K-means and K-Nearest Neighbor

Kwangwoon University MI:RU Machine Learning Study
오민성

Contents

K-means

- Algorithms
- How to find the optimal K
- Ploblems

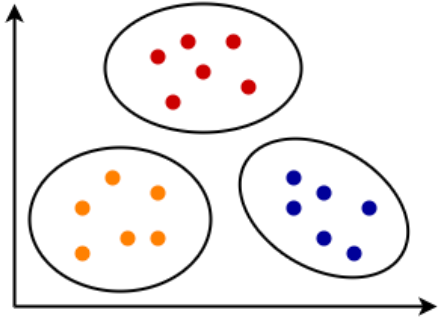
K-nearest neighbor

- Algorithms
- How to find the optimal K
- Ploblems

What algorithms is **K-means**?

- Clustering
- Unsupervised Learning

Before K-Means

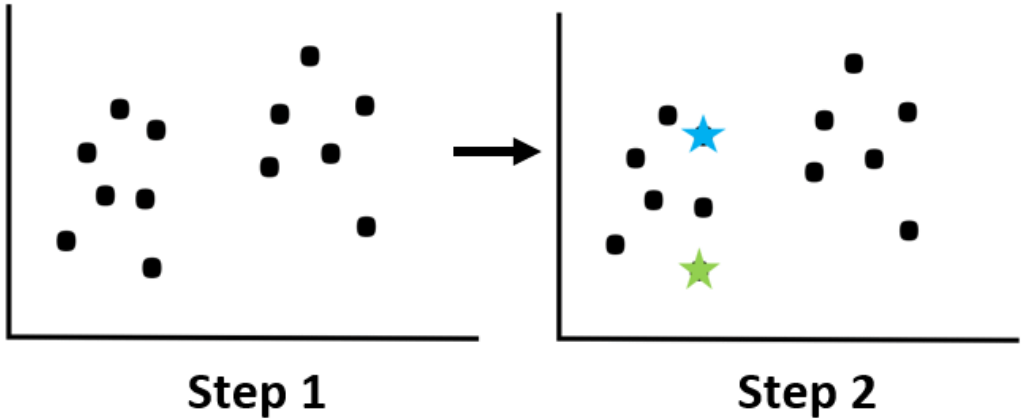


After K-Means

K-Means Clustering Algorithms

Step 1. Choose the number of clusters (K = 2)

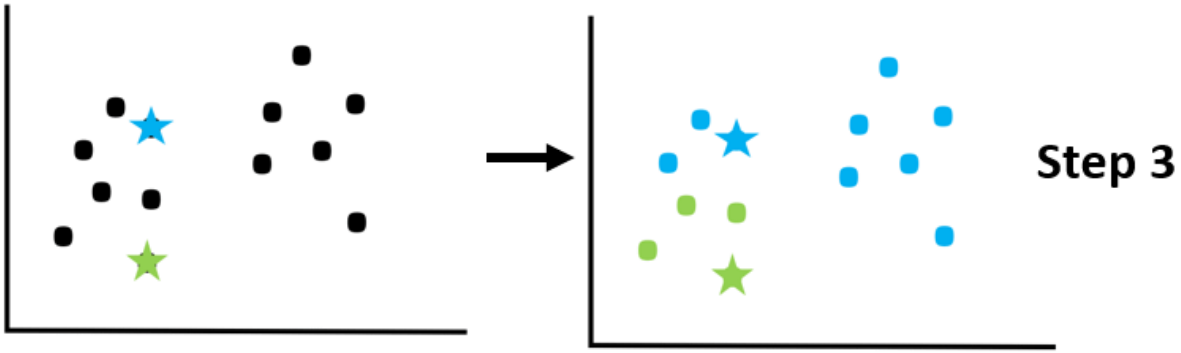
Step 2. mark 2 data points randomly for 2 clusters.



[1]Kmeans_algorithms

K-Means Clustering Algorithms

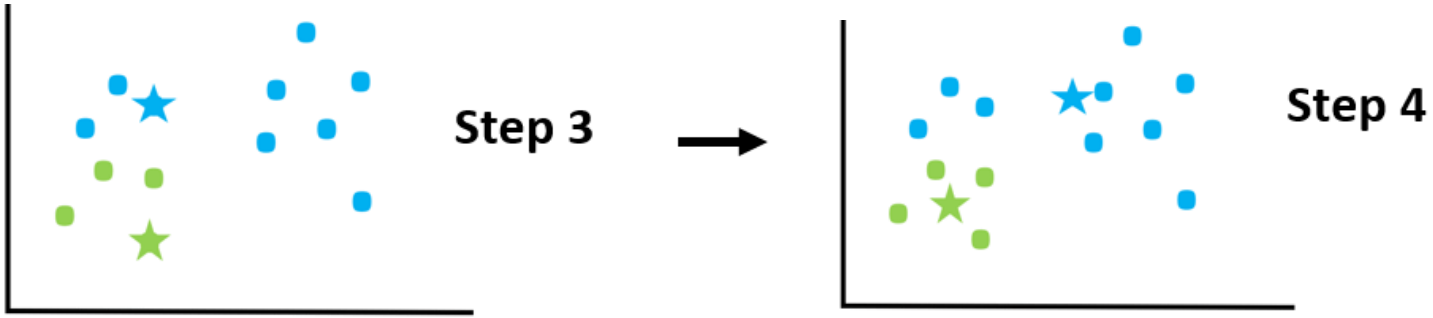
Step 3. Assign each data points to the nearest centroid.



