Data-Driven Problem Solving in Mechanical Engineering

K Nearest Neighbors

Masoud Masoumi

ME 364 - Spring 2022

Department of Mechanical Engineering
The Cooper Union for the Advancement of Science and Art

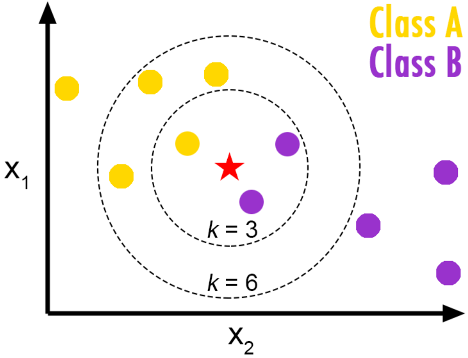
April, 2022

K Nearest Neighbors

Given a set of labeled training examples, we seek the training example which is most similar to an unlabeled point p, and then take the class label for p from its nearest labeled neighbors.

K-Nearest Neighbors is an algorithm for supervised learning, where the model is 'trained' with data points corresponding to their classes.

Once a point is to be predicted, it takes into account the 'K' nearest points to determine its class.



Implementation

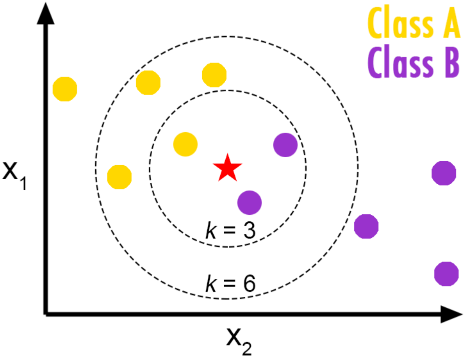
Here is how this algorithm works:

(1) Pick a value for k

(2) Calculate the distance of unknown case from all cases.

(3) Select the k points in the training data that are "nearest" to the unknown data point.

(4) Make a prediction using the most popular target variable class from the k-nearest neighbors.



Implementation

There are multiple options to find the distance in step (3), which measures the distance between points $x^{(1)}$ and $x^{(2)}$ for all n features.

- Euclidean distance
  $d = \sqrt{\sum_{i=1}^{n}(x_i^{(1)} - x_i^{(2)})^2}$
- Manhattan distance
  $d = \sum_{i=1}^{n}|x_i^{(1)} - x_i^{(2)}|$
- Minkowski distance
  $d = (\sum_{i=1}^{n}|x_i^{(1)} - x_i^{(2)}|^p)^{\frac{1}{p}}$
- Chebyshev distance
  $d = max(|x_i^{(1)} - x_i^{(2)}|)$
  $= \lim_{p\to\infty}(\sum_{i=1}^{n}|x_i^{(1)} - x_i^{(2)}|^p)^{\frac{1}{p}}$