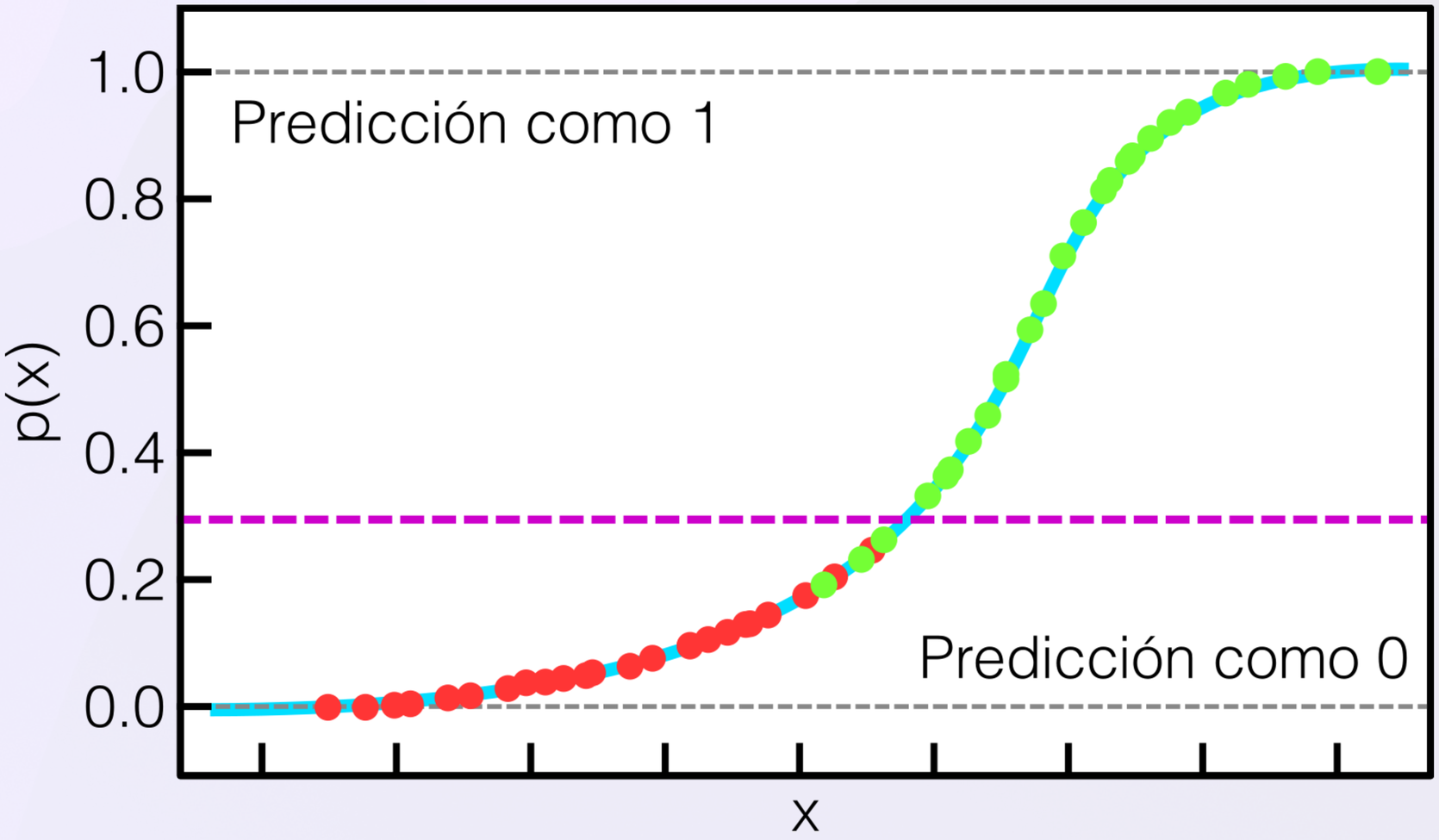
CURVA ROC



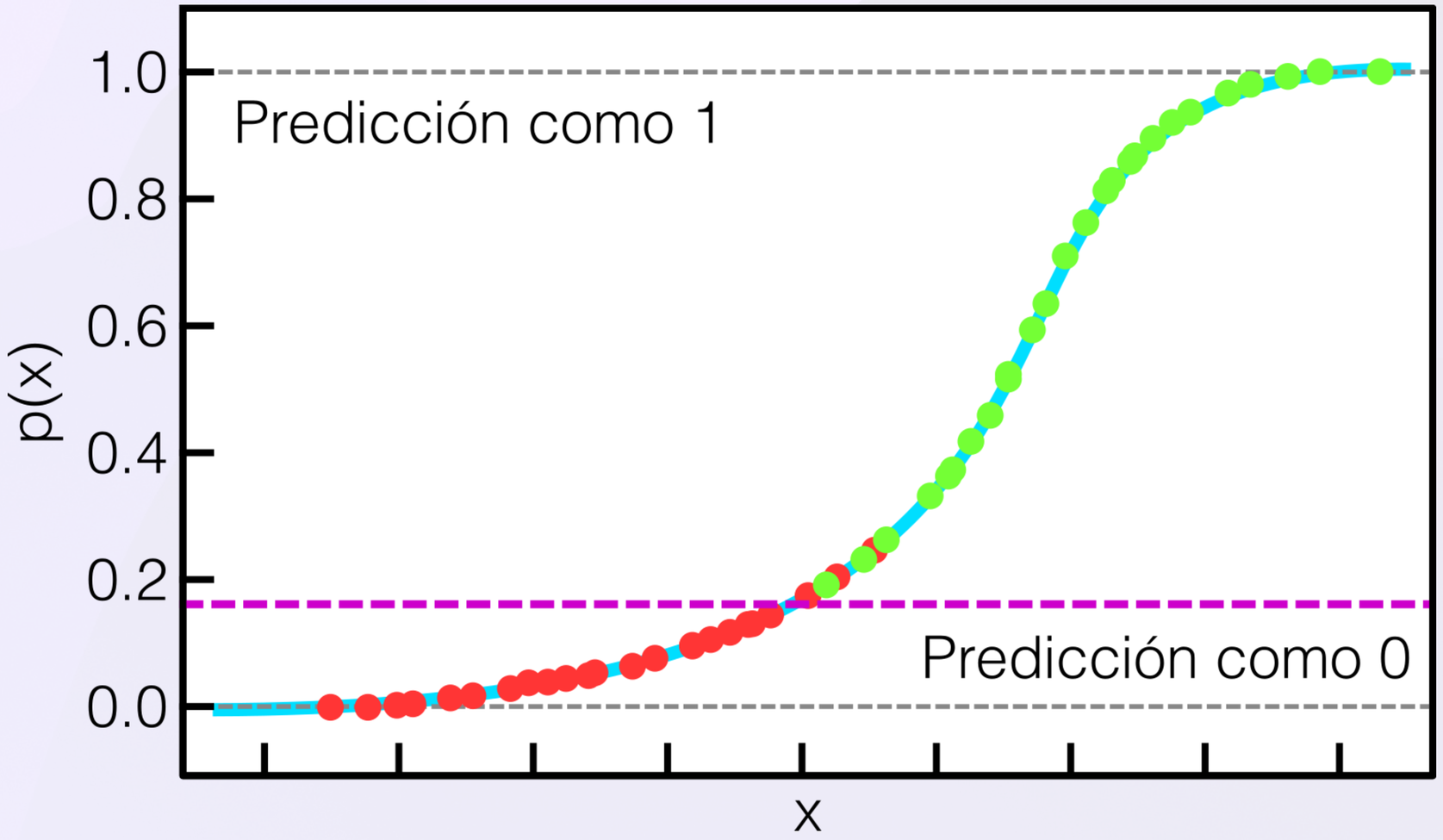
CURVA ROC



CURVA ROC



CURVA ROC