[1]Kmeans_algorithms

K-Means Clustering Algorithms

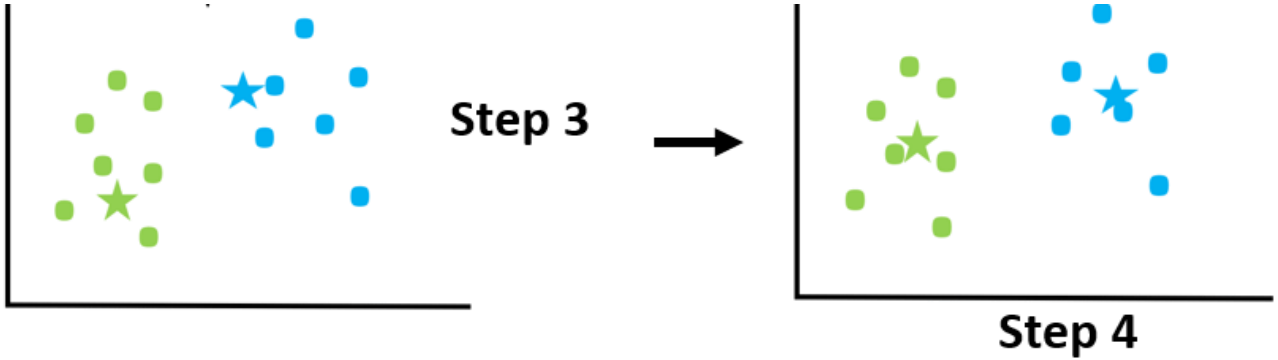
Step 4. Calculate the average of the data points belonging to each cluster and move its cluster centroid to the average location.



