Claudia Chávez

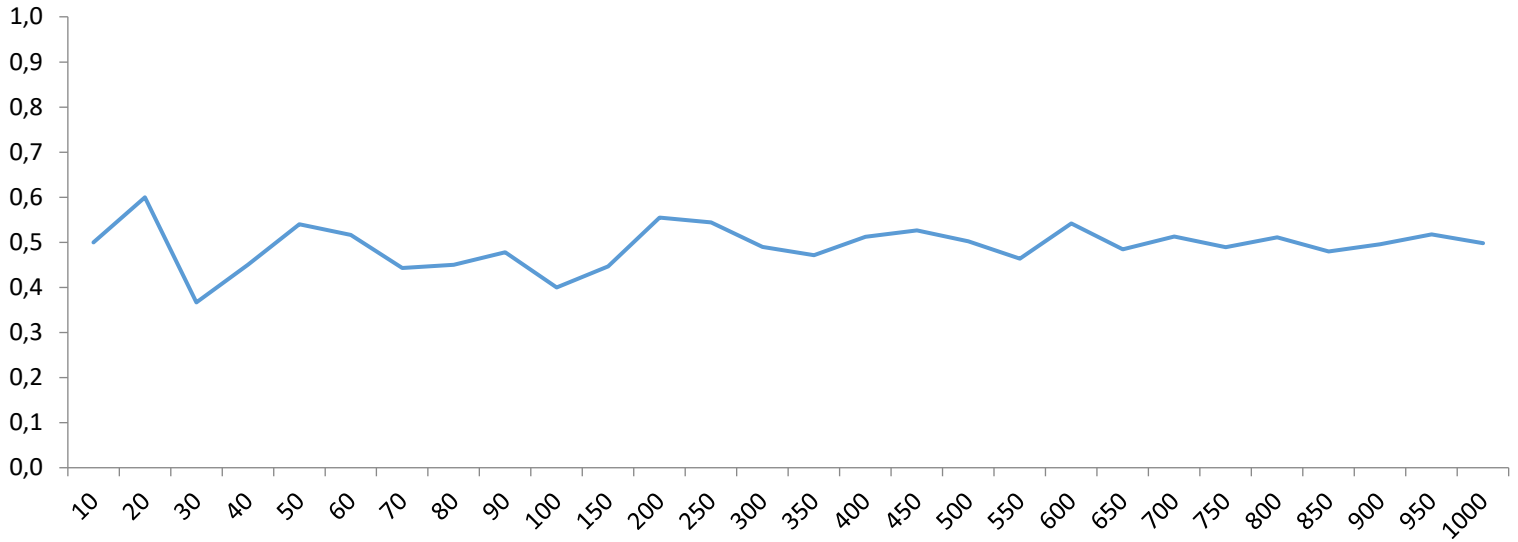
Modelos probabilísticos: Naïve Bayes

Clasificadores probabilísticos

- Basado en el teorema de Bayes
- Combina datos de ejemplo con conocimiento a priori
- Usado como clasificador, puede obtener probabilidades de pertenecer a cada clase asumiendo independencia de los predictores
- Posibilidad de construir representaciones más complejas

Antes…Probabilidad de Laplace

- Si un experimento aleatorio da lugar a un número finito de resultados posibles, se define la regla de Laplace como el cociente entre el numero de casos favorables del suceso A, y el de todos los resultados posibles:



Se lanza una moneda "normal" y se observa si sale cara o sello.

$$P(A) = \frac{\text{número de casos favorables a A}}{\text{número de casos posibles}}$$

Probabilidad Condicional e Independencia de sucesos

- Si B es un suceso aleatorio (P(B)>0) y A es otro suceso aleatorio. Llamaremos probabilidad Condicionada de A dado (|) B a:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

- Probabilidad conjunta

$$P(A \cap B) = P(B|A) * P(A)$$

Independencia

- Se define que dos sucesos, A y B, son independientes si:

$$P(A \cap B) = P(B|A) * P(A)$$

$$P(A \cap B) = P(A) * P(B)$$

$$P(B|A) = P(B)$$

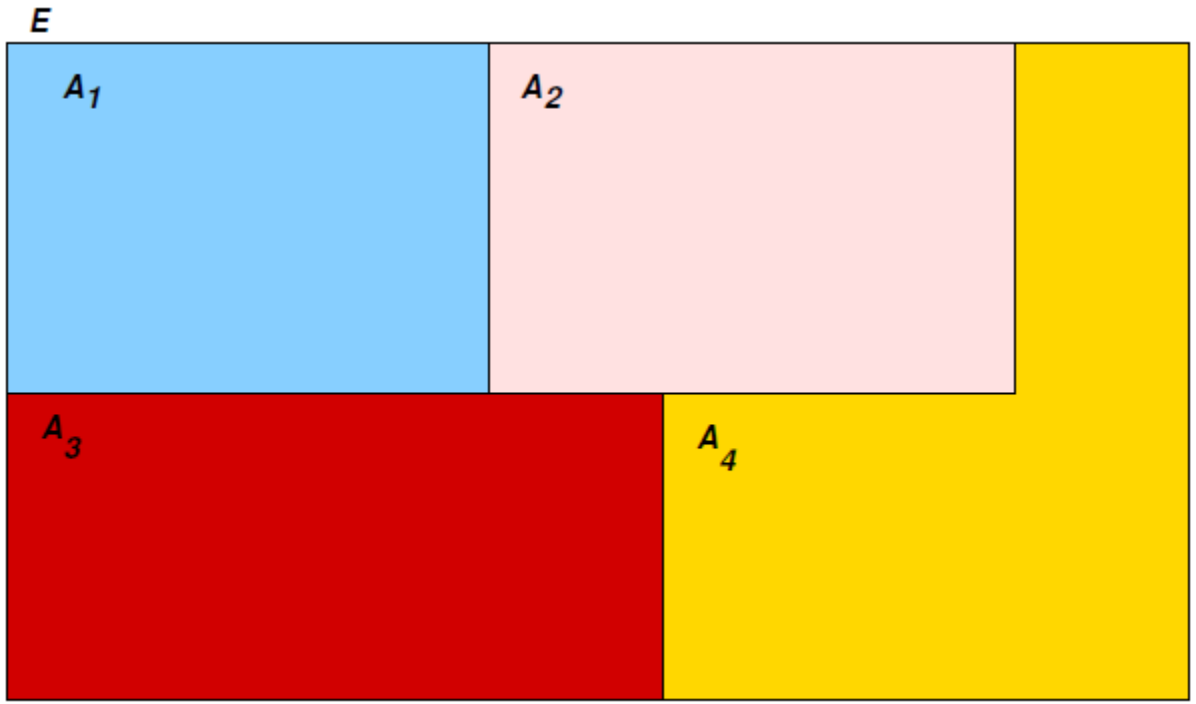
Sistema exhaustivo y excluyente

- Se dice que un sistema es exhaustivo y excluyente cuando:

$$\bigcup_{i=1}^{n} A_i = E$$

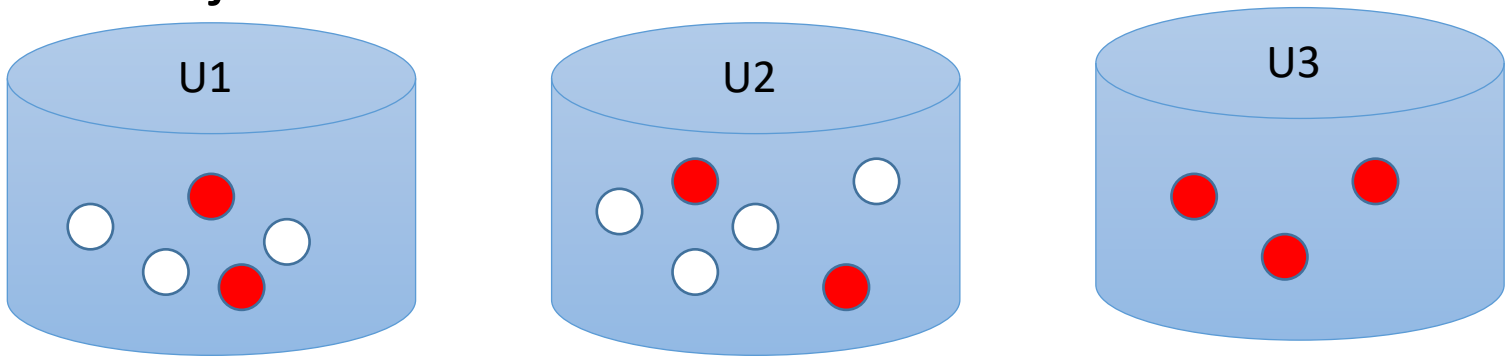
$$A_i \cap A_j = \{\}$$

Sistema exhaustivo y excluyente



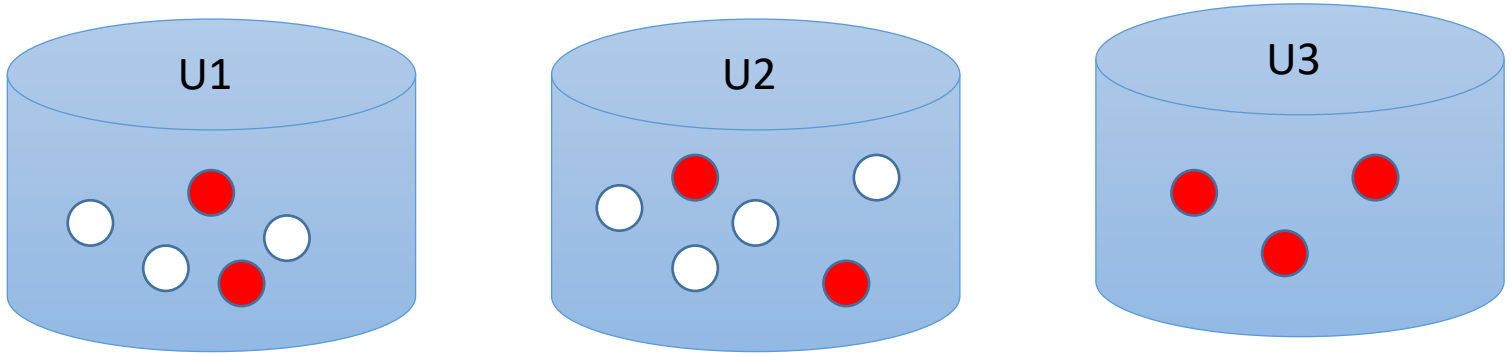
Ejemplo Teorema de Bayes

- Se tienen 3 urnas. Cada una de ellas contiene un número diferente de bolas blancas y bolas rojas:



Alguien elije al azar y con la misma probabilidad una de las tres urnas y se saca una bola. Si el resultado del experimento ha sido una bola blanca ¿Cuál es la probabilidad de que provenga de la primera urna?

Ejemplo:



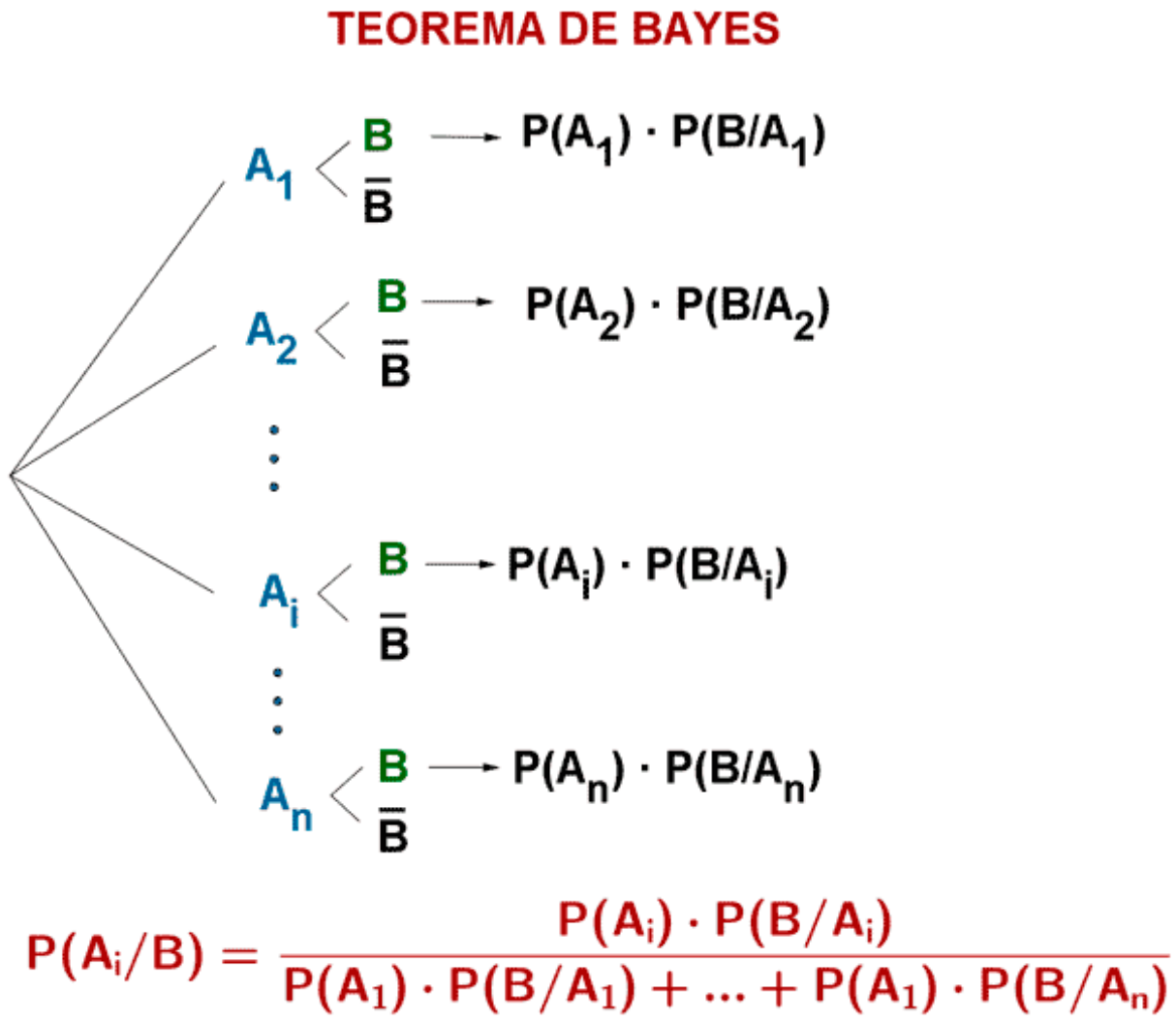
P(U1)=1/3 P(B/U1)=3/5	P(U2)=1/3 P(B/U2)=4/6	P(U3)=1/3 P(B/U3)=0

$$P(U_1 \mid B) = \frac{P(B \mid U_1)P(U_1)}{P(B \mid U_1)P(U_1) + P(B \mid U_2)P(U_2) + P(B \mid U_3)P(U_3)}$$

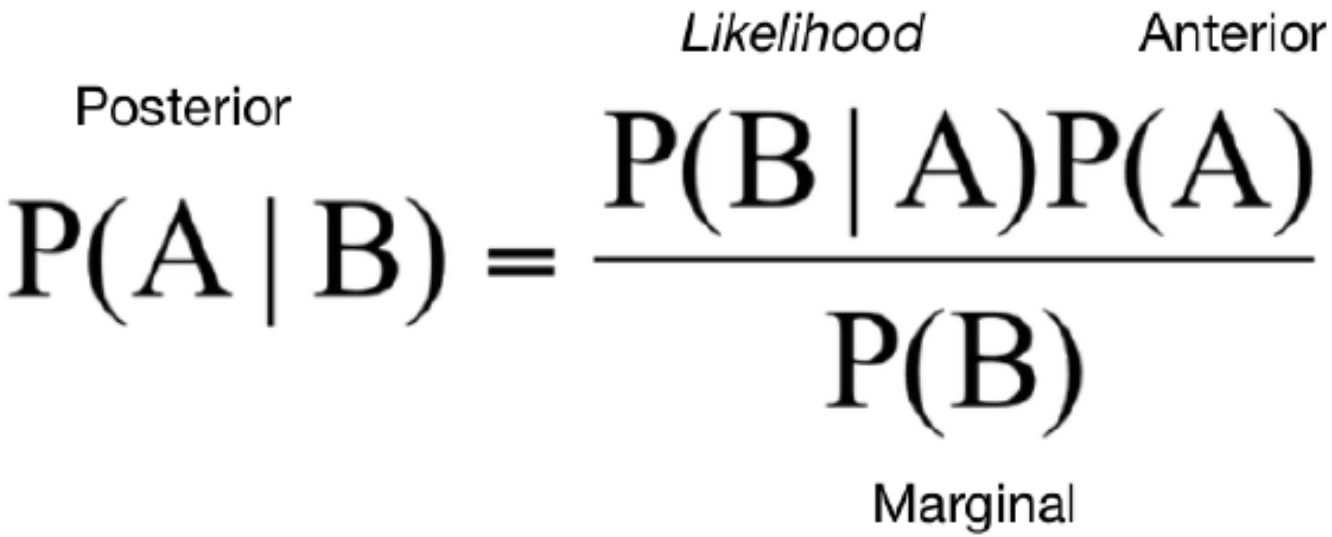
$$= \frac{\frac{3}{5} * \frac{1}{3}}{\frac{3}{5} * \frac{1}{3} + \frac{4}{6} * \frac{1}{3} + 0 * \frac{1}{3}} = \frac{9}{19} = 0{,}47$$

