MACHINE LEARNING INMERSION

ANDRÉ OMAR CHÁVEZ PANDURO

Perú

« Divide las dificultades que examinas en tantas partes como sea posible , para su mejor solución»

AGENDA

- Clasificación.
- Modelo General de los Métodos de Clasificación.
- Regresión Logística Binaria.
- Clasificación Mediante k – Vecinos más cercanos.
- Clasificador Bayesiano : Naive Bayes

CLASIFICACIÓN: DEFINICIÓN

- Dada una colección de registros (Conjunto de Entrenamiento) cada registro contiene un conjunto de variables (atributos) denominado x, con una variable (atributo) adicional que es la clase denominada y.

- El objetivo de la ***clasificación*** es encontrar un modelo (una función) para predecir la clase a la que pertenecería cada registro, esta asignación una clase se debe hacer con la mayor precisión posible.

- Un conjunto de prueba (tabla de testing) se utiliza para determinar la precisión del modelo. Por lo general, el conjunto de datos dado se divide en dos conjuntos al azar de el de entrenamiento y el de prueba.

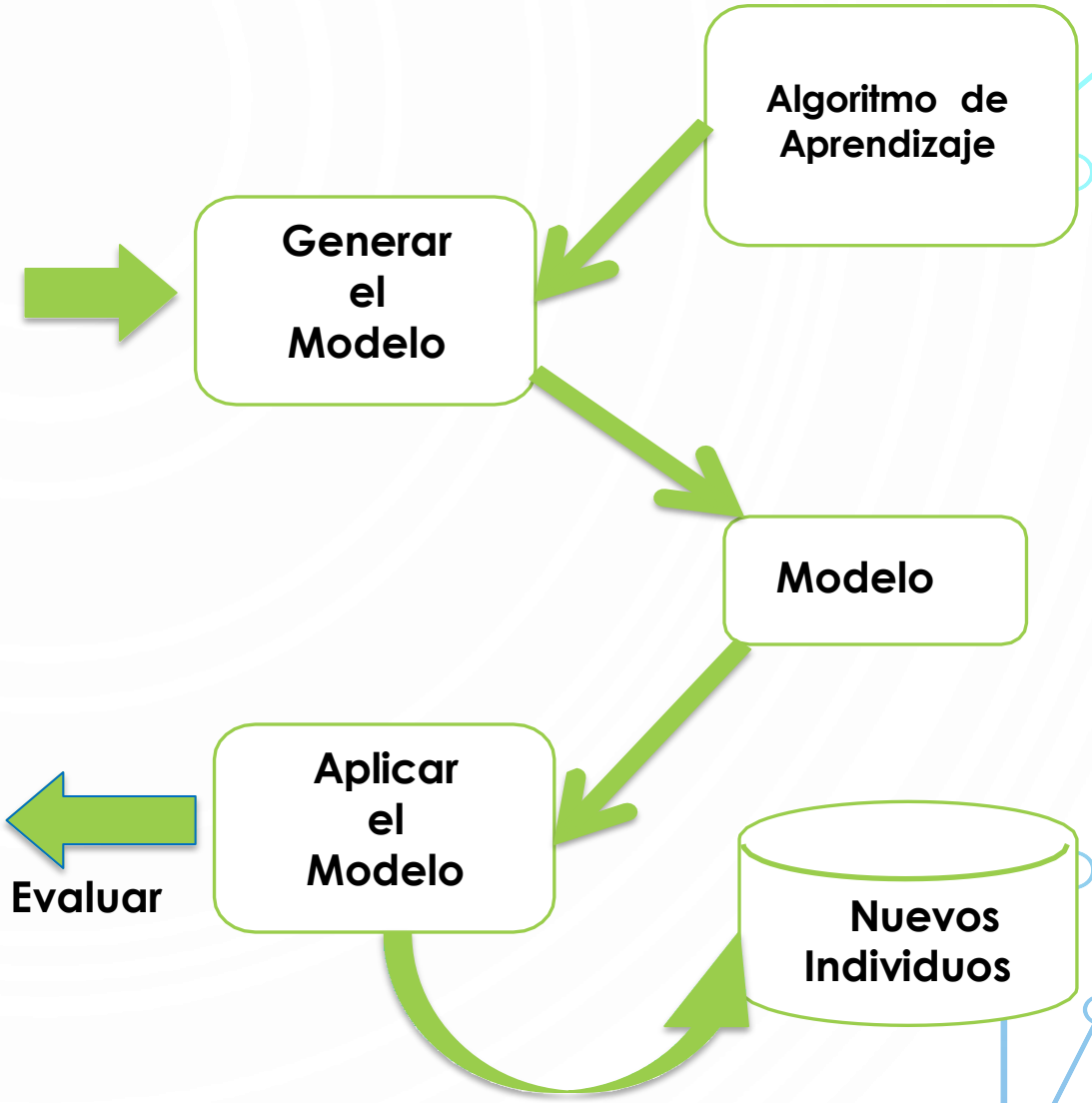
MODELO GENERAL DE LOS MÉTODOS DE CLASIFICACIÓN