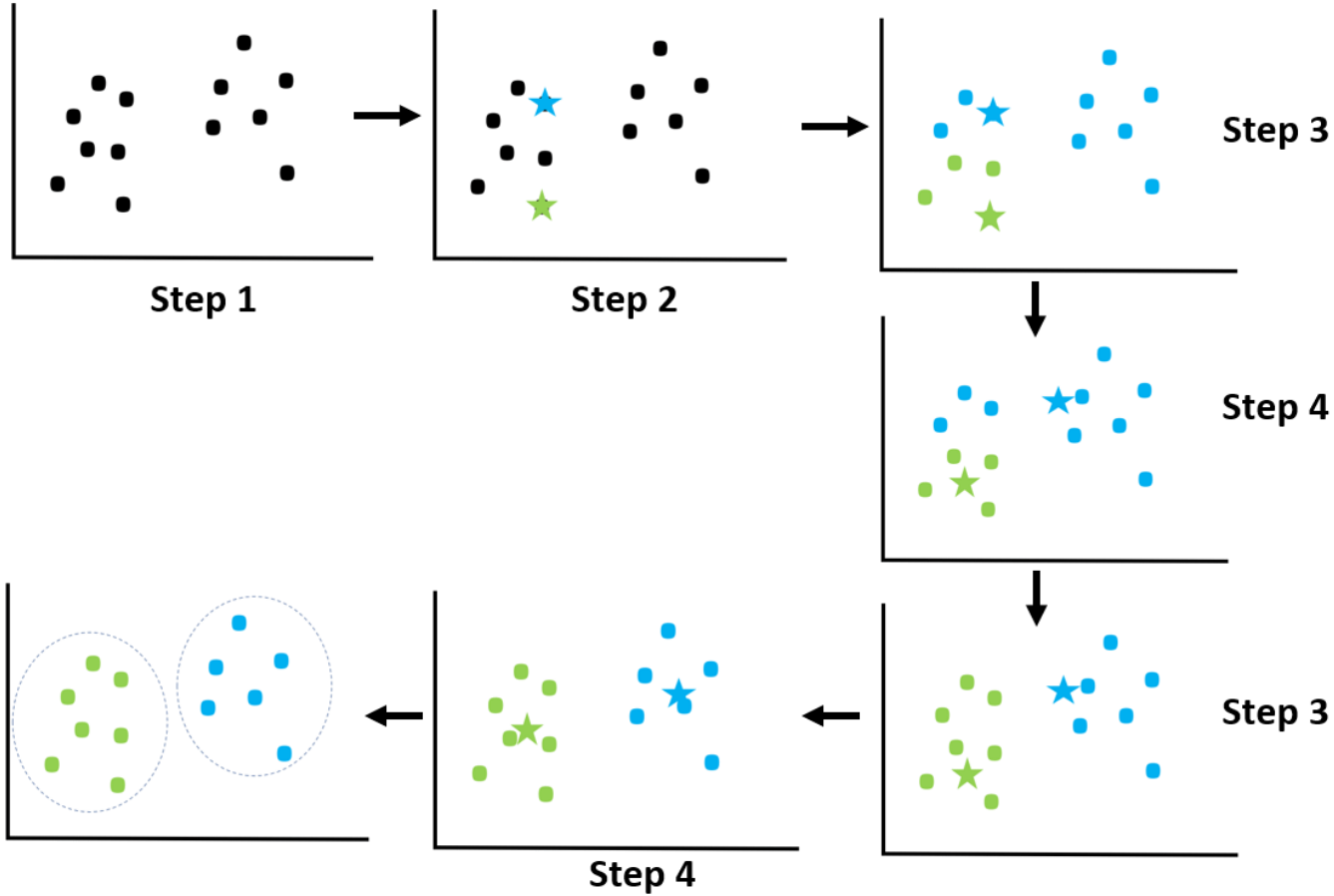
[1]Kmeans_algorithms

K-Means Clustering Algorithms

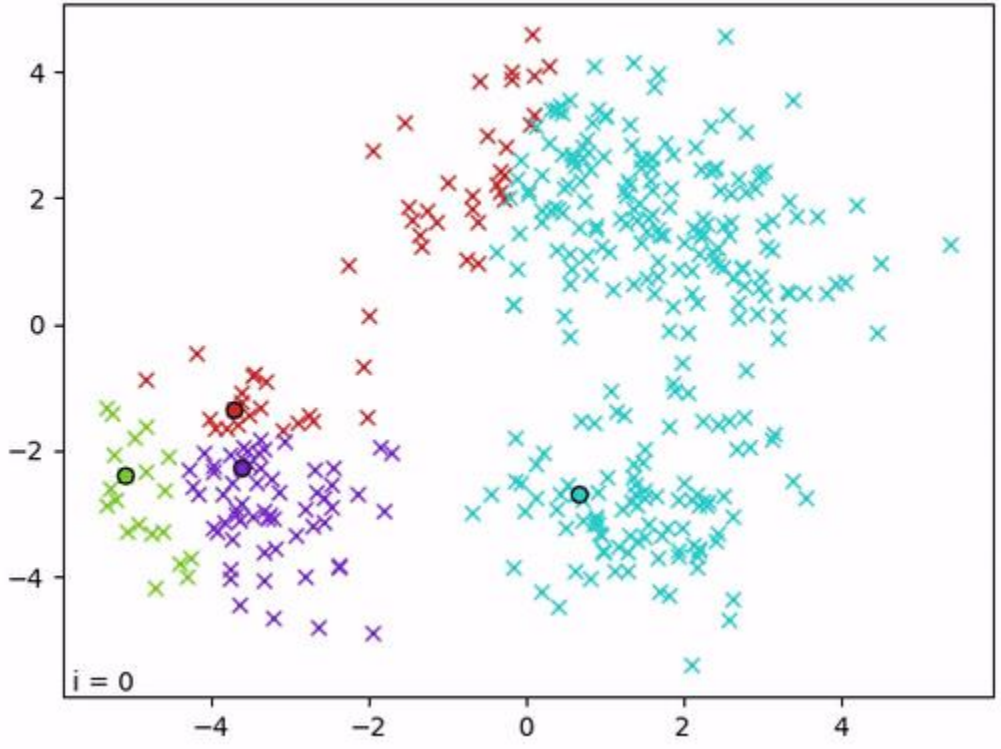
Step 5. Repeat Step 3 and Step 4 until cluster centroids do not change.