- Sea A1,A2,…, An sucesos pertenecientes a un sistema exhaustivo y excluyente. Sea B otro suceso, entonces se verifica que:

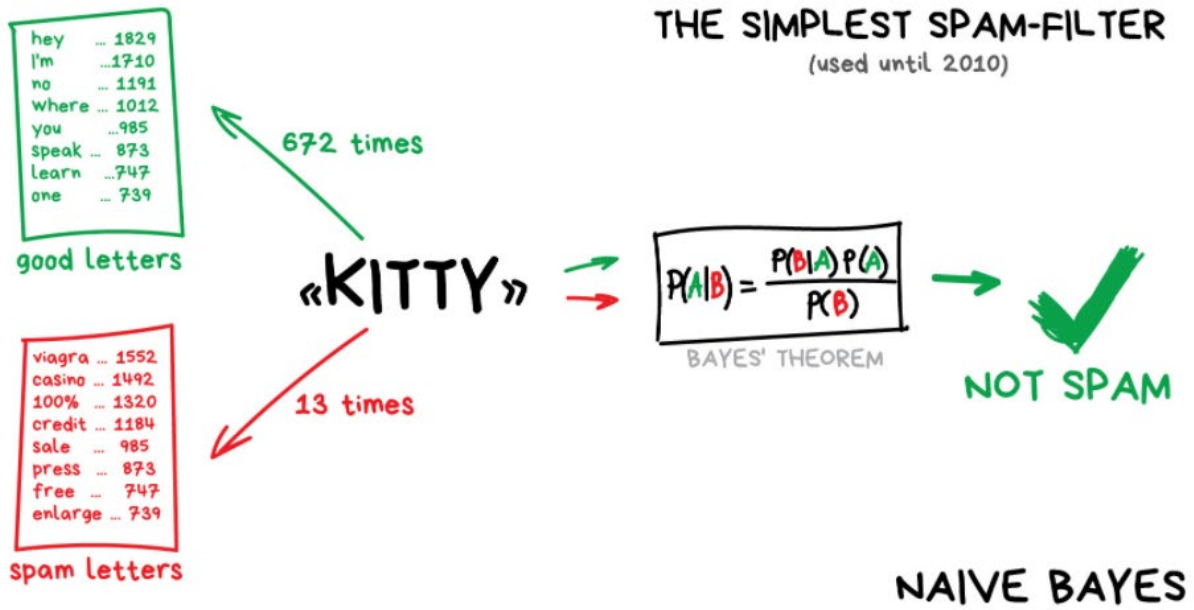
$$P(A_j \mid B) = \frac{P(B \mid A_j)P(A_j)}{\sum_{i=1}^{n} P(B \mid A_i)P(A_i)}$$



Teorema de Bayes



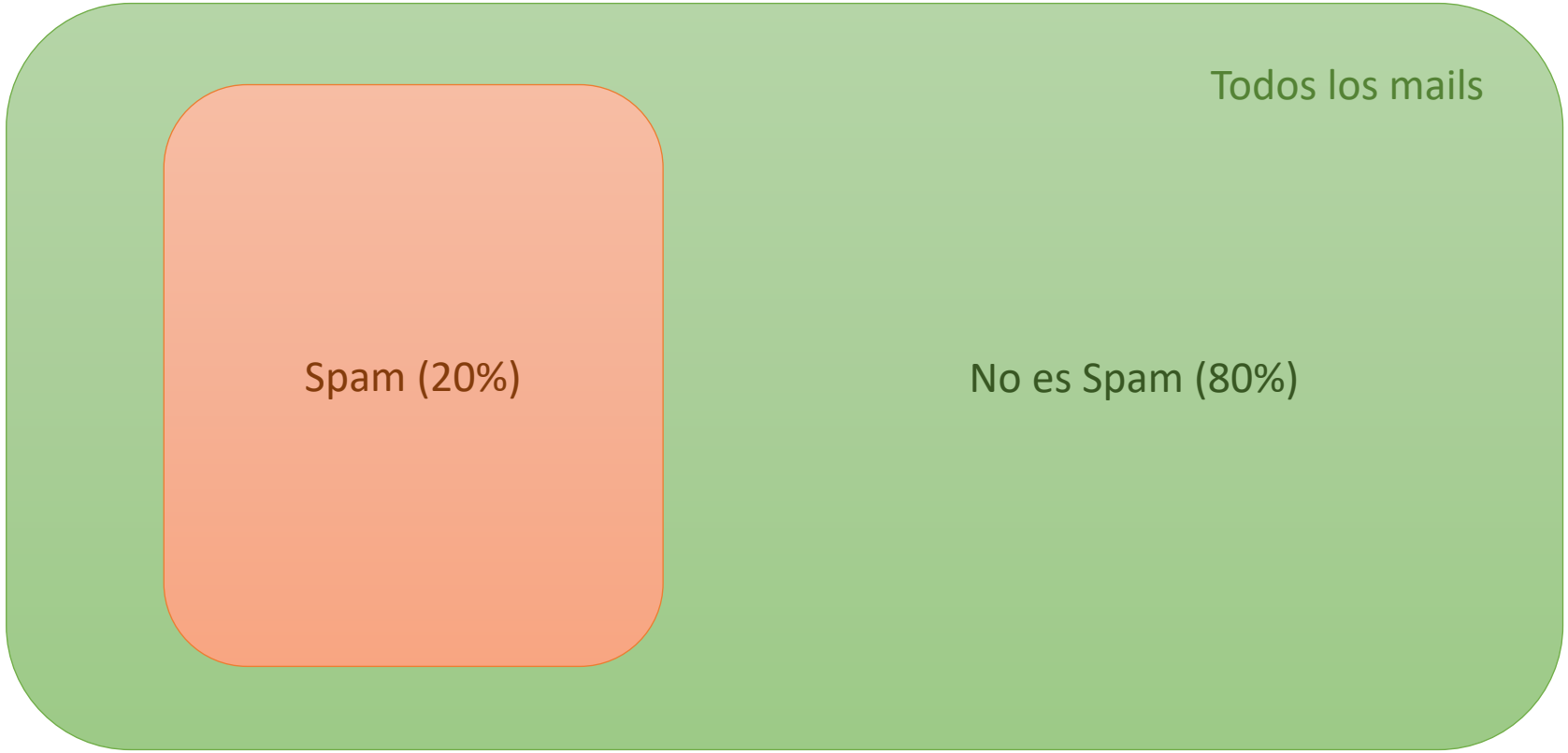
Clásico ejemplo del spam



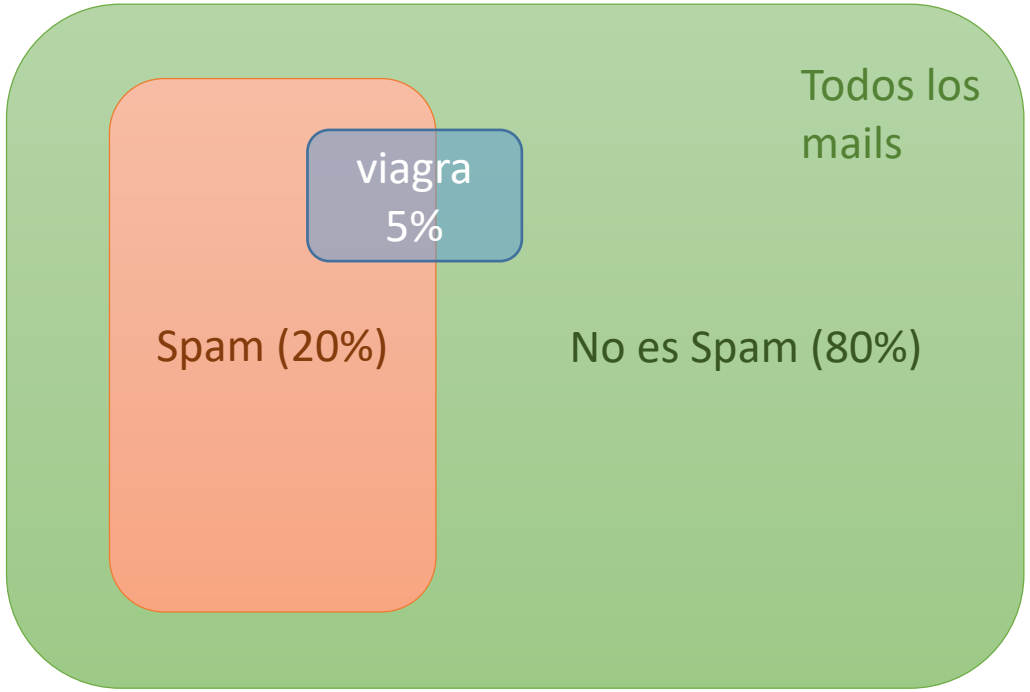
[Fuente: Vas3k Blog]