ID	REEMBOLSO	ESTADO CIVIL	INGRESOS ANUALES	FRAUDE
1	SI	SOLTERO	S/ 1,000	NO
2	SI	CASADO	S/ 5,000	NO
3	NO	CASADO	S/ 3,500	SI
4	SI	VIUDO	S/ 4,500	NO
5	NO	SOLTERO	S/ 2,000	NO
6	NO	SOLTERO	S/ 1,500	SI

Tabla de Aprendizaje

ID	REEMBOLSO	ESTADO CIVIL	INGRESOS ANUALES	FRAUDE
7	SI	SOLTERO	S/ 4,000	NO
8	SI	CASADO	S/ 5,500	NO
9	NO	CASADO	S/ 6,500	SI

Tabla de Testing

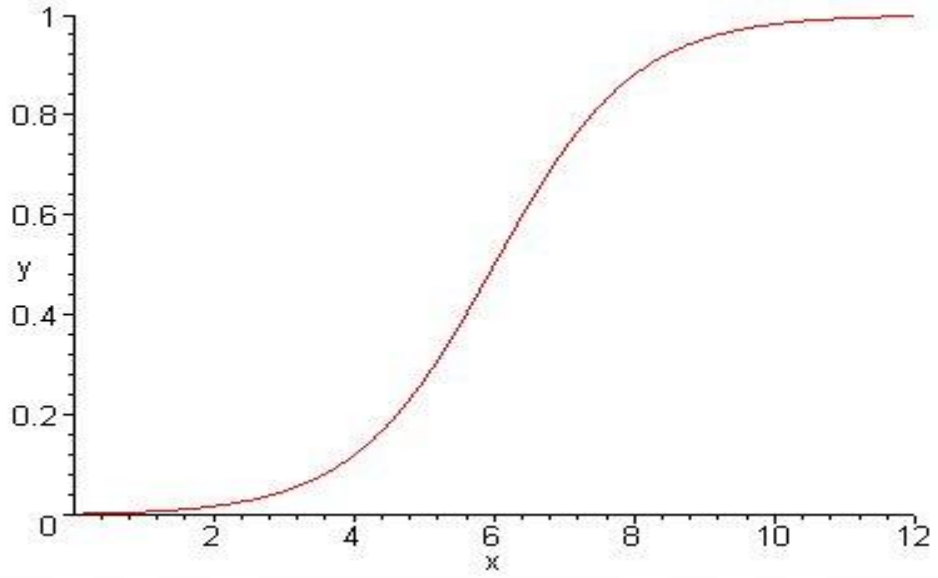


DEFINICIÓN DE CLASIFICACIÓN

- Dada una base de datos $D = \{t_1, t_2, \dots, t_n\}$ de tuplas o registros (individuos) y un conjunto de clases $C = \{C_1, C_2, \dots, C_m\}$, el **problema de la clasificación** es encontrar una función $f: D \rightarrow C$ tal que cada t_i es asignada una clase C_j.

- $f: D \rightarrow C$ podría ser una Red Neuronal, un Árbol de Decisión, un modelo basado en Análisis Discriminante, o una Red Beyesiana.

Regresión Logística Binaria



DEFINICIÓN

- Es un modelo predictivo **supervisado**.
- La regresión logística es un **modelo paramétrico** en el que la variable dependiente es cualitativa binaria.
- Es flexible en cuanto a la naturaleza de las variables explicativas, pues éstas pueden ser de **cuantitativas y categóricas.**
- Permite estudiar el **impacto** que tiene cada una de las variables independientes en la probabilidad de que **ocurra el suceso de estudio**.

MODELO DE REGRESIÓN LOGÍSTICA DICOTÓMICO

La variable Morosidad toma los siguientes valores:

"1" si el cliente es **moroso**.
"0" si el cliente es **no moroso**.

¿Es dicotómica?
¿Es cualitativa?
¿Es mutuamente excluyente?

