K-Nearest Neighbors

KNN este unul din algoritmii de invatare supervizata de machine learning, utilizat cu scopul clasificarii datelor.

Primul pas este reprezentat de definirea unui set de date:

1, y1), 2, y2),3, y3), ...,n, yn),

unde yi reprezinta clasa din care face parte vectorul de atribute i . Predictiile pentru o instanta  se fac cautand prin setul de date cele mai similare K instante (vecini), pe baza carora se va prezice carei clase ii apartine vectorul .

Pentru determinarea a K vecini ale caror atribute sunt cele mai similare cu vectorul de intrare, se realizeaza calculul distantei dintre acestea. Cele mai pupulare distante calculate sunt: distanta euclidiana, distanta hamming, distanta manhattan sau distanta minkowski. Alegerea distantei potrivite clasificarii se face pe baza proprietatilor setului de date folosit.

In figura de mai jos, punctul rosu reprezinta un vector de intrare neclasificat. Pe baza algoritmului, in functie de valorile variabilei k, vectorul ar putea apartine clasei B, pentru k=3 sau clasei A, pentru k=6.

