Ejemplo 1.- Consideraremos el siguiente ejemplo del algoritmo K vecinos más cercano, teniendo en cuenta la siguiente tabla de frecuencias, donde tenemos 3 patrones (documentos) de la clase 1 y dos de la clase 2 en el conjunto de entrenamiento. En esos documentos analizamos las veces que aparecen las palabras amor, beso, inspector y asesino.

entrenamiento	Clase 1			Clase 2		nuevos documentos	
palabras	d_1	d_2	d_3	d_4	d_5	d_6	d_7
amor	10	8	7	0	1	5	1
beso	5	6	4	1	0	6	0
inspector	2	0	0	12	8	2	12
asesino	0	1	0	20	56	0	4

Consideraremos un algoritmo 1-NN ponderado, pero considerando sólo el patrón d_1 de la clase C1 y el patrón d_4 de la clase C2

1. Para ello reemplazamos los valores de las características o variables independientes con los pesos **w** correspondientes asociados al logaritmo, considerando como pesos la proporción de veces que las palabras aparecen en los textos del conjunto de entrenamiento.

2. Luego normalizamos los vectores ponderados con los pesos **w**, de d_1, d_4, d_6 y d_7 y calculamos las distancias euclideas entre cada uno de los documentos de test d_6, d_7 y cada uno de los documentos de entrenamiento de las clases 1 y 2

Solución.-

1. “amor” y “beso” aparecen en 4 de 5 documentos, “inspector” y “asesino” en 3 de 5. En consecuencia, para las dos primeras palabras, multiplicamos las frecuencias por log5/4 = 0,1 y para las dos últimas, multiplicamos por log5/3 = 0,22.

entrenamiento	Clase 1			Clase 2		nuevos documentos	
palabras	d_1	d_2	d_3	d_4	d_5	d_6	d_7
amor	10*0,1=1	0,80	0,70	0	0,10	0,50	0,10
beso	10*0,1=0,50	0,60	0,40	0,10	0	0,60	0
inspector	2*0,22=0,44	0	0	2,64	1,76	0,22	2,64
asesino	0	0,22	0	4,4	12,32	0	0,88

2, vectores normalizados para d_1 (división por 1,2), d_4 (división por 5,13), d_6 (división por 0,81) y d_7 (división por 2,78):

normalización

$$d_1^* = \frac{d_1}{\|d_1\|} = \frac{d_1}{\sqrt{1^2 + 0,5^2 + 0,44^2}} = \frac{d_1}{1,2} \quad d_4^* = \frac{d_4}{\|d_4\|} = \frac{d_4}{\sqrt{0^2 + 0,1^2 + 2,64^2 + 4,4^2}} = \frac{d_4}{5,13}$$

$$d_6^* = \frac{d_6}{\sqrt{0,5^2 + 0,6^2 + 0,22^2 + 0^2}} = \frac{d_6}{0,81} \quad d_7^* = \frac{d_7}{\sqrt{0,1^2 + 0^2 + 2,64^2 + 0,88^2}} = \frac{d_7}{2,78}$$

palabras	d^*_1	d^*_4	d^*_6	d^*_7

amor	0,83	0	0,62	0,04
beso	0,42	0,02	0,74	0
inspector	0,37	0,51	0,27	0,95
asesino	0	0,86	0	0,32

Distancias euclideas:
d^*_1 y d^*_6: $\sqrt{0,1638} = 0,40$
d^*_4 y d^*_6: $\sqrt{1,4101} = 1,19$
d^*_1 y d^*_7: $\sqrt{1,2393} = 1,11$
d^*_4 y d^*_7: $\sqrt{0,4872} = 0,70$
Luego el patrón d_6 en un 1-NN pertenece a la clase C1, y el patrón d_7 a la clase C2

Ejercicio 2.- Supongamos que tenemos dos clases, A y B, y un nuevo documento d para clasificar. Los siguientes datos de entrenamiento están disponibles:

d_i	clase	$\cos(\mathbf{v}(d_i), \mathbf{v}(d))$
d_1	A	1
d_2	B	0,95
d_3	B	0,94
d_4	A	0,45
d_5	A	0,40
d_6	B	0,39

Supongamos que usamos el coseno como medida de distancia, es decir, cuanto más alto sea el coseno, más cerca están dos vectores.