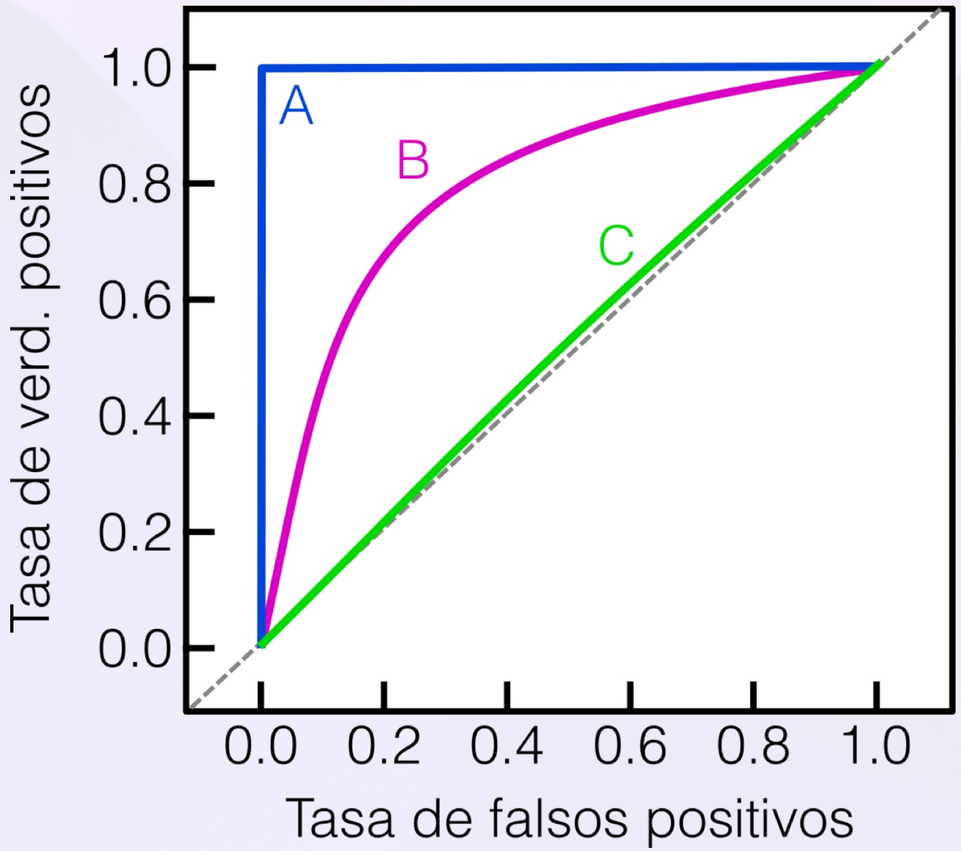


CURVA ROC

La curva ROC nos permite 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