[1]Kmeans_algorithms



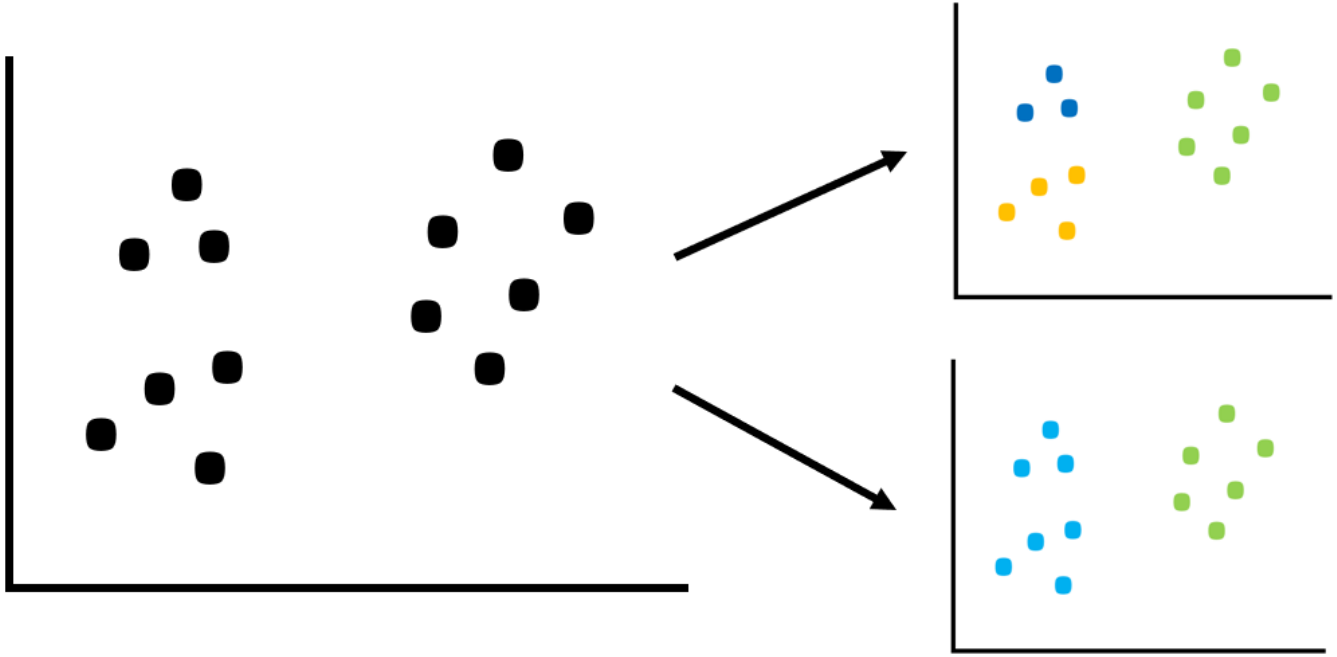
[1]Kmeans_algorithms



[2]kmeansAlgorithms

Problem

How to choose the right number of Clusters

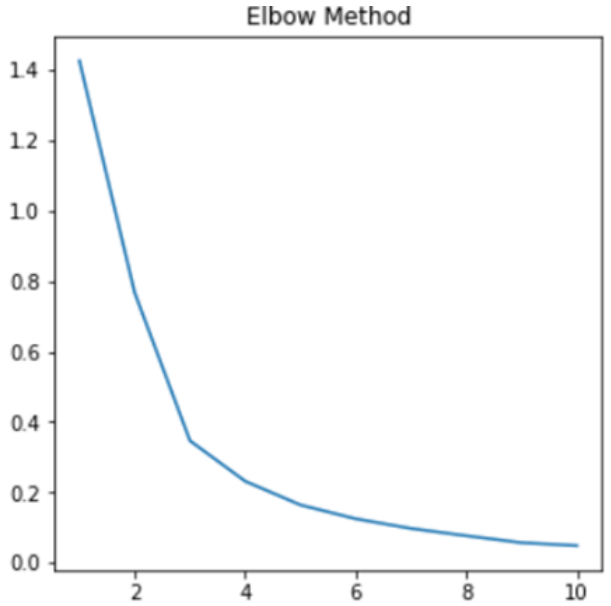


Elbow method

The sum of the squares of the distances of the points in the centroid of the cluster.

The sum is the minimum -> optimal K

But... ambiguous

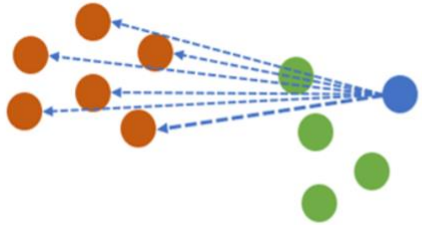


[6]optimalK

Silhouette method

a(i): avg distance between i and all other datapoints within cluster

a(i)



b(i): avg distance between i and all other datapoints outside/neighboring cluster

b(i)

