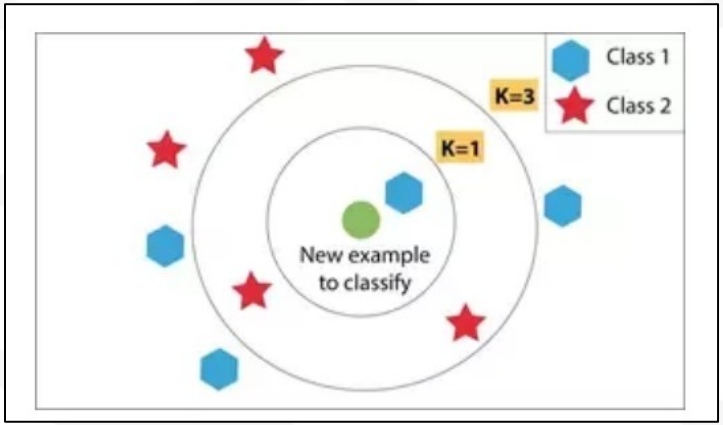
K Nearest Neighbors

K Nearest Neighbors

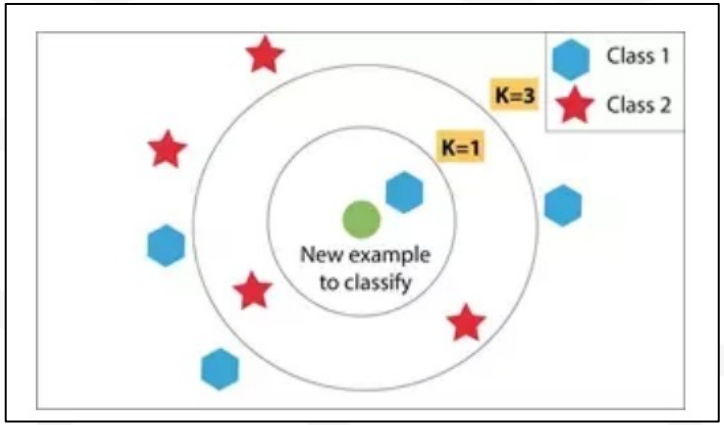
K Nearest Neighbors is typically considered a supervised learning technique that can be used for classification or regression.

We will talk about it in this context first.



K Nearest Neighbors

- The **closeness/ proximity** amongst samples of data determines their neighborhood.
- Generally, neighbors share similar characteristics and behavior that's why they can be treated as they belong to the same group.
- The K in KNN represents the K-Nearest Neighbors of the unknown data we want to classify and assign it the group appearing majorly in those K neighbors.
- For K=1, the unknown/unlabeled data will be assigned the class of its closest neighbor.



Predicting Music Taste

- We have a person named Gary who is a 23 year male and we want to predict which band he will like. Here's how we can use the KNN algorithm

- We'll have to translate gender to some numbers for the distance/ proximity relation needed for finding neighbors.

NAME	AGE	GENDER	MUSIC BAND
Amantha	19	F	Coldplay
Brendon	23	M	Coldplay
Nate	24	M	LinkinPark
Sam	30	M	LininPark
Betty	16	F	Coldplay
Christine	18	F	LinkinPark
Gin	22	F	LinkinPark
Ken	18	M	Coldplay
Susy	15	F	Coldplay

Calculating the Nearest Neighbors

$$\sqrt{(age_i - age_{Gary})^2 + (gender_i - gender_{Gary})^2}$$

NAME	AGE	GENDER	MUSIC BAND	DISTANCE^2
Amantha	19	1	Coldplay	17
Brendon	23	0	Coldplay	0
Nate	24	0	LinkinPark	1
Sam	30	0	LininPark	49
Betty	16	1	Coldplay	50
Christine	18	1	LinkinPark	26
Gin	22	1	LinkinPark	2
Ken	18	0	Coldplay	25
Susy	15	1	Coldplay	65

- Gary is Male (0) and is 23 years old
- K=3 then the K-Nearest Neighbors are

NAME	AGE	GENDER	MUSIC BAND	DISTANCE^2
Brendon	23	0	Coldplay	0
Nate	24	0	LinkinPark	1
Gin	22	1	LinkinPark	2

- LinkinPark is followed more by Gary's Neighbors so we predict that Gary will also like LinkinPark more than Coldplay

How Does This Work With Text?

- Same idea but the distances calculated are the distances between word/sentence vectors

- K-Nearest word/sentence vectors will be selected and the class having maximum frequency will be labeled to the unlabeled data.

ID	CLASS	WORD1	WORD2	WORD3
1	0	2	0	3
2	1	3	1	1
3	1	0	3	2
4	0	1	2	1

What if We Don’t Have Classes?

If our text data are unlabelled (as is often the case in NLP), we can use KNN to identify documents that are similar to a given document.

We could also use a clustering algorithm to group together similar documents (we will talk about this later on).