MODELO DE REGRESIÓN LOGÍSTICA DICOTÓMICO

- Para este modelo se considera que la variable respuesta, es una variable **dicotómica que tomas dos valores**.
- Para estos modelos dicotómicos, las dos categorías deben de ser **mutuamente excluyentes.**
- La variable respuesta se puede expresar de la siguiente forma:

$$Y_i = \begin{cases} 1, \Pr ob(Y_i = 1) = P_i \\ 0, \Pr ob(Y_i = 0) = 1 - P_i \end{cases}$$

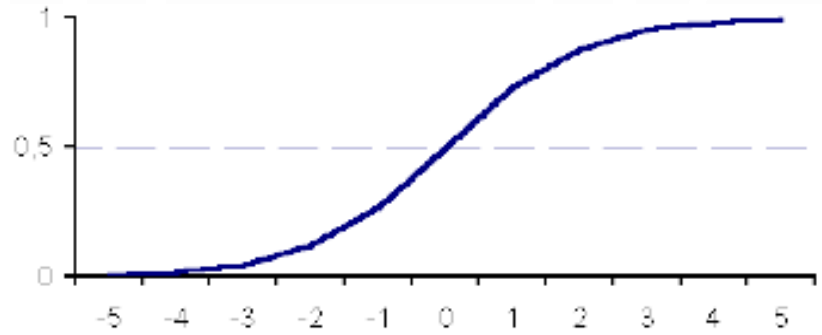
MODELO DE REGRESIÓN LOGÍSTICA DICOTÓMICO

$$p_i = f(\beta_0 + \beta_1 X_1)$$ Se calcula la función logit.

$$p_i = \frac{e^{(\beta_0+\beta_1 X_1)}}{1+e^{(\beta_0+\beta_1 X_1)}} = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1)}}$$

$$1-p_i = \frac{1}{1+e^{(\beta_0+\beta_1 X_1)}}$$

MODELO DE REGRESIÓN LOGÍSTICA DICOTÓMICO



La representación matemática del modelo es la siguiente:

$$z_i = \log \frac{P_i}{1-P_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

z_i : Variable dependiente del modelo: "Moroso" y " No Moroso"

p_i : Probabilidad de que el cliente sea "Moroso"

β_i : Coeficientes del modelo (parámetros a estimar)

x_i : Variables explicativas del modelo

MODELO DE REGRESIÓN LOGÍSTICA DICOTÓMICO

MÉTODO DE ESTIMACIÓN

- Para modelos de regresión logística, los parámetros se estiman a través de los métodos de **Máxima Verosimilitud.**
- Puesto que el modelo es no lineal, se necesita un algoritmo iterativo para esta estimación. El método iterativo que se aplica es el método de **Newton-Raphson**.

Parámetros desconocidos **Parámetros estimados**

$$z_i = \log\frac{P_i}{1-P_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \quad \Rightarrow \quad \hat{\beta}_i$$

Máxima Verosimilitud

MODELO DE REGRESIÓN LOGÍSTICA DICOTÓMICO

Odds Ratio

Es la razón entre la probabilidad de que se produzca un suceso y la probabilidad de que no se produzca ese suceso.

$$z_i = \log\frac{P_i}{1-P_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

Indica cuánto más probable es ser un cliente "Moroso" que "No Moroso"

EJEMPLO DE ODDS (CHANCE)

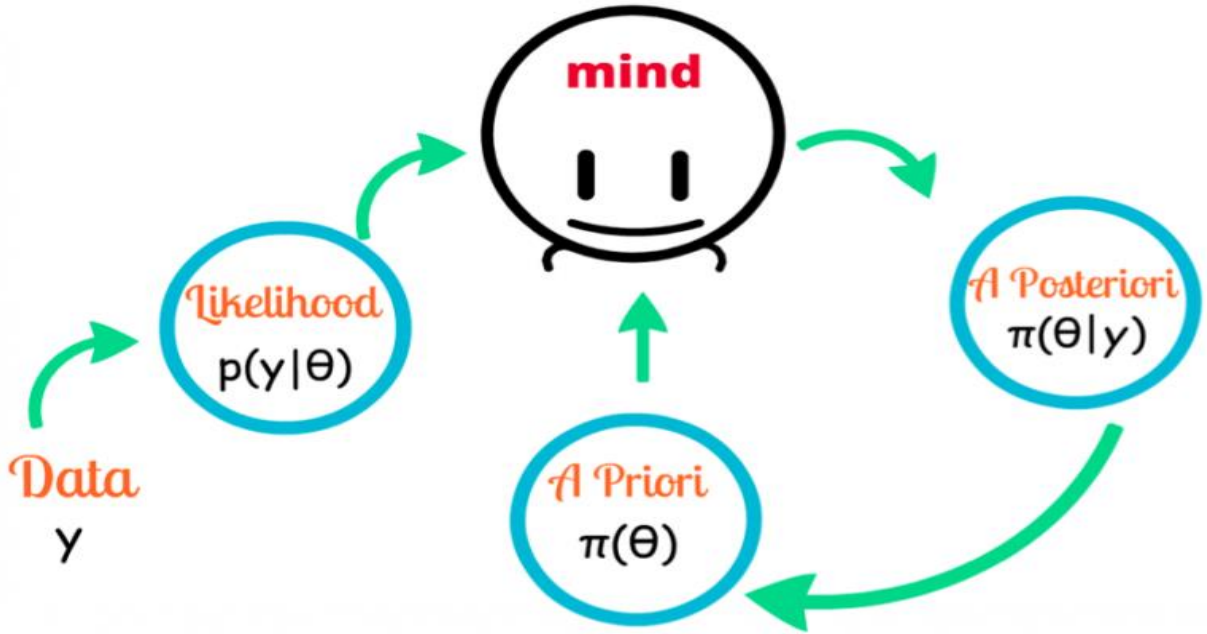
Tabla CHURN	Deudas en el SSFF? SI	NO	Total
Si	60	50	110
No	80	120	200
Total	140	170	310