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, j \neq i} d(i,j)$$

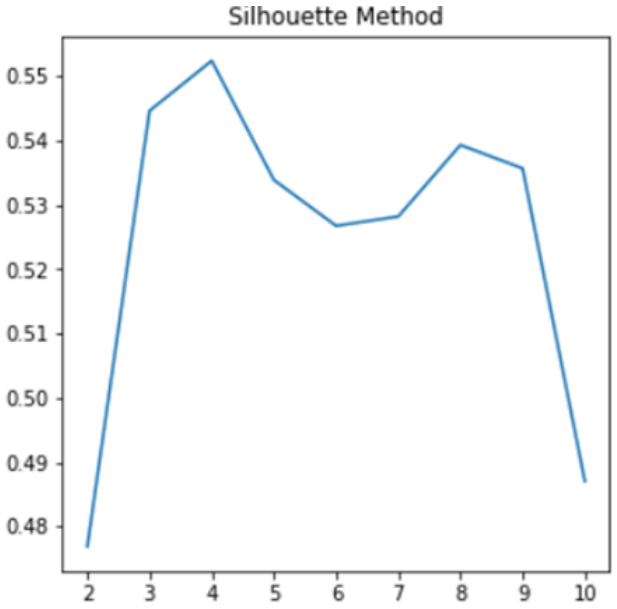
$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j)$$

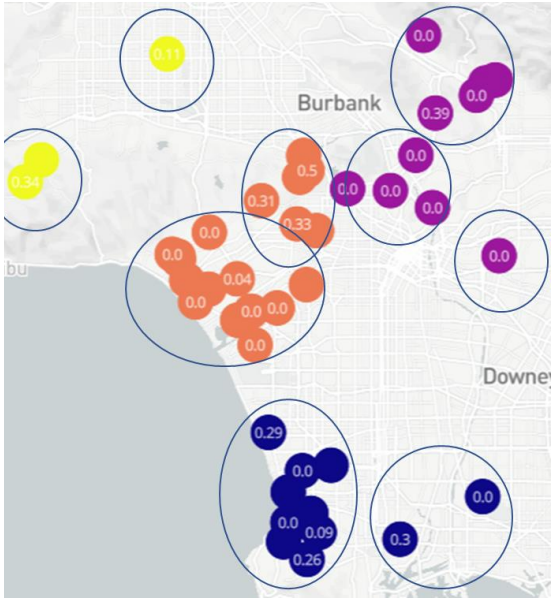
$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

