kNN uses proximity to the nearest neighbour to determine what they are.

K is the only parameter needed (how many “neighbours” we will see)

Do not make it too big.

Afbeelding met tekst, schermopname, scherm, software

Automatisch gegenereerde beschrijving

Problem with hyperparameter (k) is too low? You will only consider 1 node,

Problem with hyperparameter is too high? You will consider all nodes, which removes the purpose of k.

Elbow method: plot accuary vs the k value, when you see an elbow (straight horizontal line) then you know the right value for k.

kNN and Kmeans are not the same: kNN uses neighbour

Kmeans tries to groep by clustering data.

Kmeans 🡪unlabeled data, so no Y axis

How to use Kmeans Clustering:

Step 0: plot the unlabeled data, only features.

Step1: Choose the number of clusters to create (K value)

(in most situation you won’t visualize data)

Step 2: Randomly select K distinct data points: Our K = 3;

Step 3: Assign each remaining point to the nearest “cluster” point.

(it uses a distance metric)

Step 4: Calculate the center of each cluster points. (mean vale of each point vector)

Step5: Now assign each point to the nearest cluster center.

Repeat steps 4 and 5 until there are no more cluster reassignments.

How do we choose a reasonable value for k number of clusters? Same as kNN or something random.

How to check if good, sum of distances² to the cluster

Kmeans is not the only clustering algorithm

* Can detect very clear patterns