Siguiendo con los spam

- Supongamos que el 20% es spam



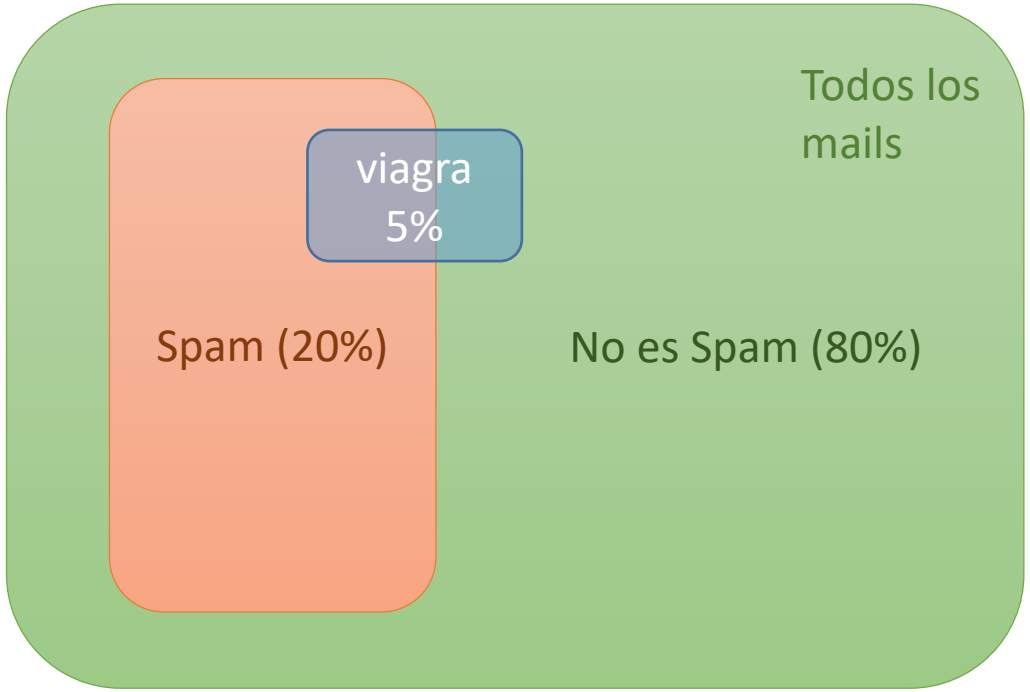
- Además se sabe que el 5% de los mail tiene la palabra “viagra”



Asumo independencia

$$P(\text{spam} \cap \text{viagra}) = P(\text{spam}) * P(\text{viagra}) = 0.2 * 0.05 = 0.01$$

- Sabemos que la independencia no es tal



	Viagra		
Frequency	Yes	No	Total
spam	4	16	20
ham	1	79	80
Total	5	95	100

	Viagra		
Likelihood	Yes	No	Total
spam	4 / 20	16 / 20	20
ham	1 / 80	79 / 80	80
Total	5 / 100	95 / 100	100

$$P(spam \cap viagra)=P(viagra/spam)*P(spam)=4/20*20/100=0.04$$

Aplicando el teorema de Bayes

	Viagra		
Likelihood	Yes	No	Total
spam	4 / 20	16 / 20	20
ham	1 / 80	79 / 80	80
Total	5 / 100	95 / 100	100

$$P(spam|Viagra) = \frac{P(Viagra|spam) * P(spam)}{P(Viagra)} = \frac{\frac{4}{20} * \frac{20}{100}}{\frac{5}{100}} = 0.8$$

¿Qué es lo que hace Nave Bayes?

- Naive significa ingenuo
- Es ingenuo debido a que asume independencia de los eventos
- Todas son igualmente importantes

$$P(c_k|x_1,\ldots,x_n) = \frac{P(x_1,\ldots,x_n|c_k) * P(c_k)}{P(x_1,\ldots,x_n)} \qquad = \frac{P(x_1|c_k)*\ldots*P(x_n|c_k) * P(c_k)}{P(x_1,\ldots,x_n)}$$

$$= \frac{1}{Z} P(c_k) \prod_{i=1}^{n} P(x_i|c_k)$$

	Viagra (W_1)		Money (W_2)		Groceries (W_3)		Unsubscribe (W_4)		
Likelihood	Yes	No	Yes	No	Yes	No	Yes	No	Total
spam	4 / 20	16 / 20	10 / 20	10 / 20	0 / 20	20 / 20	12 / 20	8 / 20	20
ham	1 / 80	79 / 80	14 / 80	66 / 80	8 / 80	71 / 80	23 / 80	57 / 80	80
Total	5 / 100	95 / 100	24 / 100	76 / 100	8 / 100	91 / 100	35 / 100	65 / 100	100

$$P\left(\text{spam} \mid W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4\right) = \frac{P\left(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4 \mid \text{spam}\right) P\left(\text{spam}\right)}{P\left(W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4\right)}$$

$$P\left(\text{spam} \mid W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4\right) \propto P\left(W_1 \mid \text{spam}\right) P\left(\neg W_2 \mid \text{spam}\right) P\left(\neg W_3 \mid \text{spam}\right) P\left(W_4 \mid \text{spam}\right) P\left(\text{spam}\right)$$

$$P\left(\text{ham} \mid W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4\right) \propto P\left(W_1 \mid \text{ham}\right) P\left(\neg W_2 \mid \text{ham}\right) P\left(\neg W_3 \mid \text{ham}\right) P\left(W_4 \mid \text{ham}\right) P\left(\text{ham}\right)$$

	Viagra (W_1)		Money (W_2)		Groceries (W_3)		Unsubscribe (W_4)		
Likelihood	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**Total**
spam	4 / 20	16 / 20	10 / 20	10 / 20	0 / 20	20 / 20	12 / 20	8 / 20	20
ham	1 / 80	79 / 80	14 / 80	66 / 80	8 / 80	71 / 80	23 / 80	57 / 80	80
Total	5 / 100	95 / 100	24 / 100	76 / 100	8 / 100	91 / 100	35 / 100	65 / 100	100

$$P(\text{spam} \mid W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) \propto P(W_1 \mid \text{spam}) P(\neg W_2 \mid \text{spam}) P(\neg W_3 \mid \text{spam}) P(W_4 \mid \text{spam}) P(\text{spam})$$

$$P(\text{ham} \mid W_1 \cap \neg W_2 \cap \neg W_3 \cap W_4) \propto P(W_1 \mid \text{ham}) P(\neg W_2 \mid \text{ham}) P(\neg W_3 \mid \text{ham}) P(W_4 \mid \text{ham}) P(\text{ham})$$

$$(4 / 20) * (10 / 20) * (20 / 20) * (12 / 20) * (20 / 100) = 0.012$$

$$(1 / 80) * (66 / 80) * (71 / 80) * (23 / 80) * (80 / 100) = 0.002$$

Tipos de algoritmos Naive Bayes

- Gaussiano
- Binomial
- Multinomial

Clasificador Gaussiano

El Teorema Central del Límite dice que la suma de un numero grande de variables aleatorias independientes idénticamente distribuidas siguen una distribución Normal.