$$\Omega deudas = \frac{\rho deudas}{\rho\,(1 - deudas)} = \frac{\frac{60}{140}}{\frac{80}{140}} = 0.75$$

$$\Omega nodeudas = \frac{\rho nodeudas}{\rho\,(1 - nodeudas)} = \frac{\frac{50}{170}}{\frac{120}{170}} = 0.42$$

$$OR = \frac{\Omega deudas}{\Omega nodeudas} = 1.78$$

MODELOS BAYESIANOS NAIVE BAYES



INTRODUCCIÓN

- Estudió el problema de la determinación de la probabilidad de las causas a través de los efectos observados.

Thomas Bayes

Definición

- Es un método importante no sólo porque ofrece un análisis cualitativo de las atributos y valores que pueden intervenir en el problema, sino porque da cuenta también de la importancia cuantitativa de esos atributos. En el aspecto cualitativo podemos representar cómo se relacionan esos atributos ya sea en una forma causal, o señalando simplemente de la correlación que existe entre esas variables (o atributos). Cuantitativamente (y ésta es la gran aportación de los métodos bayesianos).

DEFINICIÓN

$$p(C|F_1, \ldots, F_n) = \frac{p(C)\ p(F_1, \ldots, F_n|C)}{p(F_1, \ldots, F_n)}$$

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

Clasificador Naïve Bayes (cont)

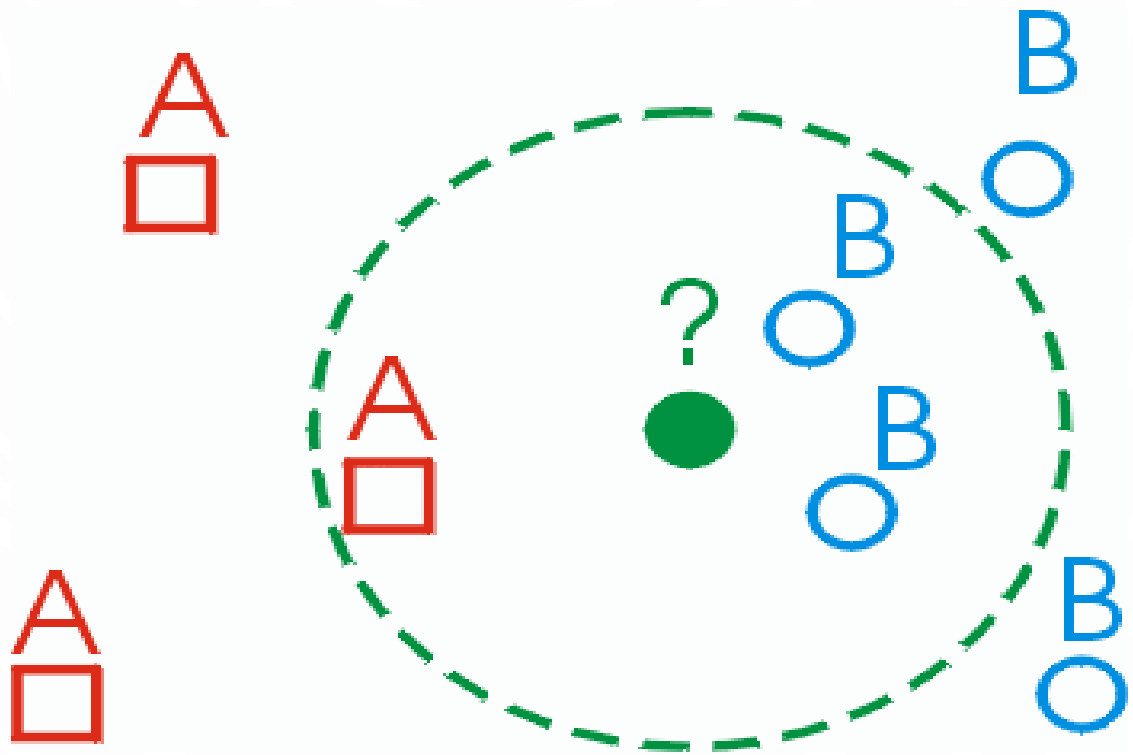
En este caso

$$P[X_j = a_j / C_i] = \frac{1}{s_j\sqrt{2\pi}} \exp[\frac{-(a_i - \bar{x}_j)^2}{2s_j^2}]$$

Donde, $\bar{x}_j$ y s_j son la media y la varianza de los valores de la variable X_j en la clase C_i.