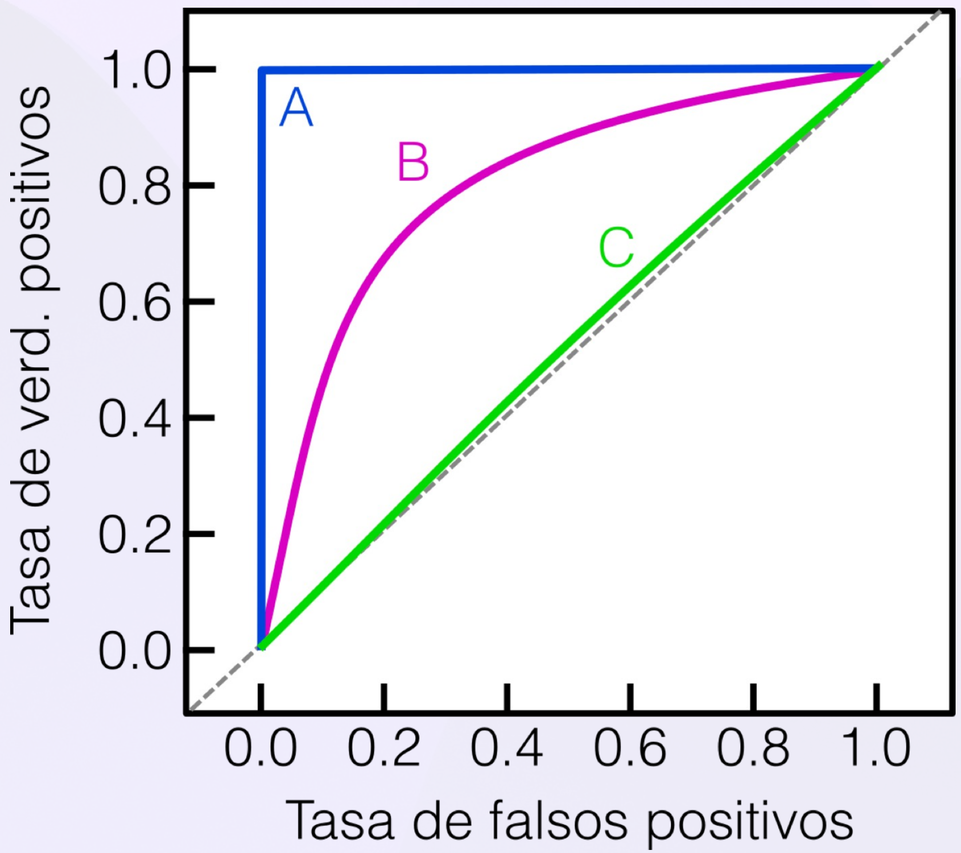
Siempre se arranca de umbral 1, donde la TPR es 0 y TFP es 0 y termina en 0 donde TVP es 1 y TFP es 1.