Si nos nos dieran los datos del coseno y si los puntos, la ecuación es la siguiente

$$\cos(\mathbf{v}(d_i), \mathbf{v}(d)) = \frac{\langle \mathbf{v}(d_i), \mathbf{v}(d) \rangle}{|\mathbf{v}(d_i)||\mathbf{v}(d)|} ,$$

donde |.| es el modulo del vector y < , > es el producto escalar de los dos vectores

¿Qué clase se asignaría a d con un clasificador k vecinos cercanos usando el coseno si
i) k = 3 y el voto es mayoritario simple;
ii) k = 5 y el voto es mayoritario simple;
iii) k = 3 con una puntuación ponderada;
iv) k = 5 con una puntuación ponderada.

Solución:
i) k = 3 y voto mayoritario simple: puntuación (A; d) = de 3, 1, puntuación (B; d) = de 3, 2, por lo tanto se le asignaría la clase B.
ii) k = 5 y voto mayoritario simple: puntaje (A; d) = de 5, 3, puntaje (B; d) = de 5, 2, por lo tanto, se le asignaría la clase A.
iii) k = 3 y una puntuación ponderada : puntuación (A; d) = 1, de d1, puntuación (B; d) = 0,95 + 0,94, de d_2 y d_3, por lo tanto, se le asignaría la clase B.
iv) k = 5 y una puntuación ponderada: puntuación (A; d) = 1 + 0,45 + 0,4=1,85, (de d_1, d_4 y d_5), puntuación (B; d) = 0,95 + 0,94=1,89, (de d_2 y d_3) por lo tanto, se le asignaría la clase B.

Ejercicio 3.-

Una costurera ha perdido la información de género de uno de sus clientes, y no sabe si hacer una falda o un pantalón. Ella planea lanzar una moneda al aire.
¿En función de la información de otros clientes, cuál debería de ser una mejor decisión usando un clasificador KNN?
El cliente al que le falta información de género: Género ----, cintura 28, cadera 34

Utilice el algoritmo K-NN para K = 3 vecinos más cercanos y complete la tabla para tomar la decisión.

Género	cintura (cm)	cadera (cm)	Distancia euclidea al cuadrado	nº de orden	pertenece al entorno (Si o No)	género del entorno
Hombre	28	32	$(28-28)^2+(34-32)^2=4$	2º	Si	H
Hombre	33	35	$(28-33)^2+(34-35)^2=26$	4º	No	
Mujer	27	33	$(28-27)^2+(34-33)^2=2$	1º	Si	M
Mujer	31	36	$(28-31)^2+(34-36)^2=13$	3º	Si	M
Patrón nuevo	28	34				

nº de miembros masculinos del vecindario: 1, de 3
nº de miembros femeninos del vecindario: 2, de 3
Clase basada en el voto mayoritario, género que recibe más visitas: Mujer

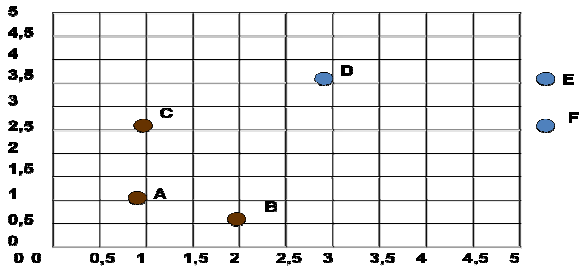
Ejercicio 4.-
Sea el siguiente conjunto de datos, con patrones de dos clases diferentes:

Nombre	X_1	X_2	Clase
A	1,0	1,0	1
B	2,0	0,5	1
C	1,0	2,5	1
D	3,0	3,5	2
E	5,5	3,5	2
F	5,5	2,5	2

Realizamos la clasificación 1-NN con validación cruzada de dejar uno fuera en los datos en el gráfico.
a) Calcule la distancia entre cada punto y su vecino más cercano utilizando la norma L_1 como medida de distancia, siendo

$$d_1(X,Y)=\sum_{i=1}^{n}|X_i-Y_i|$$

Si se dibujan los puntos, se puede identificar al vecino más cercano sin calcular todas las distancias:



$d_1(\text{A},\text{B}) = |1-2| + |1-0,5| = 1,5$; $d_1(\text{E},\text{F}) = |5,5-5,5| + |3,5-2,5| = 1$

norma L_1	A	B	C	D	E	F	nn	clase
A	0,00	**1,50**	**1,50**	4,50	7,00	6,00	B/C	1
B	**1,50**	0,00	3,00	4,00	6,50	5,50	A	1
C	**1,50**	3,00	0,00	3,00	5,50	4,50	A	1
D	4,50	4,00	3,00	0,00	**2,50**	3,50	E	2
E	7,00	6,50	5,50	2,50	0,00	**1,00**	F	2
F	6,00	5,50	4,50	3,50	**1,00**	0,00	E	2

Calcule la distancia entre cada punto y su vecino más cercano utilizando la norma L_2 como medida de distancia

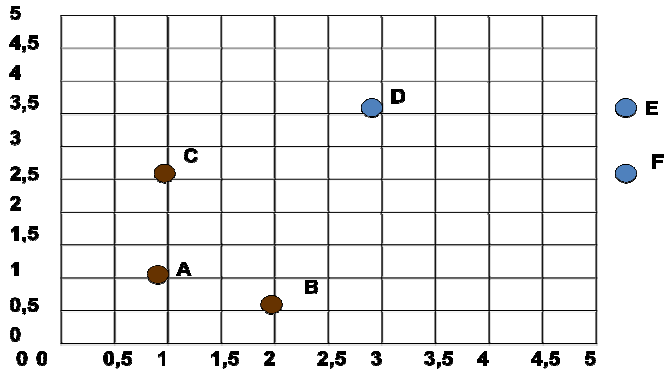
$$d_2(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}$$

$d_2(\text{A},\text{B}) = \sqrt{(1-2)^2 + (1-0,5)^2} = 1,12$; $d_2(\text{E},\text{F}) = \sqrt{(5,5-5,5)^2 + (3,5-2,5)^2} = 1$

norma L_2	A	B	C	D	E	F	nn	Clase
A	0,00	**1,12**	1,50	3,20	5,15	4,74	B	1
B	**1,12**	0,00	2,24	3,16	4,61	4,03	A	1
C	**1,50**	2,24	0,00	2,24	4,61	4,50	A	1
D	3,20	3,16	**2,24**	0,00	2,50	2,69	C	2
E	5,15	4,61	4,61	2,50	0,00	**1,00**	F	2
F	4,74	4,03	4,50	2,69	**1,00**	0,00	E	2

¿Qué puede decir sobre la clasificación si compara las dos medidas de distancia?

Las diferentes medidas de distancia pueden dar como resultado un vecino más cercano diferente y cambiar la clase a la que se asigna un punto. El punto D está más cerca de E con respecto a la norma L_1, pero más cercano a B con respecto a la norma L_2.



En cuanto a las medidas de distancia, siempre se verifica que L2 ≤ L1, esto es

$$d_2(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \leq \sum_{i=1}^{n}|X_i - Y_i| = d_1(X,Y)$$

Ejercicio 5.-
Considere un conjunto de datos con 3 clases C = {A, B, C}, con la siguiente distribución de clases N_A = 16, N_B = 32, N_C = 64. Usamos un clasificador k-NN no ponderado y establecemos k para que sea igual al número de datos, es decir, k = N_A + N_B + N_C = N
a) ¿Qué podemos decir sobre la predicción para un nuevo dato x?

Sol.-
Se clasificará como clase C. Cuando k es igual al número de puntos de datos, la vecindad de un nuevo dato contiene todos los datos del conjunto de entrenamiento, independientemente de su distancia. La clase mayoritaria en el vecindario es, por lo tanto, igual a la clase mayoritaria en el conjunto de datos.

b) ¿Qué ocurriría si utilizásemos la versión ponderada (por distancia) de k-NN?

Sol.-
Para la variante ponderada de distancia no tenemos suficiente información para responder la pregunta, ya que la distribución ponderada depende de las distancias.