K-Nearest Neighbors Algorithm Using Python

What is KNN Algorithm?

K nearest neighbors or KNN Algorithm is a simple algorithm which uses the entire dataset in its training phase. Whenever a prediction is required for an unseen data instance, it searches through the entire training dataset for k-most similar instances and the data with the most similar instance is finally returned as the prediction.

How does a KNN Algorithm work?

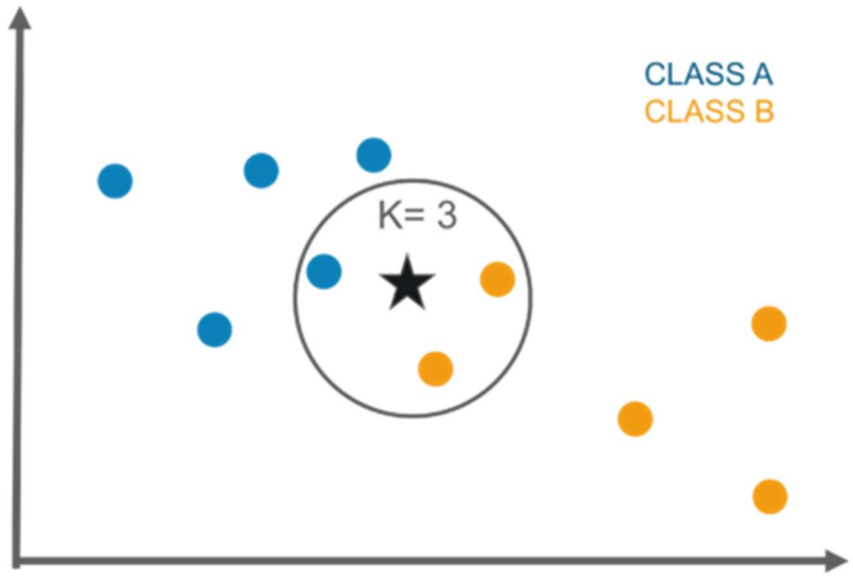
The k-nearest neighbors algorithm uses a very simple approach to perform classification. When tested with a new example, it looks through the training data and finds the k training examples that are closest to the new example. It then assigns the most common class label (among those k-training examples) to the test example.

What does 'k' in kNN Algorithm represent?

k in kNN algorithm represents the number of nearest neighbor points which are voting for the new test data's class.

If k=1, then test examples are given the same label as the closest example in the training set.

If k=3, the labels of the three closest classes are checked and the most common (i.e., occurring at least twice) label is assigned, and so on for larger ks.



kNN Algorithm Manual Implementation

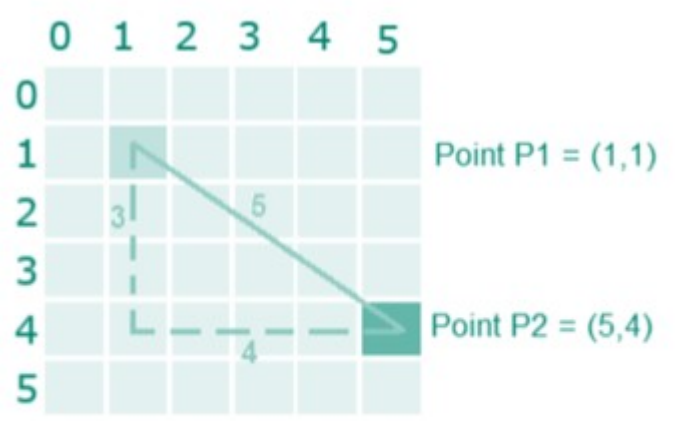
Let's consider this example,

Suppose we have height and weight and its corresponding Tshirt size of several customers. Your task is to predict the T-shirt size of Anna, whose height is 161cm and her weight is 61kg.

	A	B	C	D	E
1	Height (in cms)	Weight (in kgs)	T Shirt Size		
2	158	58	M		
3	158	59	M		
4	158	63	M		PREDICTION
5	160	59	M		Predict the T shirt size of a new customer Anna with height as 161cm and weight as 61kg
6	160	60	M		
7	163	60	M		
8	163	61	M		
9	160	64	L		
10	163	64	L		
11	165	61	L		
12	165	62	L		
13	165	65	L		
14	168	62	L		
15	168	63	L		
16	168	66	L		
17	170	63	L		
18	170	64	L		
19	170	68	L		

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:



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

SUM | =SQRT((161-A2)^2 +(61-B2)^2)

	A	B	C	D	E
1	Height (in cms)	Weight (in kgs)	T Shirt Size	Euclidean Distance	
2	158	58	M	=SQRT((161-A2)^2 +(61-B2)^2)	
3	158	59	M	SQRT(number)	
4	158	63	M		PREDICTION
5	160	59	M		Predict the T shirt size of a new customer Anna with height as 161cm and weight as 61kg
6	160	60	M		
7	163	60	M		
8	163	61	M		
9	160	64	L		
10	163	64	L		
11	165	61	L		
12	165	62	L		
13	165	65	L		
14	168	62	L		
15	168	63	L		
16	168	66	L		
17	170	63	L		
18	170	64	L		
19	170	68	L		

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

	A	B	C	D	E	F
1	Height (in cms)	Weight (in kgs)	T Shirt Size	Euclidean Distance	Ranks	
2	158	58	M	4.242640687		
3	158	59	M	3.605551275		
4	158	63	M	3.605551275		PREDICTION
5	160	59	M	2.236067977	4	with height as 161cm and weight as 61kg
6	160	60	M	1.414213562	1	
7	163	60	M	2.236067977	3	
8	163	61	M	2	2	For K = 5
9	160	64	L	3.16227766	5	Find the nearest neighbors
10	163	64	L	3.605551275		So, look for top 5 values in ascending order
11	165	61	L	4		
12	165	62	L	4.123105626		
13	165	65	L	5.656854249		
14	168	62	L	7.071067812		
15	168	63	L	7.280109889		
16	168	66	L	8.602325267		
17	170	63	L	9.219544457		
18	170	64	L	9.486832981		
19	170	68	L	11.40175425		
20						

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

Since for K = 5, we have 4 Tshirts of size M, therefore according to the kNN Algorithm, Anna of height 161 cm and weight, 61kg will fit into a Tshirt of size M.

	A	B	C	D	E	F
1	Height (in cms)	Weight (in kgs)	T Shirt Size	Euclidean Distance	Ranks	
2	158	58	M	4.242640687		
3	158	59	M	3.605551275		
4	158	63	M	3.605551275		PREDICTION
5	160	59	M	2.236067977	4	and weight as 61kg
6	160	60	M	1.414213562	1	
7	163	60	M	2.236067977	3	
8	163	61	M	2	2	For K = 5
9	160	64	L	3.16227766	5	Find the nearest neighbors
10	163	64	L	3.605551275		So, look for top 5 values in ascending order
11	165	61	L	4		
12	165	62	L	4.123105626		Since for K = 5, we have 4 M size Tshirts, so according to kNN Algorithm
13	165	65	L	5.656854249		Anna of height 161 cm and weight 61 kg will fit into a Tshirt of size M
14	168	62	L	7.071067812		
15	168	63	L	7.280109889		
16	168	66	L	8.602325267		
17	170	63	L	9.219544457		
18	170	64	L	9.486832981		
19	170	68	L	11.40175425		
20						

Implementation of kNN Algorithm using Python

- Handling the data
- Calculate the distance
- Find k nearest point
- Predict the class
- Check the accuracy