- **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 me permite encontrar el valor umbral que mejor resultado me dé.

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

CURVA ROC

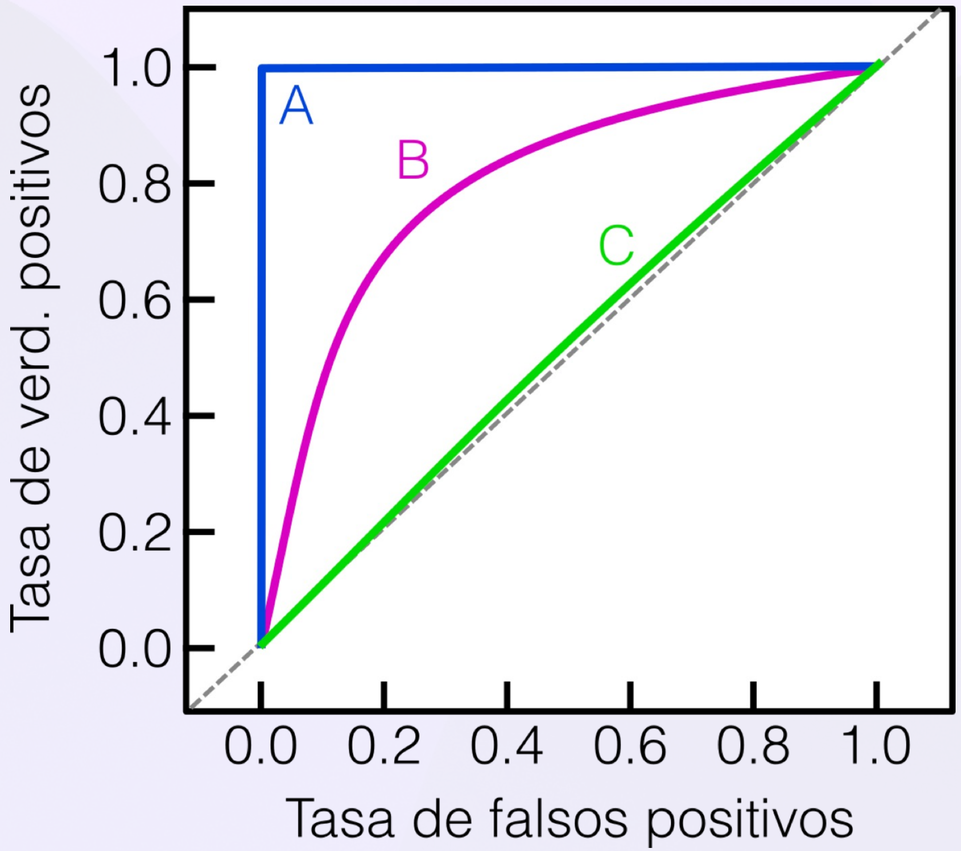


Si quiero bajar a una métrica a esta curva, podemos calcular el área bajo la curva (AUC).

- **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.


VAMOS A PRÁCTICAR UN POCO...

REGRESIÓN LOGISTICA 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 que llamamos K.

Para hacer esto, se elige una de las clases como base y sobre esta se construye la probabilidad de las demás clases:

$$P(Y = k \mid X = x) = \frac{e^{w_{k0} + w_{k1}x_1 + \ldots + w_{kn}x_n}}{1 + \sum_{l=1}^{K-1} e^{w_{l0} + w_{l1}x_1 + \ldots + w_{ln}x_n}}$$

$$P(Y = K \mid X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} e^{w_{l0} + w_{l1}x_1 + \ldots + w_{ln}x_n}}$$

Y las chances son: $ln\left(\frac{P(Y = k \mid X = x)}{P(Y = K \mid X = x)}\right) = w_{k0} + w_{k1}x_1 + \ldots + w_{kn}x_n$

REGRESIÓN LOGÍSTICA MULTI-CLASE

Que clase K se elige como base no importa. Los coeficientes van a cambiar con dos modelos entrenados con el mismo dataset, pero diferente elección de clase para la base.

