K-NN Algorithm

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems.

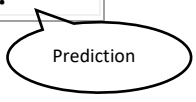
Let's take an example:

Step 1: Table of Input data -

- Most mobile devices are equipped with different kind of sensors.
- We can use the data sent from Gyroscope and Accelerometer sensors to categorize any motion:
    - 3 numbers from Accelerometer sensor
    - 3 numbers from Gyroscope sensor

- This is the training data and the test data:

Accelerometer Data			Gyroscope Data			Fall (+), Not (-)
x	y	z	x	y	z	+/-
1	2	3	2	1	3	-
2	1	3	3	1	2	-
1	1	2	3	2	2	-
2	2	3	3	2	1	-
6	5	7	5	6	7	+
5	6	6	6	5	7	+
5	6	7	5	7	6	+
7	6	7	6	5	6	+
7	6	5	5	6	7	??



Step 2:

Suppose we determine K = 8 (we will use 8 nearest neighbors) as parameter of this algorithm.

- Then we calculate the distance between the query-instance and all the training samples.
  - Because we use only quantitative X_i, we can use Euclidean distance.
- Suppose the query instance have coordinates (X_1^q, X_2^q) and the coordinate of training sample is (X_1^t, X_2^t) then square Euclidean distance is
  - $$d_{tq}^2 = (X_1^t - X_1^q)^2 + (X_2^t - X_2^q)^2$$

N = 8
K = √N =√8 = 3

Step 3: Find the distance using the formula in step 2-

Accelerometer Data			Gyroscope Data			Fall (+), Not (-)	Dist. Acce	Dist Gyro.
x	y	z	x	y	z	+/-		
1	2	3	2	1	3	-	56	50
2	1	3	3	1	2	-	54	54
1	1	2	3	2	2	-	70	45
2	2	3	3	2	1	-	45	56
6	5	7	5	6	7	+	6	0
5	6	6	6	5	7	+	5	2
5	6	7	5	7	6	+	8	2
7	6	7	6	5	6	+	4	3
7	6	5	5	6	7	+		

Step 4: Find the K-nearest neighbors –

We include a training sample as nearest neighbors if the distance of this training sample to the query instance is less than or equal to the K-th smallest distance.

If the distance of the training sample is below the K-th minimum, then we gather the category Y of this nearest neighbors' training samples.
Some special case happens in our example that the 3rd until the 8 th minimum distance happen to be the same.

- In this case we directly use the highest K=8 because choosing arbitrary among the 3rd until the 8th nearest neighbors is unstable.

The KNN prediction of the query instance is based on simple majority of the category of nearest neighbors.

- In our example, the data is only binary, thus the majority can be taken as simple as counting the number of ‘+' and ‘-‘ signs.
    - If the number of plus is greater than minus, we predict the query instance as plus and vice versa.
    - If the number of plus is equal to minus, we can choose arbitrary or determine as one of the plus or minus.