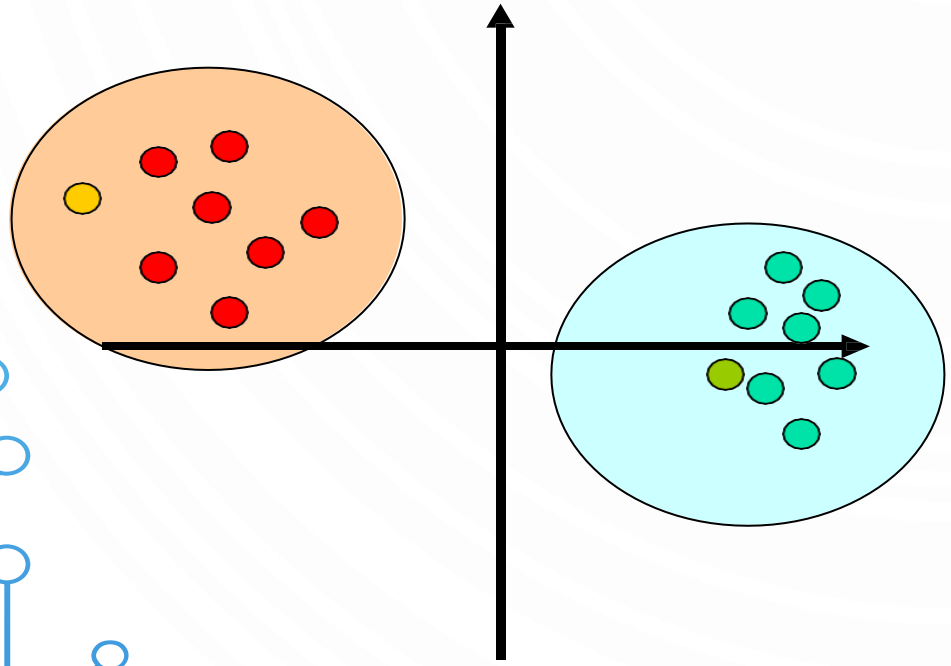
La libreria e1071 de R contiene una funcion **naiveBayes** que calcula el clasificador naïve Bayes, tanto para datos discretos como continuos.

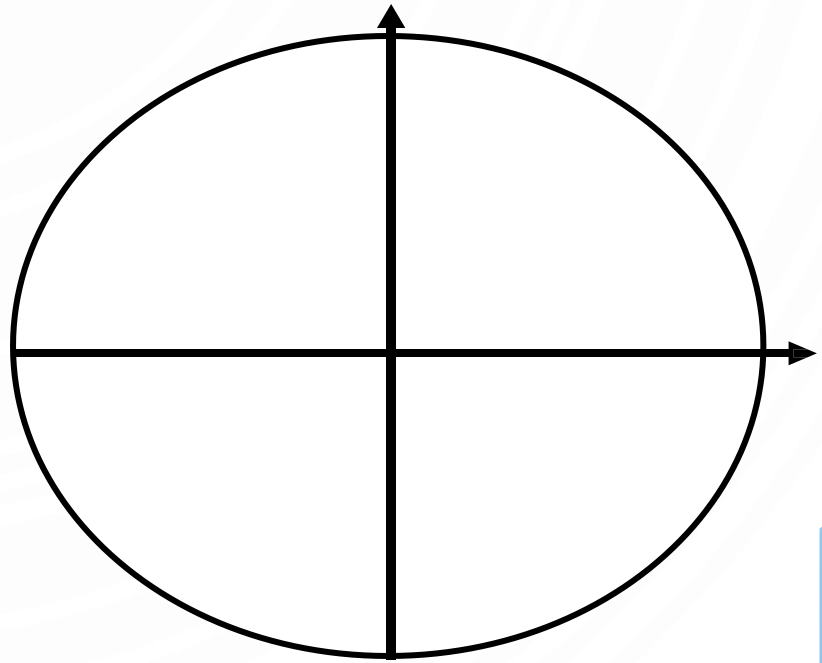
K – Vecinos más cercanos



CLASIFICACIÓN MEDIANTE K-VECINOS MÁS CERCANOS.