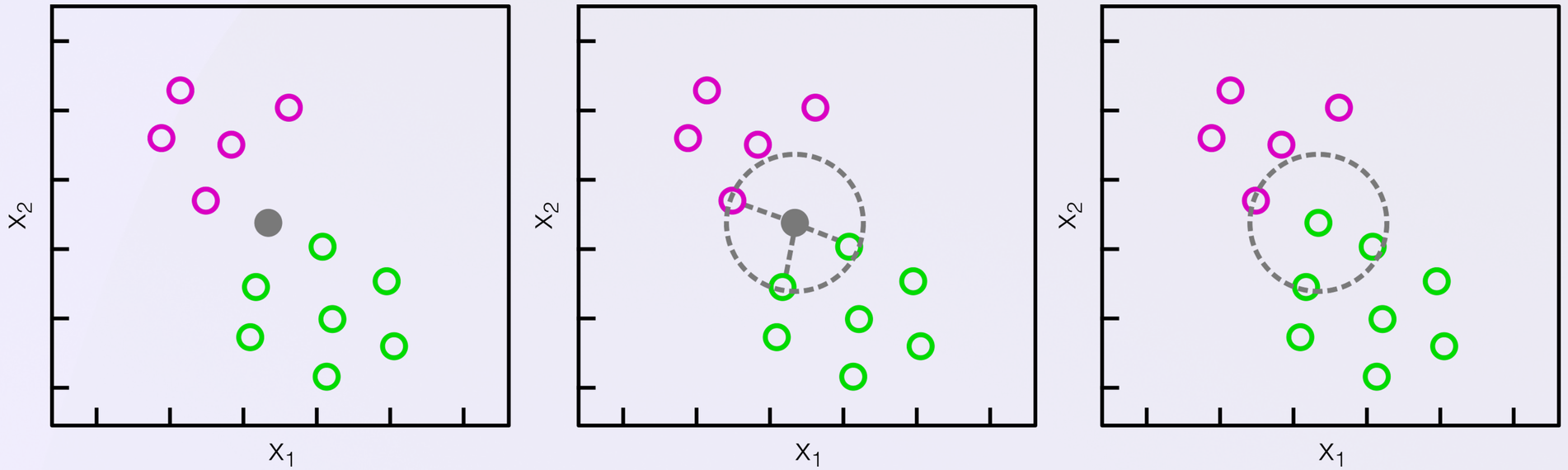
Lo importante es que los valores predichos de probabilidad se van a mantener igual.

CLASIFICADOR KNN

KNN

El clasificador de k vecinos más cercanos (KNN o k-NN), es un algoritmo que utiliza la proximidad de sus vecinos para hacer clasificaciones sobre la agrupación de un punto.

La idea se basa de la **suposición** de que se pueden encontrar puntos similares cerca uno del otro en base a votación de pluralidad (se elige la clase en función de la moda de la clase de sus vecinos).



KNN

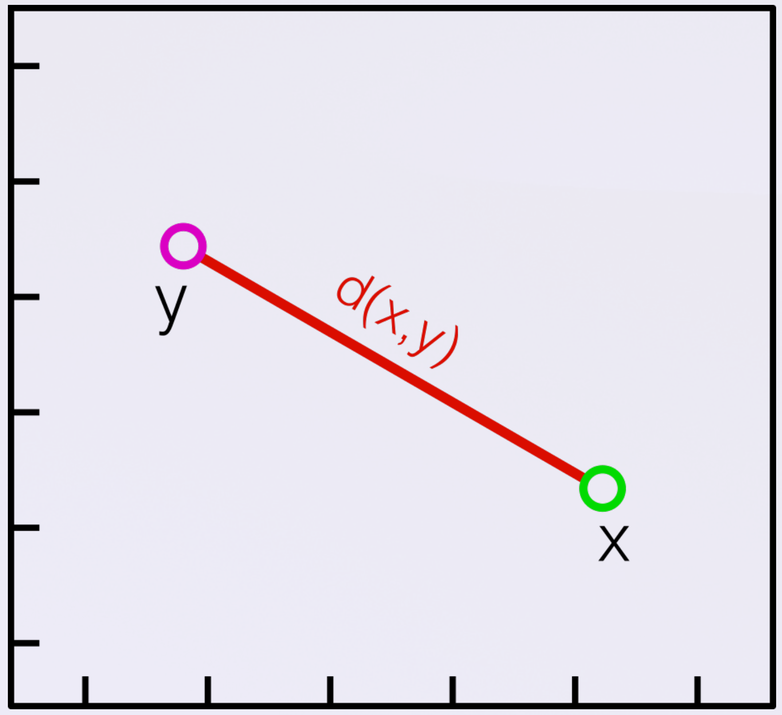
Como vimos, este algoritmo se fija en la distancia entre observaciones. ¿Ahora la pregunta es como medimos la distancia?

