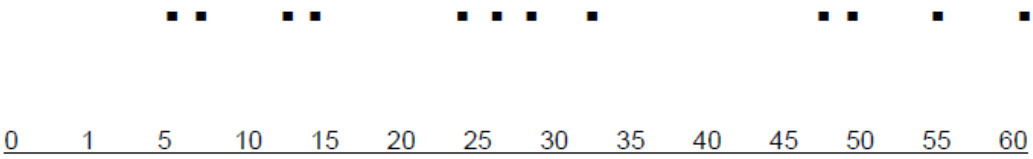
Clasificación por distancias

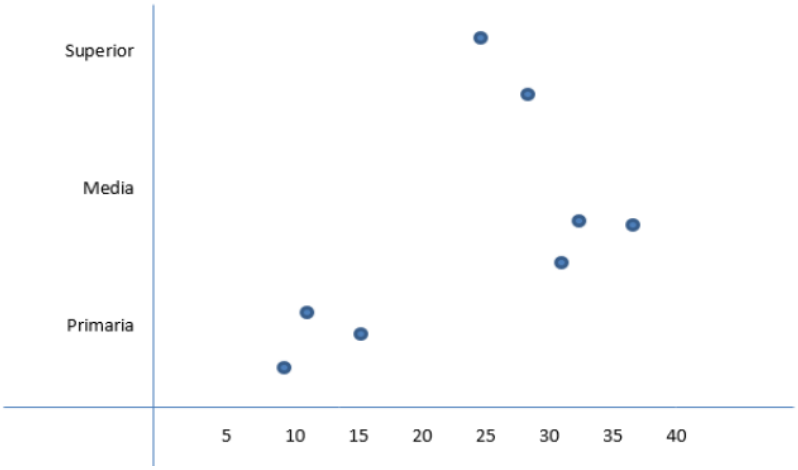
La clasificación por distancias se basa en la idea de que entre más próximos estén los elementos en el espacio, más semejantes serán y por ende, más afines a clasificarse dentro del mismo grupo. Esto por supuesto puede no ser cierto en una amplia variedad de casos; sin embargo, dentro de la llamada minería de datos, es posible retomar la idea a partir de que se comparen atributos afines entre todos los participantes.

Para ilustrar lo anterior, supongamos un grupo de personas de las cuales se compara el atributo edad. Cada una de las personas se representa mediante un punto sobre una recta numérica. Para el caso, se toman 12 personas.



Los cuatro primeros puntos están próximos entre sí, al igual que los segundos y terceros grupos de cuatro. Podemos inferior entonces tres grupos o "clústeres" y llamar al primero "jóvenes", al segundo "adultos" y al tercero "adultos mayores". Geométricamente, a menor distancia, mayor afinidad a un clúster.

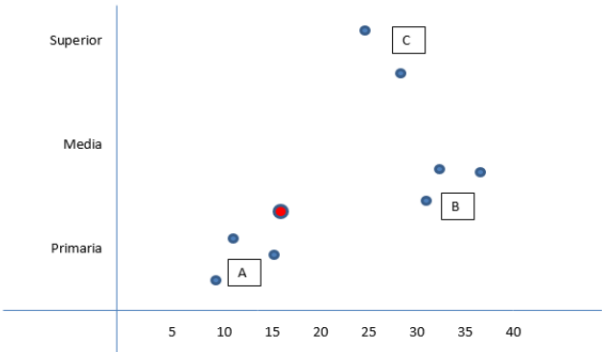
Si se tomara además del atributo edad, el atributo "grado académico", tendríamos un diagrama similar al siguiente:



Con dos atributos, el plano se convierte en bidimensional. Los puntos siguen representando a las personas, el eje "X" contiene el atributo edad y el eje "Y", el atributo nivel académico. Cada punto empareja una edad con un nivel académico para cada persona. Los puntos más cercanos entre sí pueden clasificarse en un grupo determinado.

Entonces ¿cómo resuelve el algoritmo de las distancias el problema de la clasificación? La respuesta es: Dado un punto x que representa a una instancia no clasificada o invisible, dicha instancia se clasificará en el grupo al cual se encuentre más cercano.

Por ejemplo, en la siguiente gráfica, vemos que la instancia no clasificada o invisible (que se resalta con un tamaño mayor), se clasifica en el grupo A:

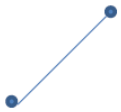


Para calcular la distancia entre las diversas instancias, se puede hacer uso de diferentes métodos. Aquí se explican tres: la distancia Euclides, la Manhattan y la máxima dimensión.

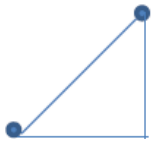
Distancia Euclides

La distancia de Euclides se calcula a partir del conocido teorema de Pitágoras.

Sean dos puntos, colocados en el plano de esta forma:



La menor distancia entre ambos puntos es una línea recta.



Para calcular la longitud de dicha línea; podemos equipararla a una hipotenusa de un triángulo rectángulo y calcular su longitud a partir de la suma de los cuadrados de los catetos de dicho triángulo:

Los catetos son las coordenadas x y y respectivas de cada punto en su plano cartesiano. Esto equivale, utilizando el lenguaje de Inteligencia de Negocios, a valores de los atributos de las instancias.

Longitud de cateto 1: x2 – x1

Longitud de cateto 2: y2 – y1

Distancia de Euclides: $\sqrt{(x2-x1)^2+(y2-y1)^2}$

3.3.2 Distancia Manhattan

La distancia Manhattan es una variación de la distancia de Euclides, que toma simplemente la suma de la longitud de los catetos. Está inspirada en la distancia entre dos lugares urbanos, donde para ir de un punto a otro, es necesario doblar varias esquinas o "cuadras".