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

- El valor esperado es: $E\{x\} = \mu$
- La varianza es: $Var(x) = \sigma^2$
- La desviación estandar es: $\sigma_x = \sigma$

Clasificador binomial

- Una distribución binomial da la probabilidad de observar r eventos de n muestras independientes con dos posibles resultados.

$$P(r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{(n-r)}$$

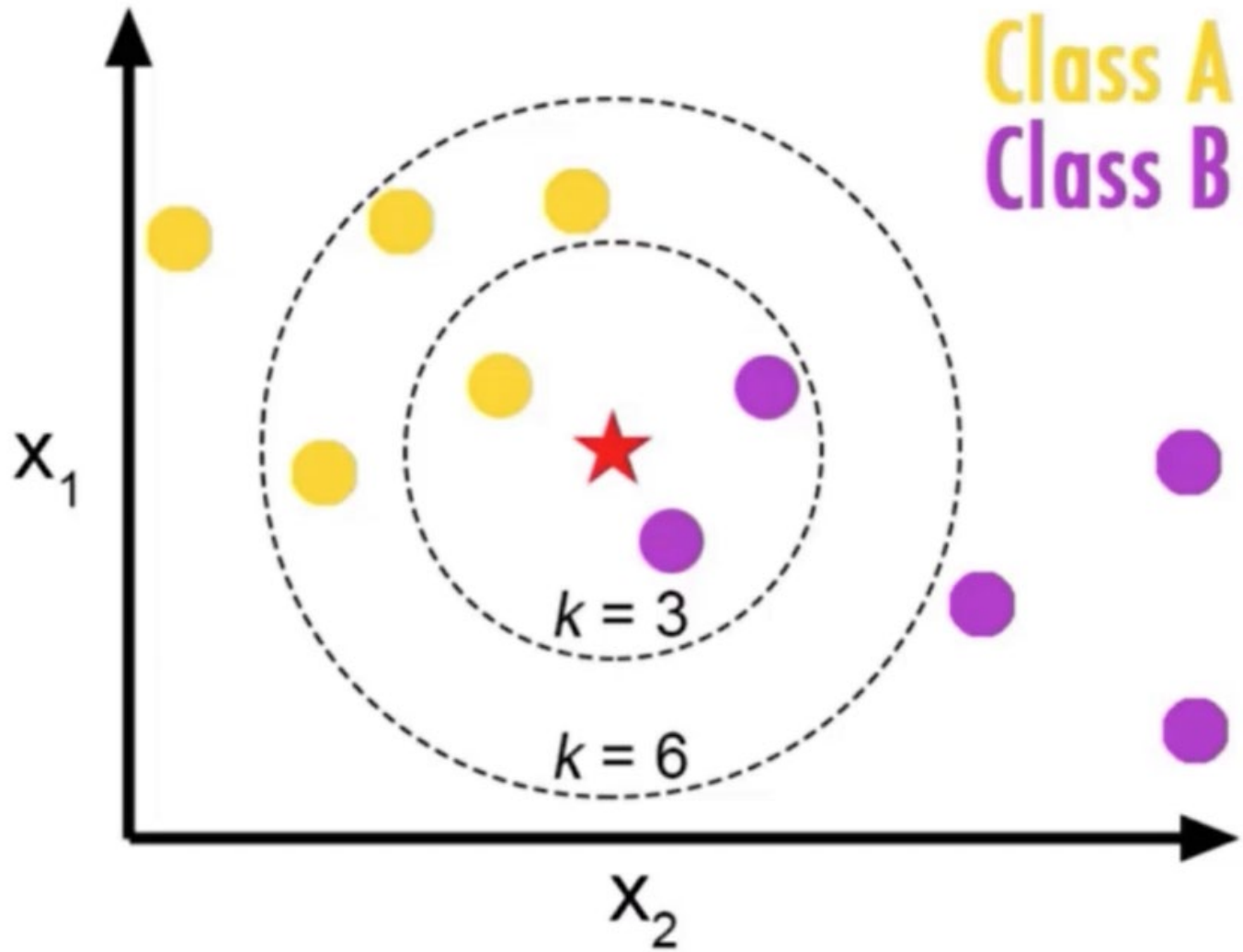
- El valor esperado es: $E\{x\} = np$
- La varianza es: $Var(x) = np(1-p)$
- La desviación estandar es: $\sigma_x = \sqrt{np(1-p)}$

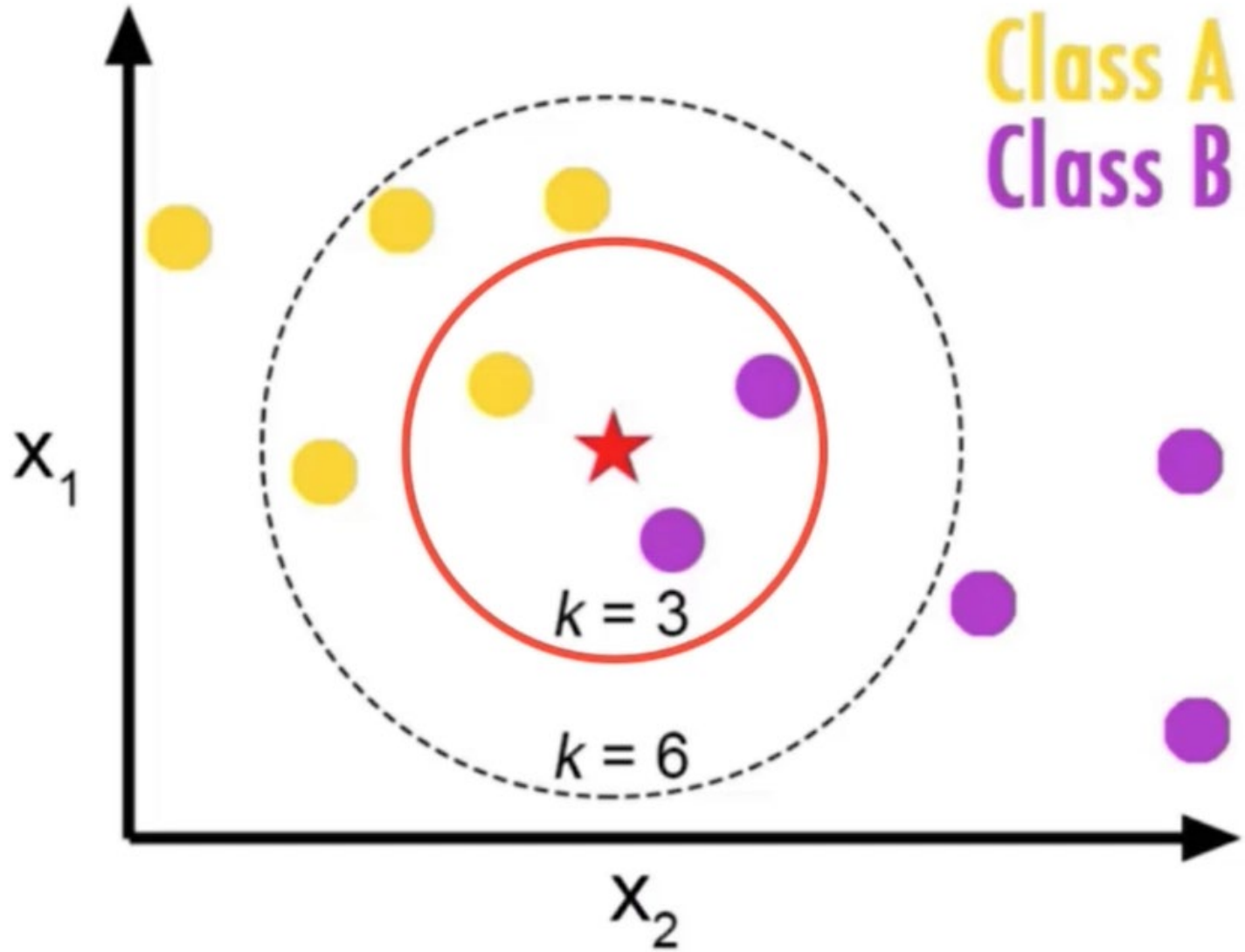
Si n es grande, se aproxima a una distribución Normal