Hay múltiples maneras de medir distancia entre dos puntos geométricos. Vamos a definir algunas.

KNN

Distancia euclidiana (modulo 2): Es la más conocida, es la mínima distancia (una recta) entre dos puntos en un espacio euclidiano. Es adecuada para datos numéricos continuos.

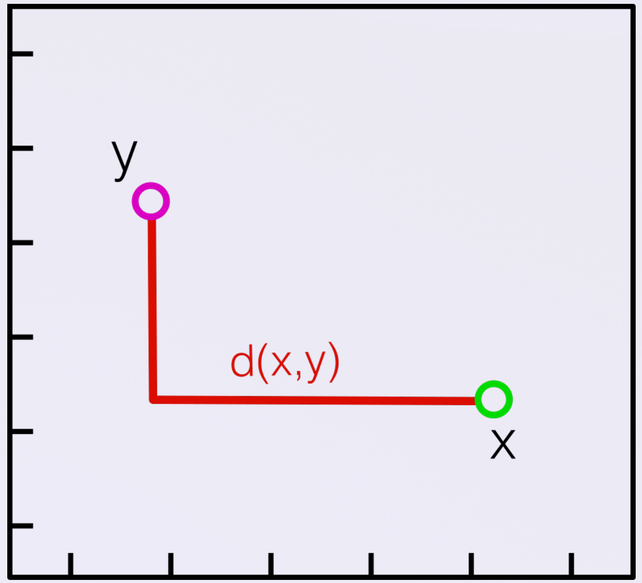
$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$



KNN

Distancia de Manhattan (modulo 1): Es la medida del valor absoluto entre dos puntos. Se conoce también como distancia taxi o de cuadra de ciudad, ya que mide distancias como en una ciudad. Es adecuada para datos que pueden tener correlaciones no lineales y no sigue la suposición de varianzas iguales en todas las dimensiones.

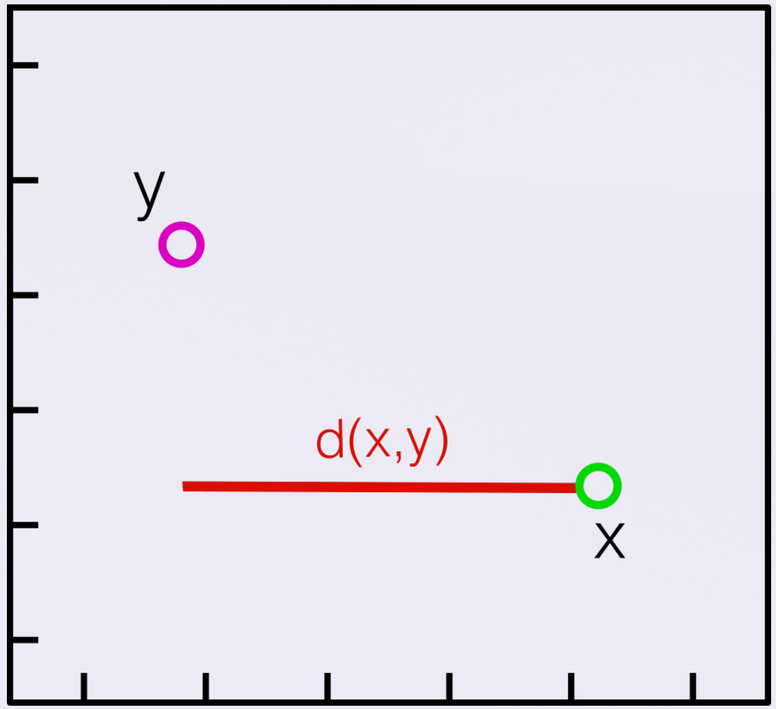
$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$



KNN

Distancia de Chebyshev (modulo infinito): Se calcula como la diferencia máxima entre las coordenadas de dos puntos. Es adecuada cuando las dimensiones son independientes y la distancia máxima es relevante.

