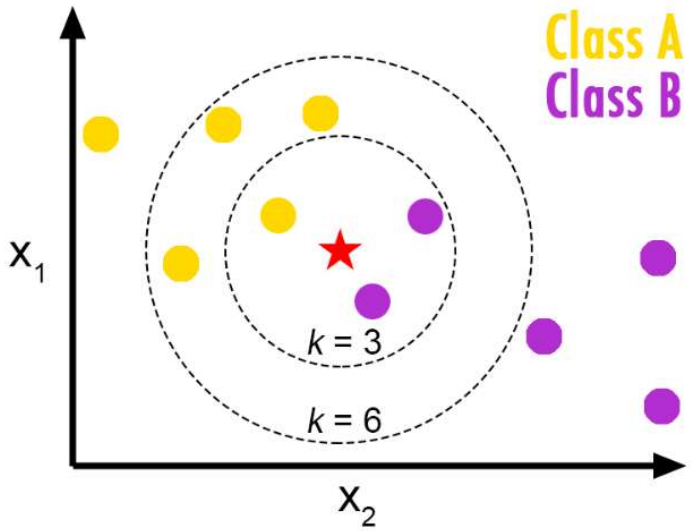
K Nearest Neighbor



About

- The name of the algorithm, originates from the philosophy of kNN – i.e., people having similar mindset tend to stay close to each other.
- In the same way, as a part of kNN algorithm, the unknown and unlabelled data which comes for a prediction problem is judged on the basis of training data set of elements which are similar to the unknown elements.
- k in kNN algorithm represents the number of nearest neighbour points which are voting for the new test data's class.

kNN Algorithm Manual Implementation

- Load the data
- Initialize K to your chosen number of neighbours
- For each example in the data
  - Calculate the distance between the query example and the current example from the data.
  - Add the distance and the index of the example to an ordered collection
- Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances.
- Pick the first K entries from the sorted collection.
- Get the labels of the selected K entries, then return mode of the lables
- setting K-Nearest Neighbor algorithm forms a majority vote between the K most similar instances to a given unseen observation.

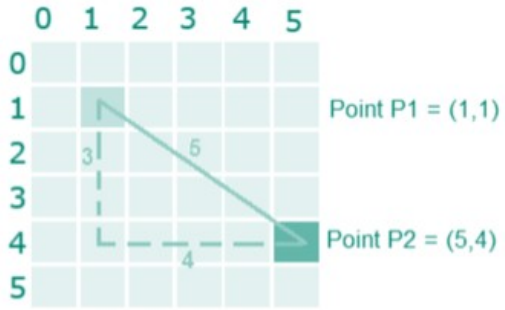
Example Dataset contains the Height and weight for T-Shirt Size

Height (in cms)	Weight (in kgs)	T Shirt Size
158	58	M
158	59	M
158	63	M
160	59	M
160	60	M
163	60	M
163	61	M
160	64	L
163	64	L
165	61	L
165	62	L
165	65	L
168	62	L
168	63	L
168	66	L
170	63	L
170	64	L
170	68	L

predict the T-shirt size of **Anna**, whose height is **161cm** and her weight is **61kg.**

Step1: Calculate the Euclidean distance between the new point and the existing points

For example, Euclidean distance between point P1(1,1) and P2(5,4) is



Euclidean distance = $\sqrt{(5-1)^2 + (4-1)^2} = 5$

Step 2: Choose the value of K and select K neighbors closet to the new point. In this case, select the top 5 parameters having least Euclidean distance

Step 3: Count the votes of all the K neighbors / Predicting Values

Advantages

- The algorithm is simple and easy to implement.
- There's no need to build a model, tune several parameters, or make additional assumptions.
- The algorithm is versatile. It can be used for classification, regression, and search.

Disadvantages

- The algorithm gets significantly slower as the number of examples and/or predictors/independent variables increase.


KNN BE LIKE
"Show me your friends, and I'll tell you who you are."