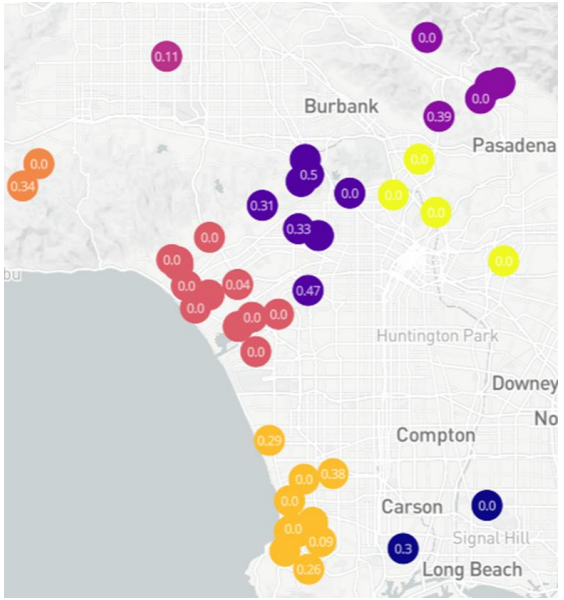
-1 <= S(I) <= 1

*CI = number of data in cluster
d(i,j) = distance from i to j

[6]optimalK

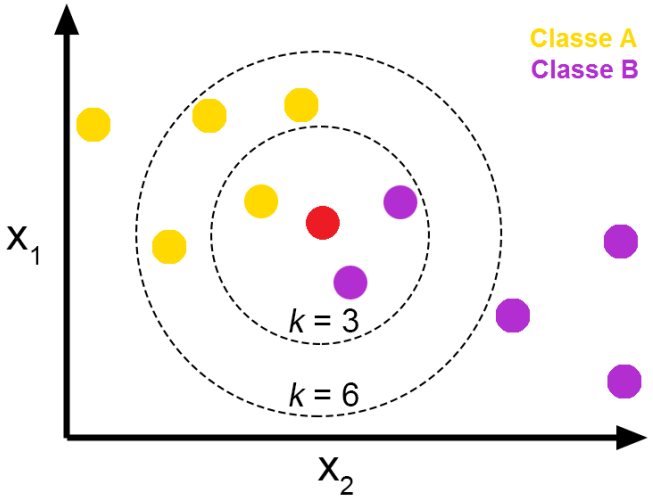

Silhouette Method
0.55
0.54
0.53
0.52
0.51
0.50
0.49
0.48
2
3
4
5
6
7
8
9
10


Burbank
Downey

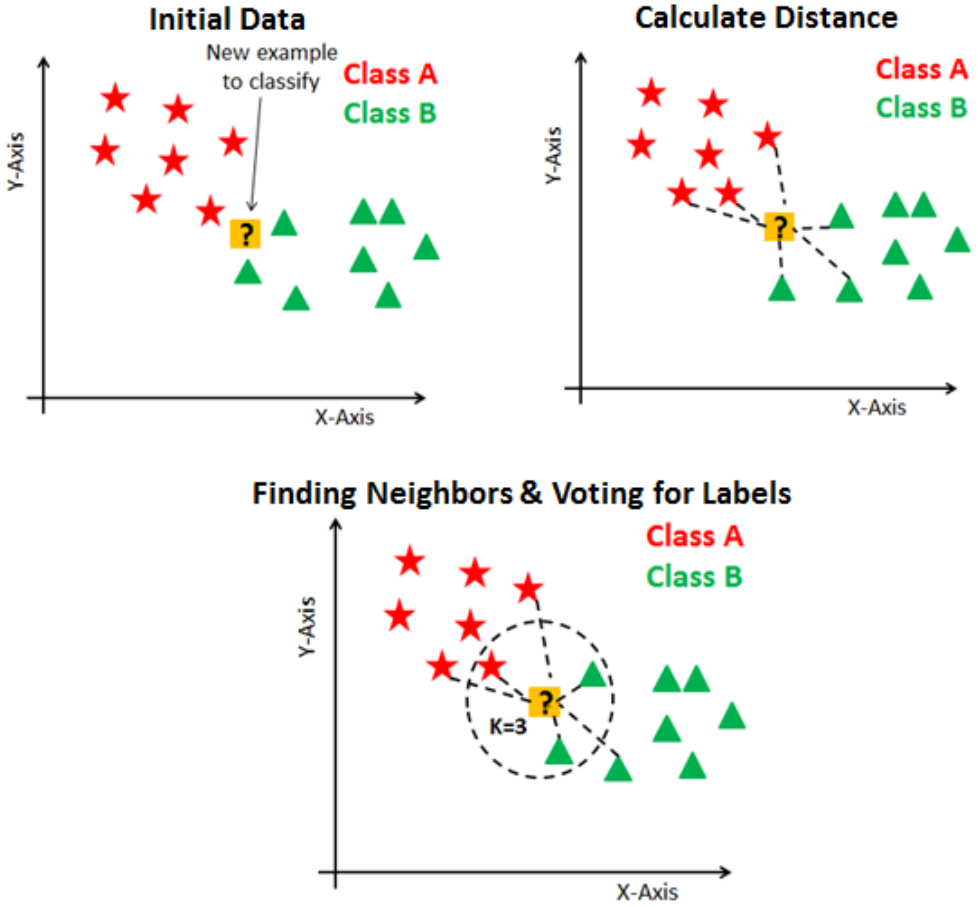

Burbank
Pasadena
Huntington Park
Downey
Compton
Carson
Signal Hill
Long Beach

[6]optimalK

What algorithms is **K-NearestNeighbor**?



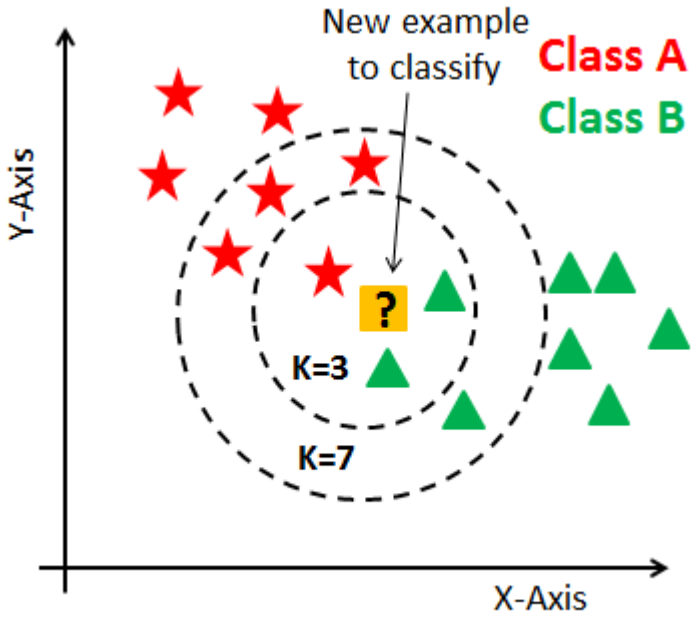
- Classification: Predict by majority vote
- Regression: Predict by mean value