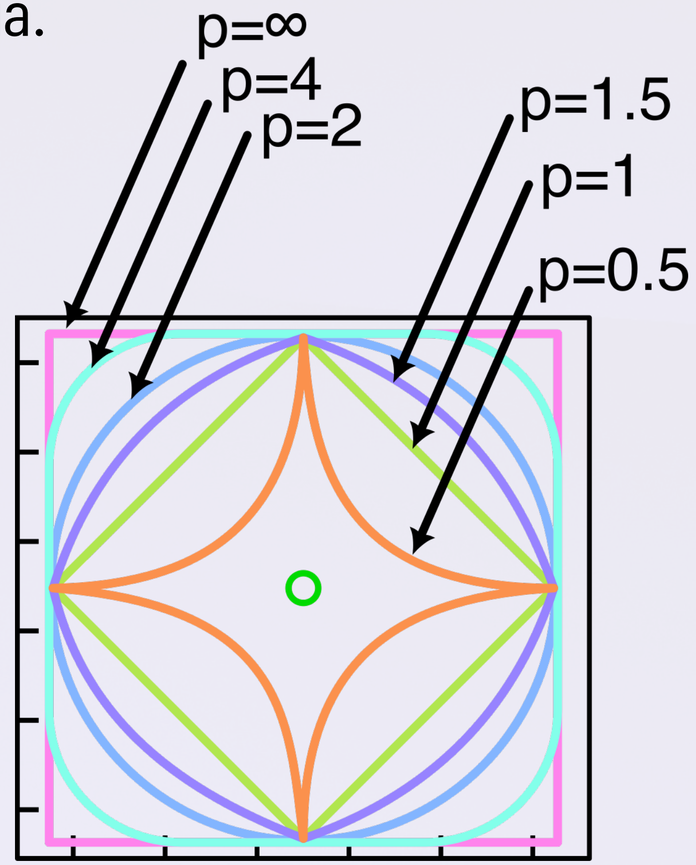
$$d(x, y) = \max_i \left(|x_i - y_i|\right)$$



KNN

Distancia de Minkowski: Es una medida generalizada que incluye las anteriores. Posee un parámetro, p, es la que permite variar el tipo de distancia.

$$d_p(x, y) = \left(\sum_{i=1}^{n} (x_i - y_i)^p \right)^{1/p}$$

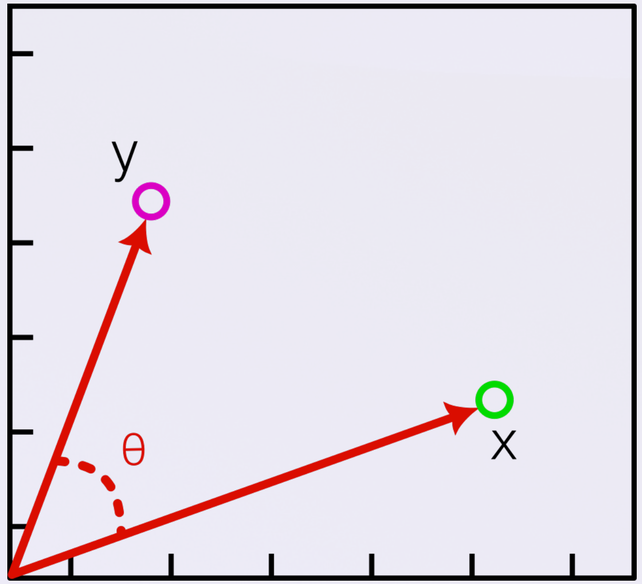


KNN

Distancia Coseno: La similitud coseno mide la similitud entre dos vectores como el coseno del ángulo entre ellos, y la distancia es 1 menos la similitud coseno. Es adecuada para datos donde la magnitud de los vectores es irrelevante, pero si su orientación.

$$d_c(x, y) = 1 - S_c(x, y)$$

$$S_c(x, y) = cos(\theta) = \frac{x \cdot y}{||x||\ ||y||}$$



KNN

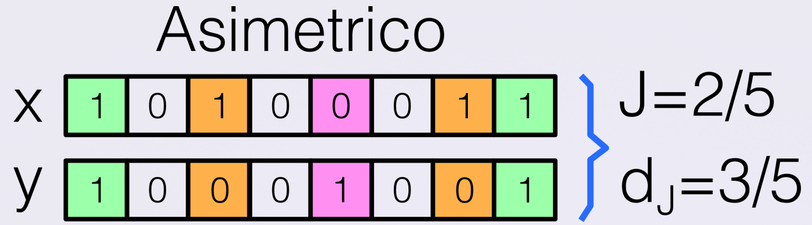
Distancia de Canberra: Es una métrica de distancia ponderada que se utiliza comúnmente para datos numéricos y pondera más las diferencias en las dimensiones donde los valores son pequeños. Es la distancia de Manhattan ponderada.