Longitud de cateto 1: $x2 - x1$

Longitud de cateto 2: $y2 - y1$

Distancia Manhattan = $|x2 - x1| + |y2 - y1|$

Las barras verticales simbolizan "valor absoluto", ya que una longitud no tiene signo negativo.

3.3.3 Distancia Máxima Dimensión

La distancia máxima dimensión, consiste en considerar como distancia entre los dos puntos, el mayor valor de los sumandos de la distancia Manhattan:

Distancia Máxima Dimensión: Max ($|x2 - x1|$, $|y2 - y1|$)

Donde Max es una función que devuelve el mayor valor de sus argumentos.

1. Una decisión relacionada con la compra o no de un producto toma en cuenta como parámetros el precio de compra, la distancia recorrida para el surtido, el tiempo de caducidad y el tiempo de garantía. La siguiente tabla de entrenamiento muestra posibles decisiones en base a los parámetros mencionados

precio de compra	distancia recorrida surtido	tiempo de caducidad	tiempo de garantía	Decisión
25	187	5	2	no comprar
27	192	7	1	no comprar
18	200	6	3	no comprar
19	215	6	3	comprar
20	178	7	2	no comprar
22	192	8	2	comprar
26	193	9	1	no comprar
28	182	4	2	comprar
27	173	5	1	comprar
25	189	3	2	comprar
20	190	8	3	comprar
30	187	6	3	no comprar
29	197	4	3	no comprar
28	174	9	1	no comprar
18	186	8	2	no comprar
20	189	7	2	no comprar
17	191	8	1	comprar
33	129	9	1	comprar
32	130	9	3	comprar
27	199	7	2	no comprar

Se pide:

a) Especificar la decisión a tomar en base a: precio de compra: 28; distancia recorrida: 147; tiempo de caducidad: 1; tiempo de garantía: 2. Utilice las distancias medidas según:
   - **a. distancia de Euclides.**
   - **b. distancia Manhattan.**
   - **c. distancia de la máxima dimensión.**

a. **1 – Distancia de Euclides**

Se debe calcular las respectivas distancias con respecto a la instancia invisible (28, 147, 1, 2)

$$\sqrt{(25-28)^2+(187-147)^2+(5-1)^2+(2-2)^2}$$

Esto se hace porque se toma como base las filas, esto es para la primera fila, debe hacerse lo mismo para todas las siguientes filas.

Sugerir hacer primero las operaciones internas y luego la raíz.

- Entonces al tener el resultado de todas las filas se ordenan de menor a mayor (de la más cercana a la más lejana)
- Se toman los cinco primeros valores (los 5 más cercanos) , entonces como se ve la mayoría gana, en este caso la desición es "comprar"

19.24	comprar
20.35	comprar
26.34	comprar
28.18	no comprar
32.57	no comprar

b. **distancia Manhattan.**

Para éste literal el procedimiento es el mismo, las barras significan valor absoluto y lo que se debe hacer es cambiar el signo si el resultado es negativo.

$|25-28|+|187-147|+|5-1|+|2-2|$ el resultado se calcula igual, con los 5 más cercanos o menores.

c. **distancia de máxima dimensión**

$\text{Máximo}(|25-28|, |187-147|, |5-1|, |2-2|)$ Igual, el mismo procedimiento. Así los demás y elegir según los 5 más cercanos

$\text{Máximo}(3, 40, 4, 0) = 40$

b) Si el precio de compra es de 18; la distancia recorrida de 192; el tiempo de caducidad es de 3 y el tiempo de garantía es de 5, especifique la decisión a partir de la tabla de entrenamiento normalizada y utilizando:
   a. distancia de Euclides.
   b. distancia Manhattan.
   c. distancia de la máxima dimensión.

2. La elección de una ruta de distribución depende de la proximidad del destino, nivel de tráfico (porcentaje) y de la urgencia de la entrega (escala de 5 a 10). Diferentes combinaciones de estas condiciones permiten elegir entre la ruta A, B, C o D.

proximidad	nivel tráfico	urgencia de entrega	ruta
25	90	5	A
27	92	7	C
18	20	6	C
19	15	6	D
20	78	7	B
22	92	8	A
26	93	9	A
28	82	4	B
27	73	5	B
25	89	3	C
20	90	8	C
30	87	6	C
29	97	4	B
28	74	9	B
18	86	8	A
20	89	7	A
17	91	8	D
33	29	9	A
32	30	9	D
27	99	7	D

Se pide:

c) Especificar la decisión a tomar en base a: proximidad: 28, nivel de tráfico: 88, urgencia de entrega: 5. Utilice las distancias medidas según:
   a. distancia de Euclides.
   b. distancia Manhattan.
   c. distancia de la máxima dimensión.

d) Si la proximidad es de 27, el nivel de tráfico es de 90 y la urgencia de entrega de 10, especifique la decisión a partir de la tabla de entrenamiento:
   a. distancia de Euclides.
   b. distancia Manhattan.
   c. distancia de la máxima dimensión.

e) Si la proximidad es de 56, el nivel de tráfico es de 60 y la urgencia de entrega es de 7, especifique la decisión a partir de la tabla de entrenamiento.
   a. distancia de Euclides.
   b. distancia Manhattan.
   c. distancia de la máxima dimensión.