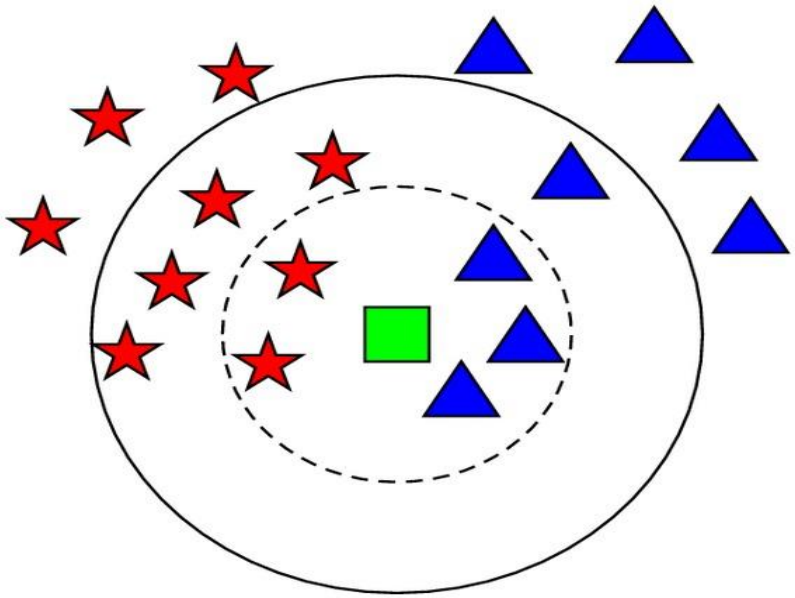
- Análisis de vecino más próximo es un método de clasificación de casos basado en su similaridad con otros casos. En aprendizaje de máquinas, se ha desarrollado como una forma de reconocer patrones de datos sin requerir una coincidencia exacta con patrones o casos almacenados.





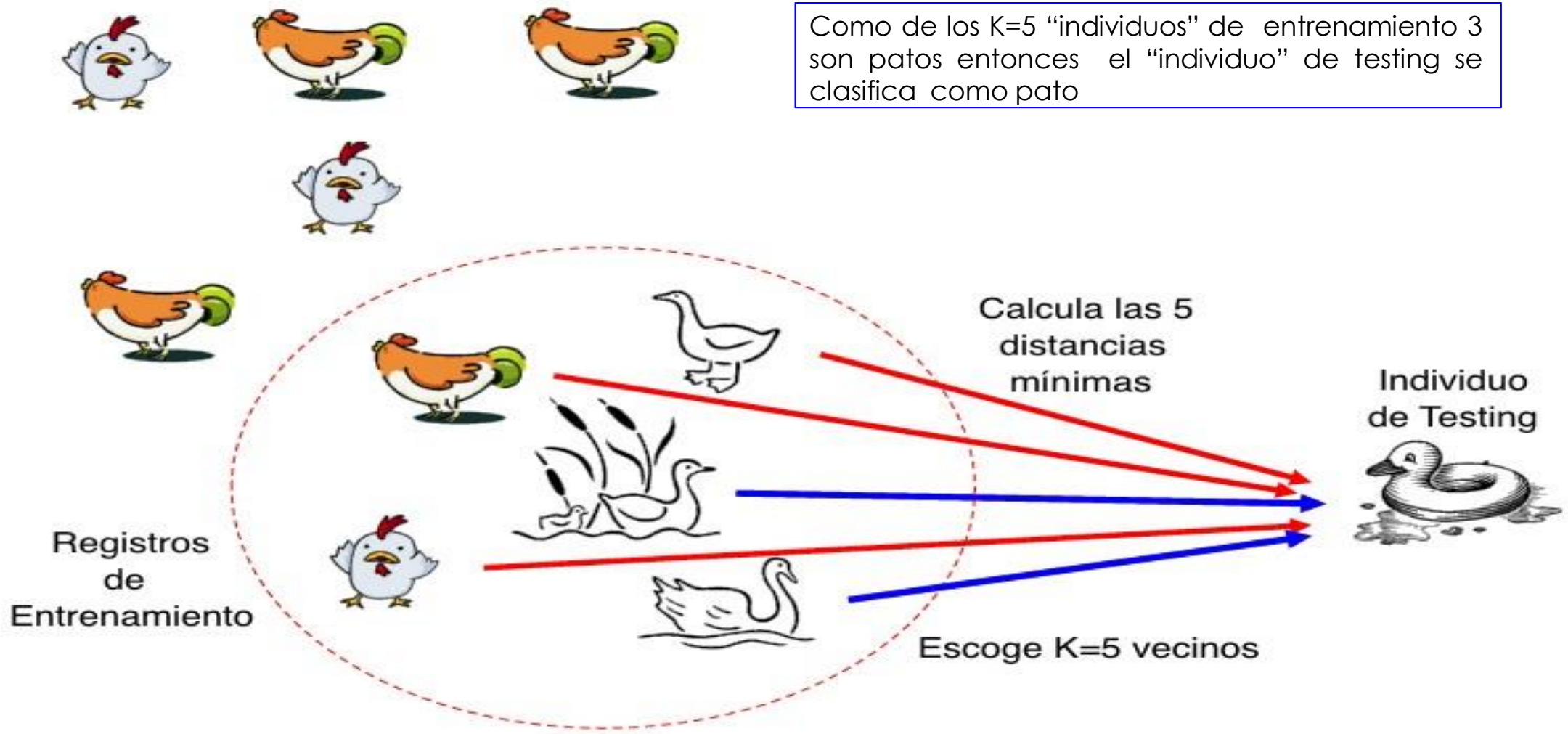
CLASIFICACIÓN MEDIANTE K-VECINOS MÁS CERCANOS.

