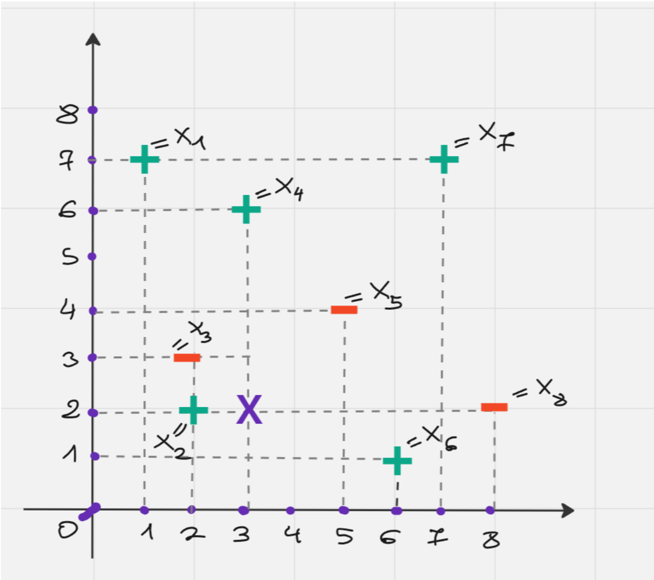
Algorytm KNN.

Klasyfikacja nowego obiektu.
Na płaszczyźnie wśród punktów o znanych wartościach "+" i "-" pojawił się nowy punkt, jak na rysunku poniżej.



Zadanie polega na przypisaniu nowego punktu X wartości "+" lub "-" w oparciu o modę (dominantę) wartości sąsiadujących z nim punktów. W tym celu:

1. Obliczamy odległości wszystkich punktów X_i od punktu X.

   $X_1 = (1,7) \Rightarrow d_1 = d(X_1, X) = ((1,7),(3,2)) = \sqrt{2^2+5^2} = \sqrt{29}$,
   $X_2 = (2,2) \Rightarrow d_2 = d(X_2, X) = 1 = \sqrt{1}$,
   $X_3 = (2,3) \Rightarrow d_3 = d(X_3, X) = \sqrt{2}$,
   $X_4 = (3,6) \Rightarrow d_4 = 4 = \sqrt{16}$,
   $X_5 = (5,4) \Rightarrow d_5 = \sqrt{8}$,
   $X_6 = (6,1) \Rightarrow d_6 = \sqrt{10}$,
   $X_7 = (7,7) \Rightarrow d_7 = \sqrt{41}$,
   $X_8 = (8,2) \Rightarrow d_8 = 5 = \sqrt{25}$.

2. Porządkujemy odległości rosnąco
   Ponieważ
   $$1 < \sqrt{2} < \sqrt{8} < \sqrt{10} < \sqrt{16} < \sqrt{25} < \sqrt{29} < \sqrt{36},$$
   to
   $$d_2 < d_3 < d_5 < d_6 < d_4 < d_8 < d_1 < d_7.$$
   Możemy więc stworzyć wektor
   $$D = [d_2, d_3, d_5, d_6, d_4, d_8, d_1, d_7]$$
   uporządkowanych odległości.

3. Ustalamy k i klasyfikujemy nowy punkt w oparciu o k najbliższych mu sąsiadów. Jeśli:

   - $k = 1$, to wybieramy pierwszego najbliższego sąsiada - jest on oddalony od X o 1. Innymi słowy, wybieramy pierwszy element wektora D. Ten element to d_2. Oznacza to, że najbliższym sąsiadem jest X_2, którego etykieta y_2 to "+", zatem w tym przypadku etykieta nowego punktu to "+". Sytuację tę można przedstawić schematycznie następująco:
   $$k = 1 \to D[0] = d_2 \mapsto X_2 \mapsto y_2 = " - " \Rightarrow y = y_2 = " + ".$$

- $k = 2$, to wybieramy dwóch najbliższych sąsiadów. Oznacza to, że interesują nas 2 pierwsze elementy wektora D. Zatem

$$k = 2 \to D[0,1] = [d_2, d_3] \mapsto [X_2, X_3] \mapsto [y_2, y_3] = ["+", "-"].$$

Jak widać, jest to sytuacja nierozstrzygająca, możemy wtedy np. zwiększyć liczbę sąsiadów.

- $k = 3$, co oznacza, że wybieramy 3 pierwsze elementy wektora D. Zatem:

$$k = 3 \to D[0,1,2] = [d_2, d_3, d_5] \mapsto [X_2, X_3, X_5] \mapsto [y_2, y_3, y_5].$$

A ponieważ

$$[y_2, y_3, y_5] = ["+", "-", "+"],$$

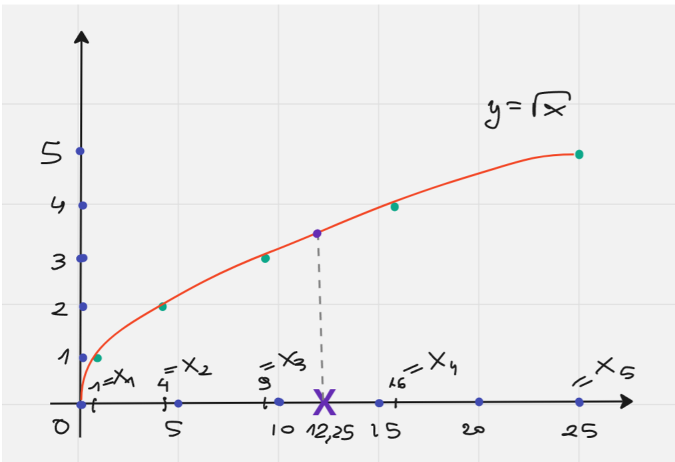
to

$$y = M_o(y_2, y_3, y_5) = "+",$$

gdzie M_o oznacza dominantę danego zbioru.

Zadanie regresji.

Regresja polega na przewidywaniu wartości ciągłej zmiennej zależnej w oparciu o znane wartości zmiennych niezależnych.



Zadanie polega na wyznaczeniu wartości dla nowego punktu $X = 12,25$. W tym celu:

1. Powtarzamy kroki 1 i 2, aby otrzymać uporządkowany wektor odległości

$$D = [d_3, d_4, d_2, d_1, d_5],$$

gdzie
$d_3 = d(X, X_3) = 3,25$,
$d_4 = d(X, X_4) = 3,75$,
$d_2 = d(X, X_2) = 8,25$,
$d_1 = d(X, X_1) = 12,25$,
$d_5 = d(X, X_5) = 12,75$.

2. Następnie wybieramy k i ustalamy wartość nowego punktu jako średnią arytmetyczną μ z wartości k najbliższych sąsiadów.

   - Dla $k = 1$ wybieramy pierwszego najbliższego sąsiada. Ponieważ średnia arytmetyczna z jednej liczby jest równa tej liczbie to
   $$k = 1 \to D[0] = d_3 \mapsto X_3 \mapsto y_3 = 3,25 \Rightarrow y = y_3 = 3,25.$$
   - Dla $k = 2$ obliczamy średnią arytmetyczną z dwóch najbliższych sąsiadów a zatem
   $$k = 2 \to D[0,1] = [d_3, d_4] \mapsto [X_3, X_4] \mapsto [y_3, y_4] = [3,4] \Rightarrow y = \mu(y_3, y_4) = \frac{3+4}{2} = 3,5.$$
   - Analogicznie dla $k = 3$, $y = \frac{y_3+y_4+y_2}{3} = \frac{3+4+2}{3} = 3$, itd.