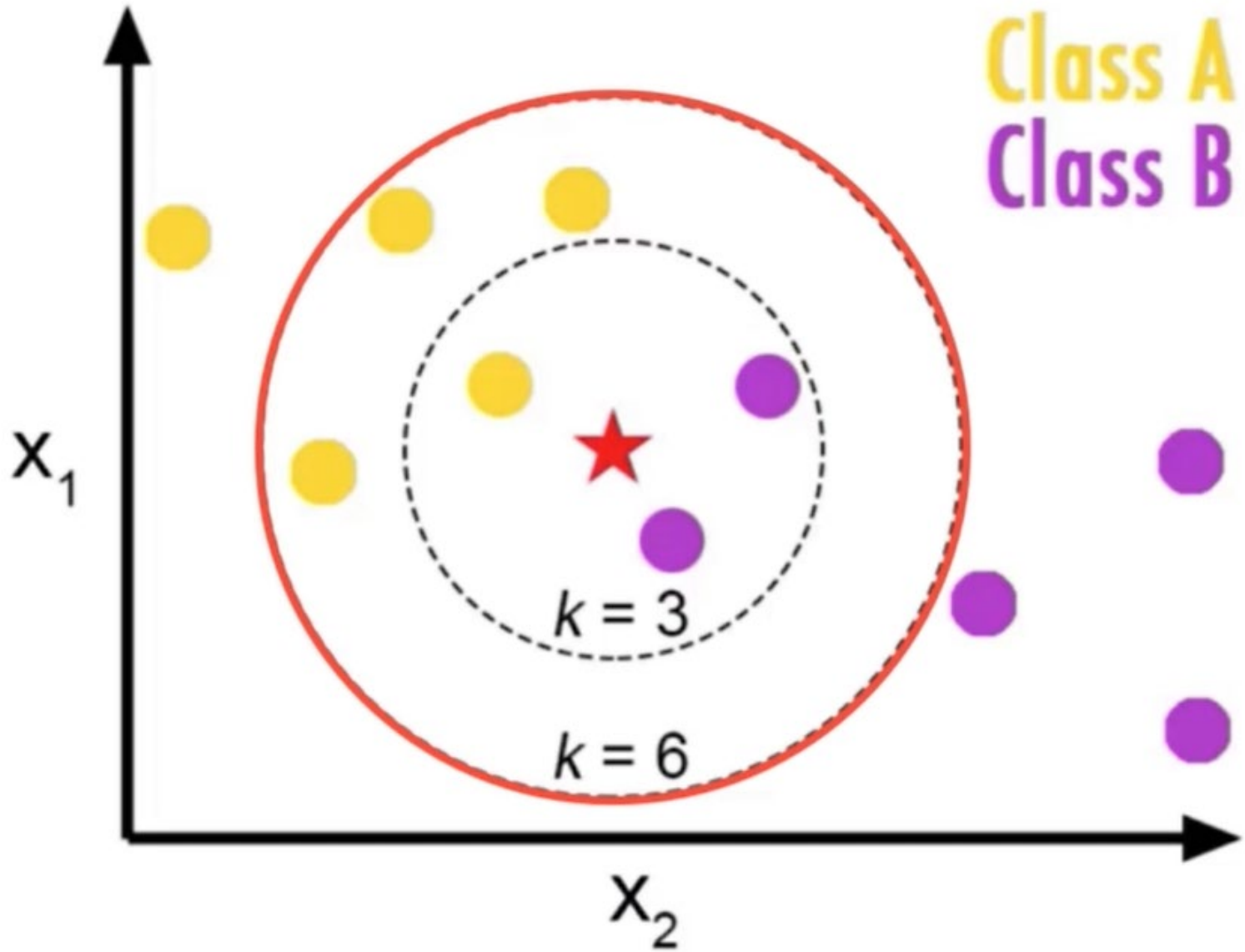
Multinomial es una extensión del clasificador binomial

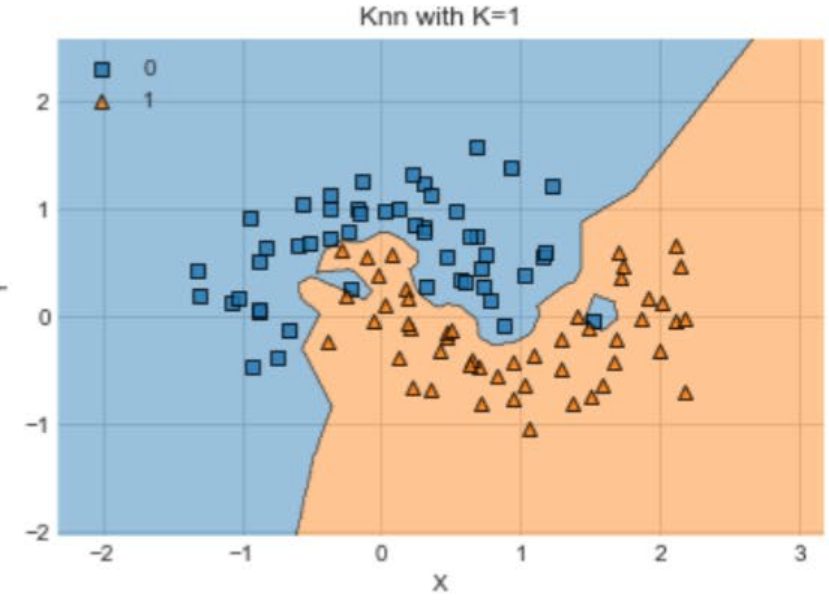
K-Nearest Neighbors (vecinos próximos)

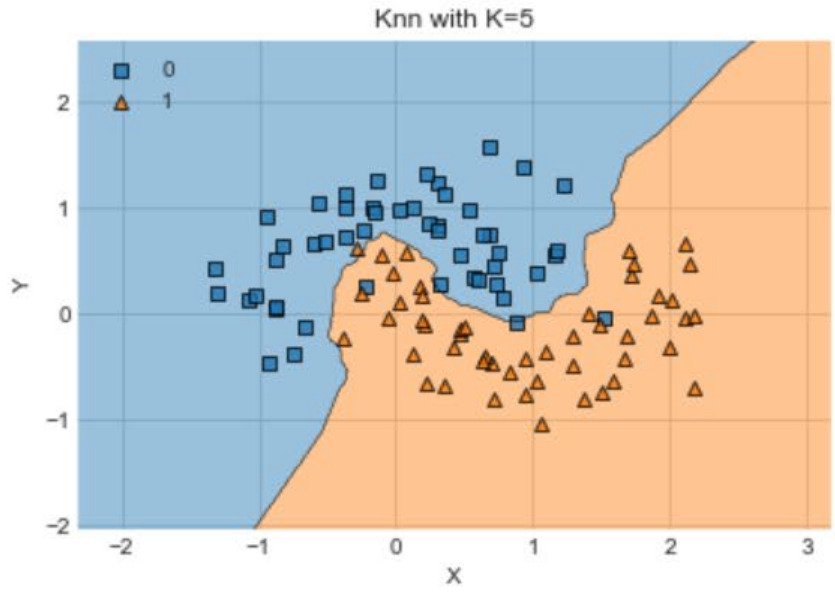
- Método no paramétrico
- Clasificador basado en los vecinos mas cercanos
    - Buscar vecinos mas cercanos, k (impar) vecinos mas cercanos
    - Se le asigna la clase mas común de los vecinos encontrados
- Predice la clase especifica acorde a una “votación” de los vecinos más cercano al individuo a clasificar
- Tiene un costo de memoria alto