- Los casos similares están cercanos entre sí y los casos no similares están distantes entre sí. Además, la distancia entre dos casos es una medida de sus diferencias.

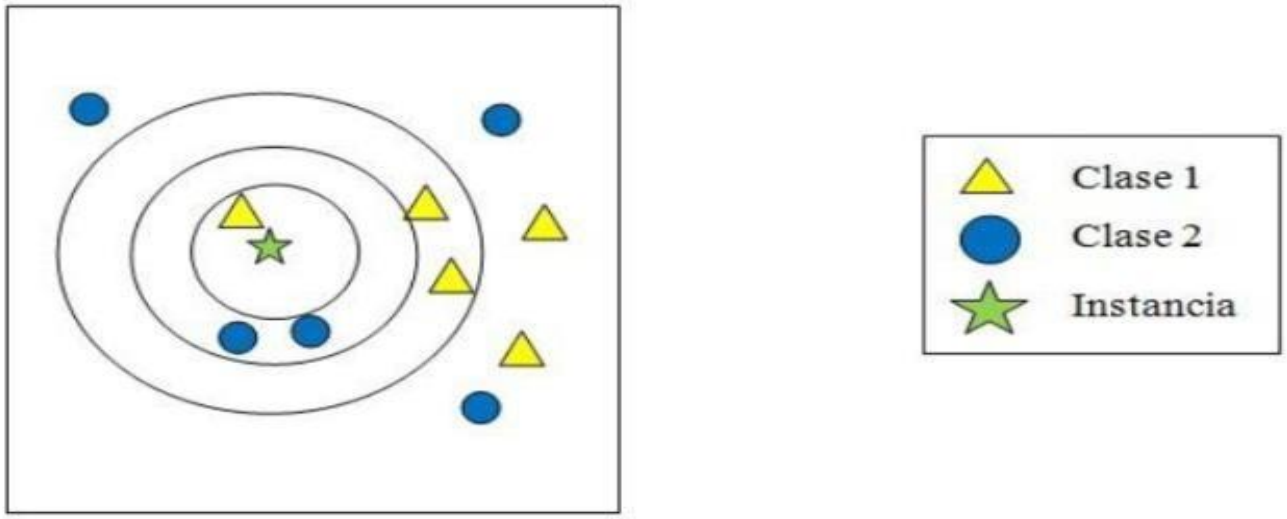


CLASIFICACIÓN MEDIANTE K-VECINOS MÁS CERCANOS : IDEA INTUITIVA

Como de los K=5 "individuos" de entrenamiento 3 son patos entonces el "individuo" de testing se clasifica como pato



CLASIFICACIÓN MEDIANTE K-VECINOS MÁS CERCANOS



Para K=1 (círculo más pequeño), la clase de la nueva instancia sería la Clase 1, ya que es la clase de su vecino más cercano, mientras que para K=3 la clase de la nueva instancia sería la Clase 2 pues habrían dos vecinos de la Clase 2 y solo 1 de la Clase 1

Clasificación mediante k-vecinos más cercanos : Elección de la distancia

Distancias y disimilaridades		Métrica	Euclídea
Euclídea	$\sqrt{\sum_{t=1}^{T}(x_{it}-x_{jt})^2}$	Si	Si
Manhattan	$\sum_{t=1}^{T}\lvert x_{it}-x_{jt}\rvert$	Si	No
Bray-Curtis	$\frac{\sum_{t=1}^{T}\lvert x_{it}-x_{jt}\rvert}{\sum_{t=1}^{T}(x_{it}+x_{jt})}$	Si	No
Canberra	$\sum_{t=1}^{T}\frac{\lvert x_{it}-x_{jt}\rvert}{(x_{it}+x_{jt})}$	Si	No
Minkowski	$\sqrt[q]{\sum_{t=1}^{T}\lvert x_{it}-x_{jt}\rvert^q}$	Si	Si
Mahalanobis	$\sqrt{\sum_{t=1}^{T}\sum_{l=1}^{T}(x_{it}-x_{jt})\sigma_{il}^{-1}(x_{il}-x_{jl})}$	Si	Si

Clasificación mediante k-vecinos más cercanos : Algoritmo

COMIENZO

Entrada: $D = \{(\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_N, c_N)\}$