[4]K-NN_algorithm

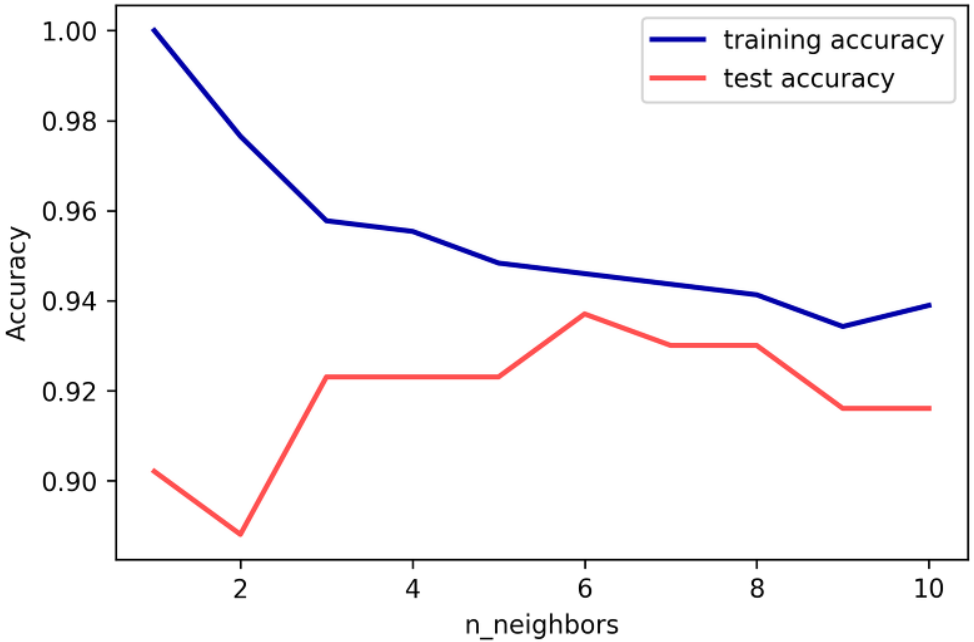
Problem

How to choose the optimal K



[4]K-NN_algorithm

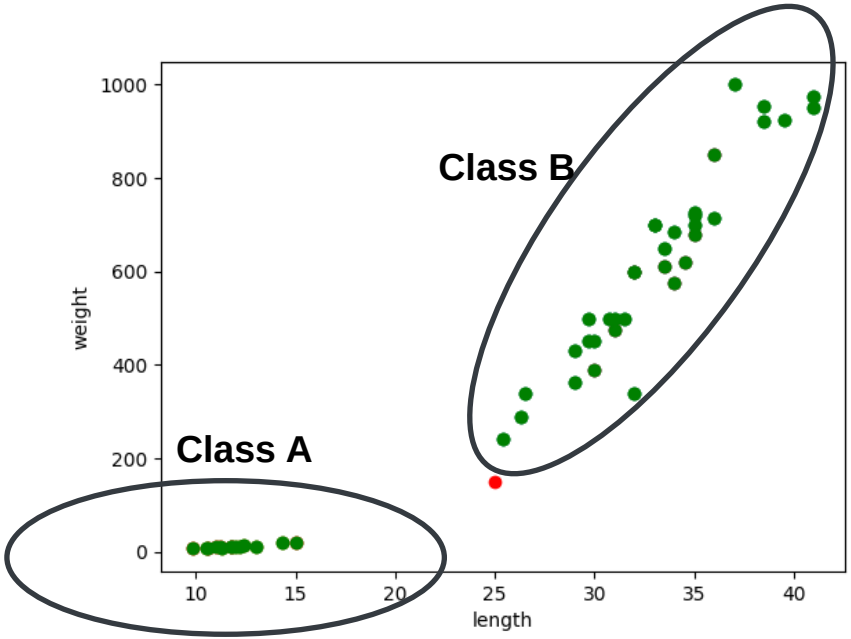
Find optimal K : Parameter tuning