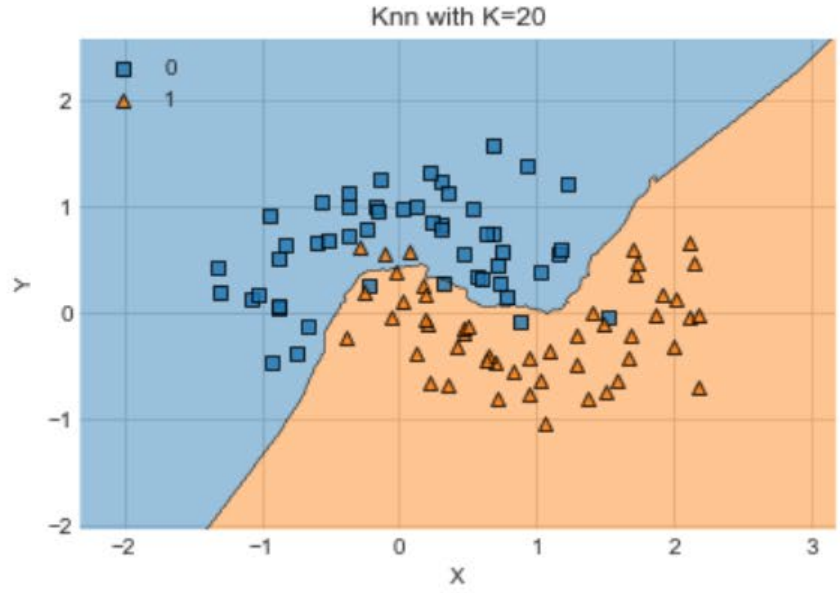

Class A
Class B
x_1
x_2
$k = 3$
$k = 6$

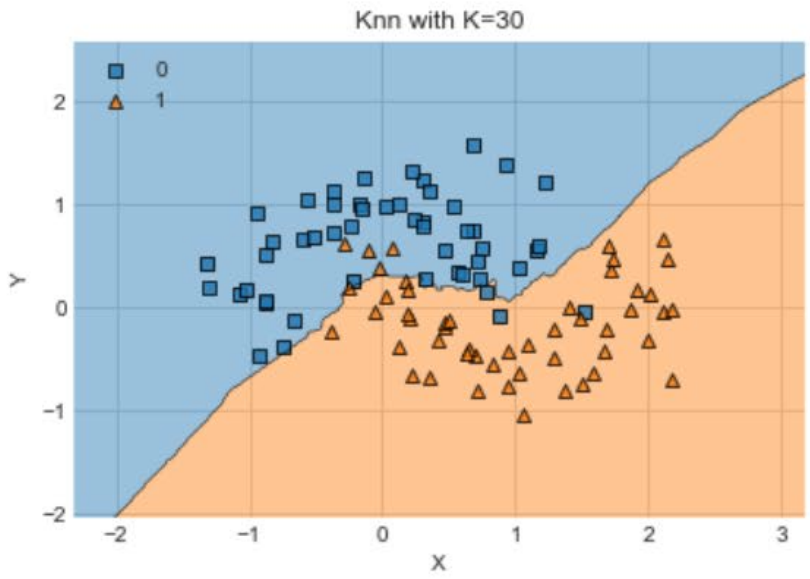

Class A
Class B
x1
x2
k = 3
k = 6

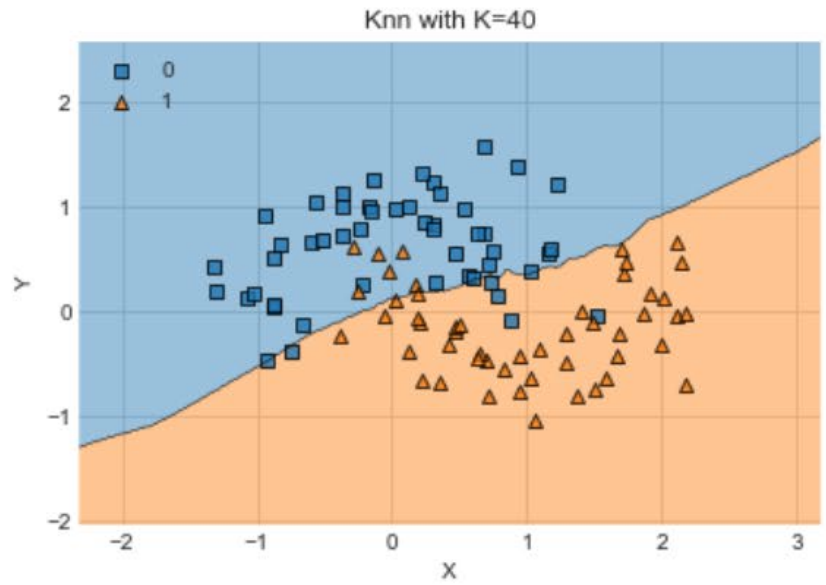

Class A
Class B
x_1
x_2
$k = 3$
$k = 6$

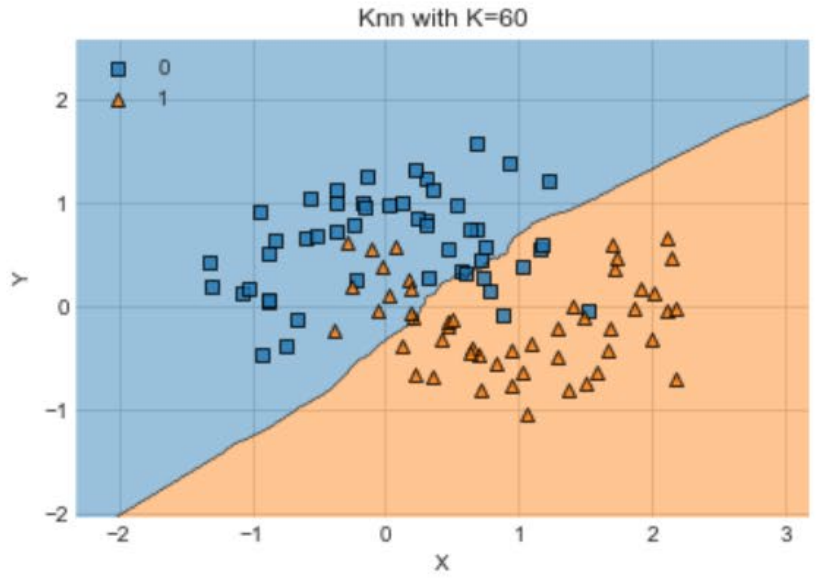

Knn with K=1
0
1
X
Y


Knn with K=5
0
1
X
Y

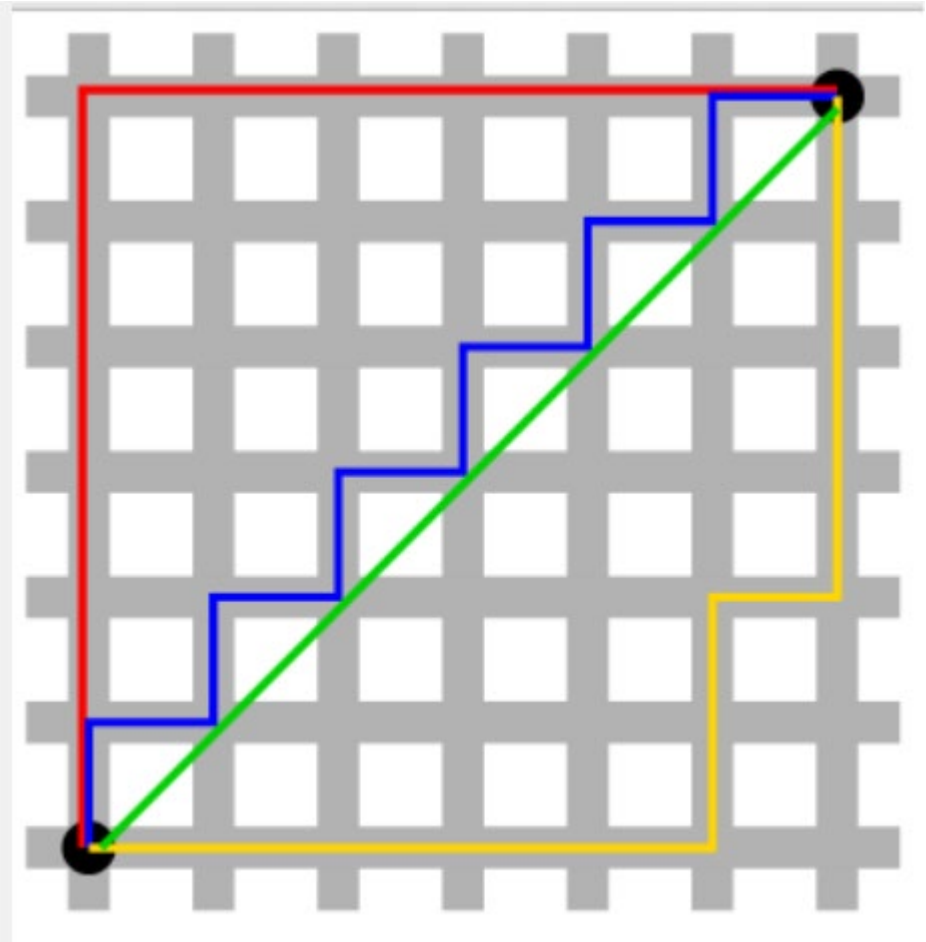

Knn with K=20
0
1
X
Y


Knn with K=30
0
1
X
Y


Knn with K=40
0
1
X
Y


Knn with K=60
0
1
X
Y

Distancias más usadas

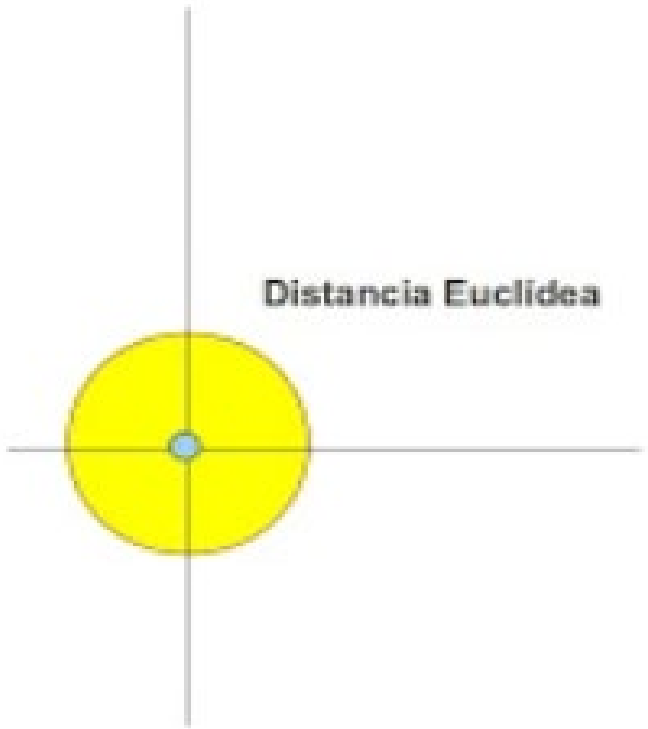


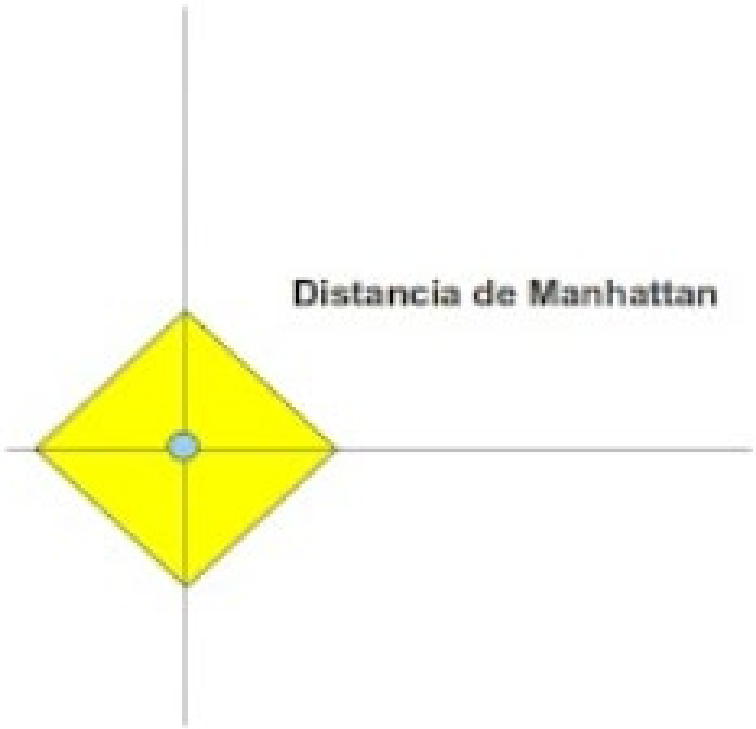
- Distancia euclidea

$$dist(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2}$$

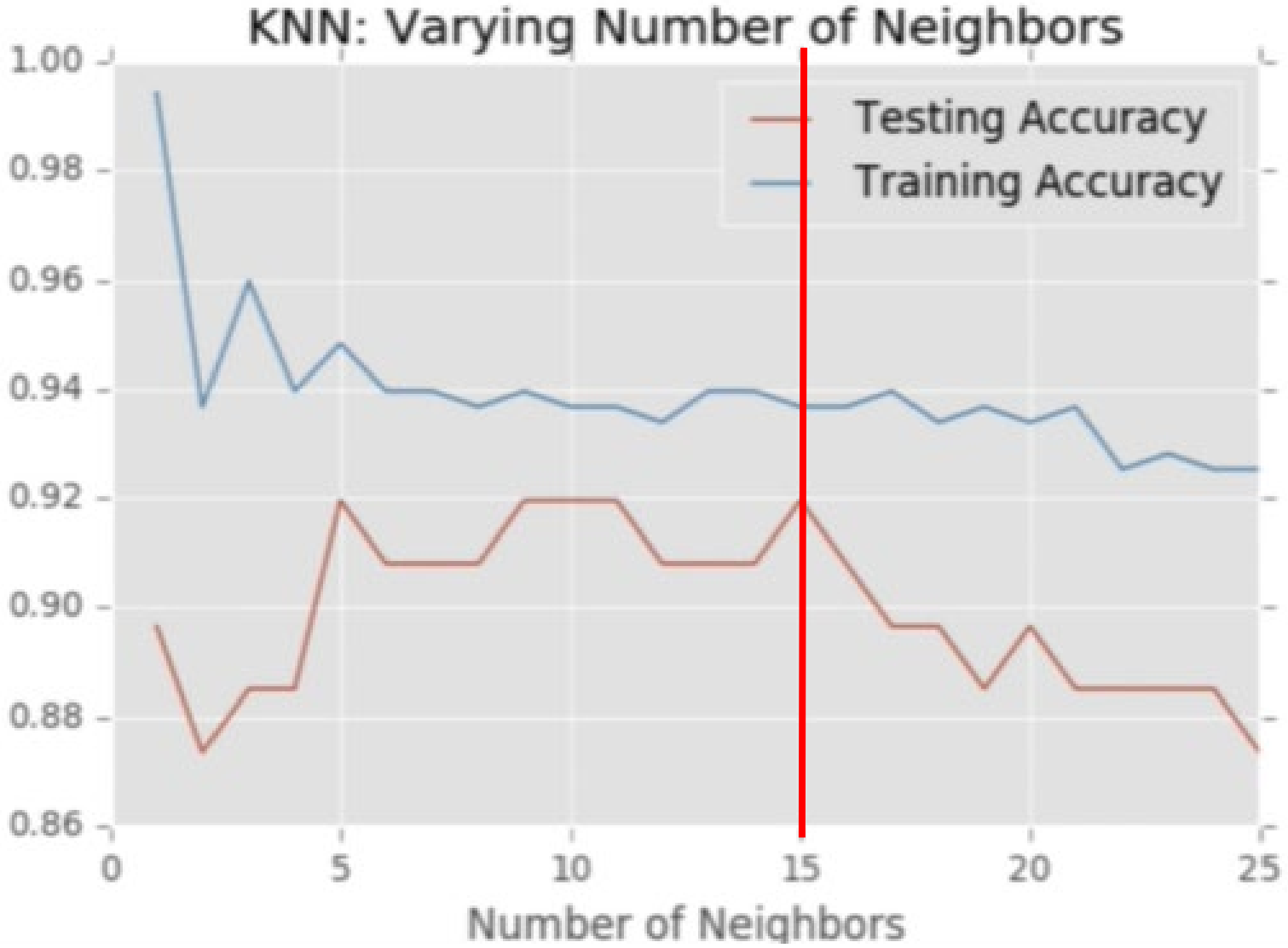
- Distancia manhattan

$$dist(x, y) = (x_1 - y_1) + (x_2 - y_2) + \ldots + (x_n - y_n)$$


Distancia Euclídea


Distancia de Manhattan

¿Cómo elegir el k?



Ventajas	Desventajas
Muy simple	Consume mucha memoria (malo para datos masivos)
Puedes utilizar muchas clases	No va bien con datos con muchas dimensiones
Fácil de añadir mas datos	No va bien con variables categóricas
Pocos parámetros	

Tarea

1. Realizar un análisis exploratorio de los datos.
2. Aplicar tres de los algoritmos desarrollados en clases justificando su elección.
3. Aplicar evaluación de cada uno de ellos con medidas vistas en clases y cross validation.