$\mathbf{x} = (x_1, \ldots, x_n)$ nuevo caso a clasificar

PARA todo objeto ya clasificado (x_i, c_i)

calcular $d_i = d(\mathbf{x}_i, \mathbf{x})$

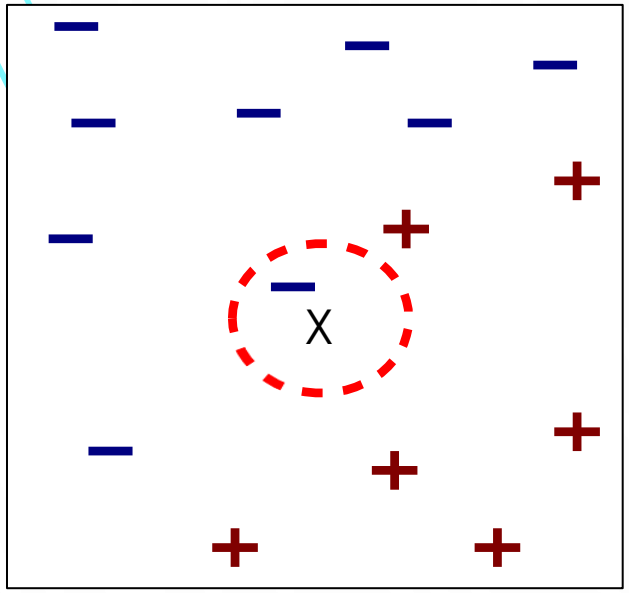
Ordenar $d_i (i = 1, \ldots, N)$ en orden ascendente

Quedarnos con los K casos $D_{\mathbf{x}}^K$ ya clasificados más cercanos a $\mathbf{x}$

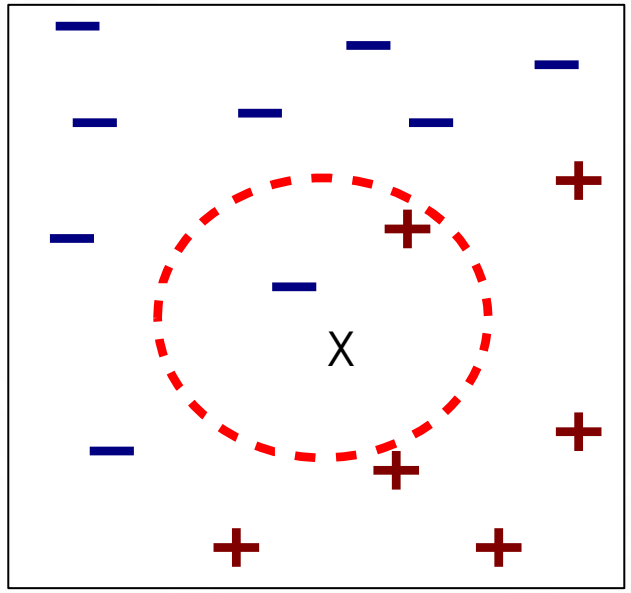
Asignar a $\mathbf{x}$ la clase más frecuente en $D_{\mathbf{x}}^K$

FIN

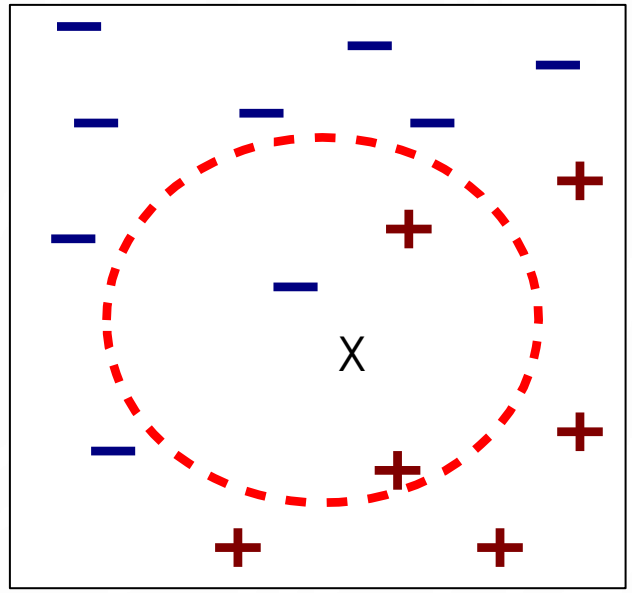
CLASIFICACIÓN MEDIANTE K-VECINOS MÁS CERCANOS : ¿ELECCIÓN DEL K ÓPTIMO?



(a) 1 - Vecino más cercano.



(b) 2 - Vecinos más cercanos.



(c) 3- Vecinos más cercanos.

- Si K es muy pequeño el modelo será muy sentitivo a puntos que son atípicos o que son ruido (datos corruptos)
- Si K es muy grande, el modelo tiende a asignar siempre a la clase más grande.

GRACIAS
POR SU ATENCIÓN