$$d(x, y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

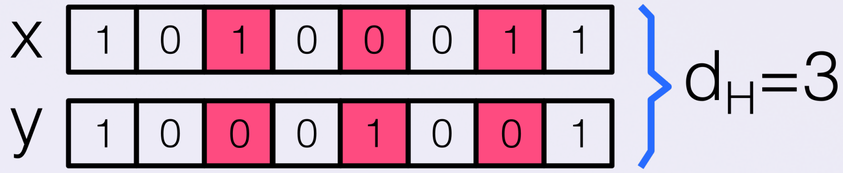
KNN

Distancia de Jaccard: Se utiliza comúnmente en conjuntos o datos binarios. Mide la similitud entre dos conjuntos como el tamaño de su intersección dividido por el tamaño de su unión.

$$J(x,y) = \frac{|x \cap y|}{|x \cup y|} \qquad d_J(x,y) = 1 - J(x,y)$$



Distancia de Hamming: Se usa típicamente con vectores booleanos, en donde se mide la cantidad de elementos del vector que son diferentes entre sí.



KNN

Distancia de Gower: Es una métrica de distancia que puede manejar datos mixtos (numéricos y categóricos) y tiene en cuenta la escala de las variables y la similitud entre las categorías. Esta entre 0 y 1.

- Datos numéricos: Se calcula usando la distancia de Manhattan pero para cada atributo se la divide por el rango de valores (poblacional o de muestra).
- Datos categóricos: Si el atributo es igual es 0, sino es 1.

KNN

Dada la metrica de distancia, debemos definir el valor de k, que es quien define con cuantos vecinos se usará para determinar la clasificación de un punto.

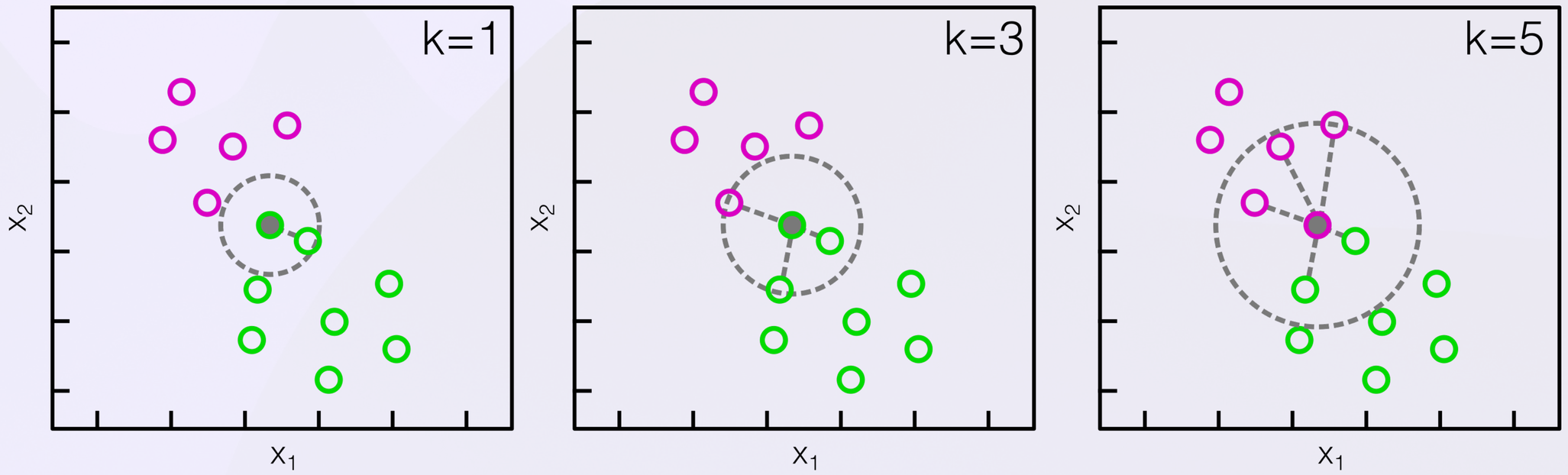
Por ejemplo, si k=1, la observación se asignará a la misma clase de su vecino mas cercano.

Definir k, el cual es un hiperparámetro justo al tipo de distancia elegida, es un acto de equilibrio.

Valores bajos de k pueden tener una varianza alta, pero un sesgo bajo, y valores altos de k un sesgo alto y poca varianza.

En general, se recomienda tener un **número impar** para k para evitar empates en la clasificación.

KNN




VAMOS A PRÁCTICAR UN POCO...