4. Seleccionar el mejor método para el problema presentado.

Tarea (CASEN 2017)

- *Feature:*
  - *Sexo [sexo]*
  - *Edad [edad]*
  - *Estado civil [ecivil]*
  - *Escolaridad [esc]*
  - *Nivel educacional [educ]*
  - *Dependencia administrativa [depen]*
  - *Condición de actividad [activ]*
  - *Indicador de materialidad [indmat]*
  - *Indicador de Saneamiento[indsan]*
  - *Calidad global de la vivienda [calglobviv]*
  - *Hacinamiento [hacinamiento]*
  - *Sistema de salud previsional [s12]*
  - *Tipo de contrato [o16]*
  - *Empleado[o1]*
  - *Ocupación u Oficio[oficio1]*
  - *Ingreso Total[ytot]*

Variable	Descripción de la Variable	Categoría observada
pobreza	**Situación de pobreza por ingresos**	1 Pobres extremos
		2 Pobres no extremos
		3 No pobres
		Blancos
		Total

Entrega

- Formato de entrega: Notebook con código python y desarrollo del análisis.
- Fecha de entrega: Lunes 15 Julio 23:59
- Tarea deben enviarlas a mi correo: claudia.chavez@usach.cl/cchavezo@gmail.com

Evaluación

- Se entregará feedback en Notebook

Realiza análisis exploratorio de datos, explicando los pasos realizados.	10%
Justifica la selección de features y/o observaciones	5%
Aplica los tres modelos de forma apropiada	30%
Evalúa con todos los indicadores de rendimiento	20%
Utiliza cross validation	15%
Interpreta bien los indicadores	20%