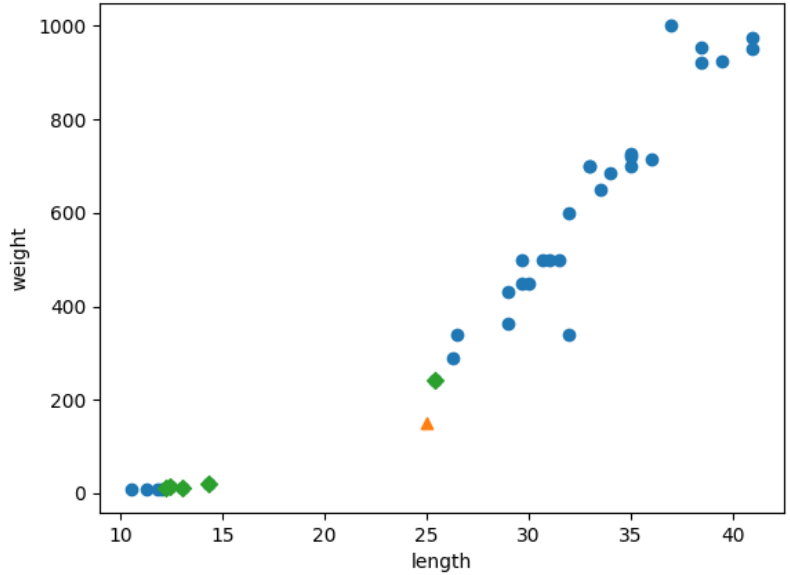


[4]K-NN_algorithm

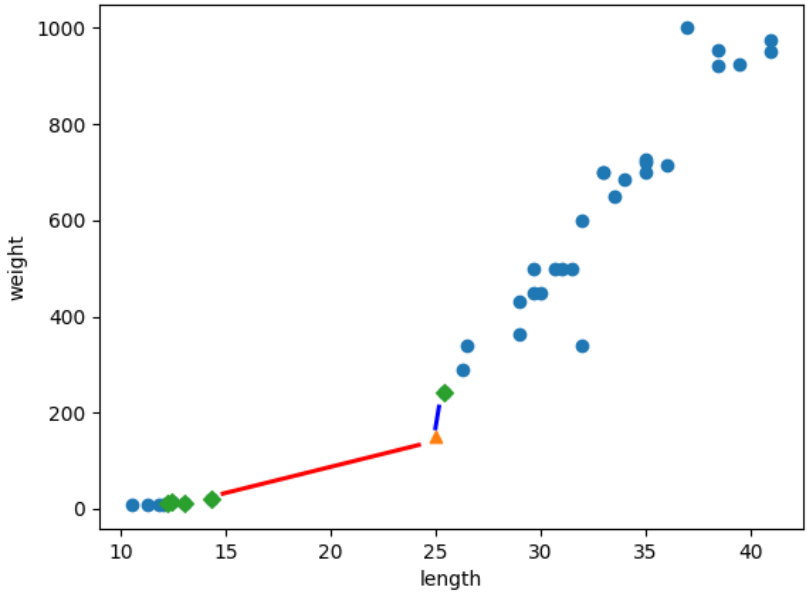
Problem2

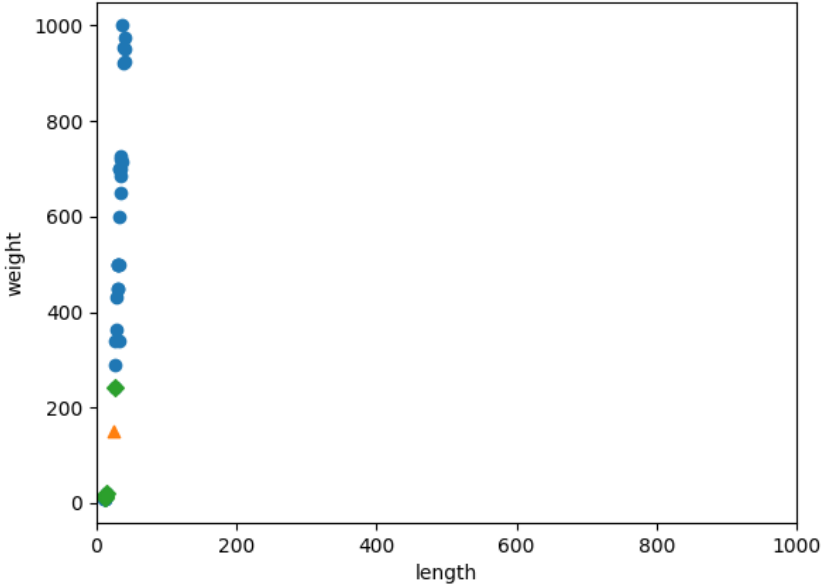
(K=5)



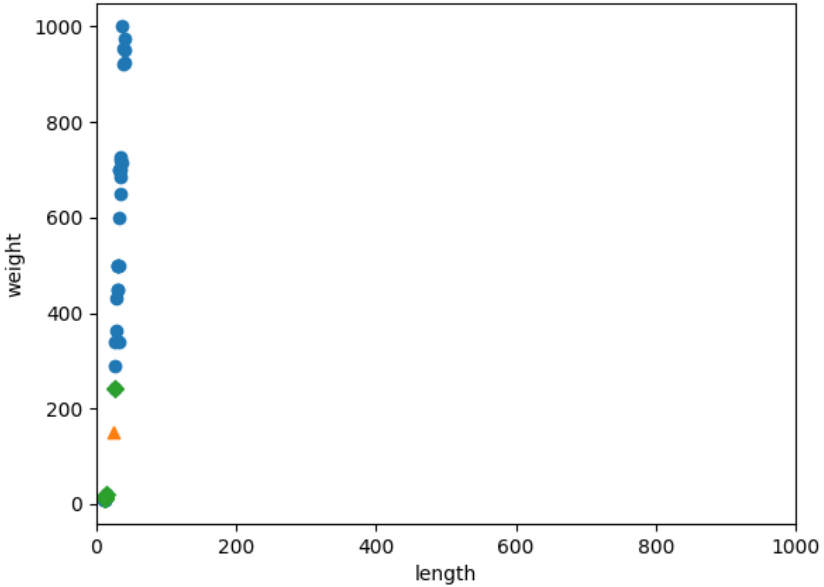


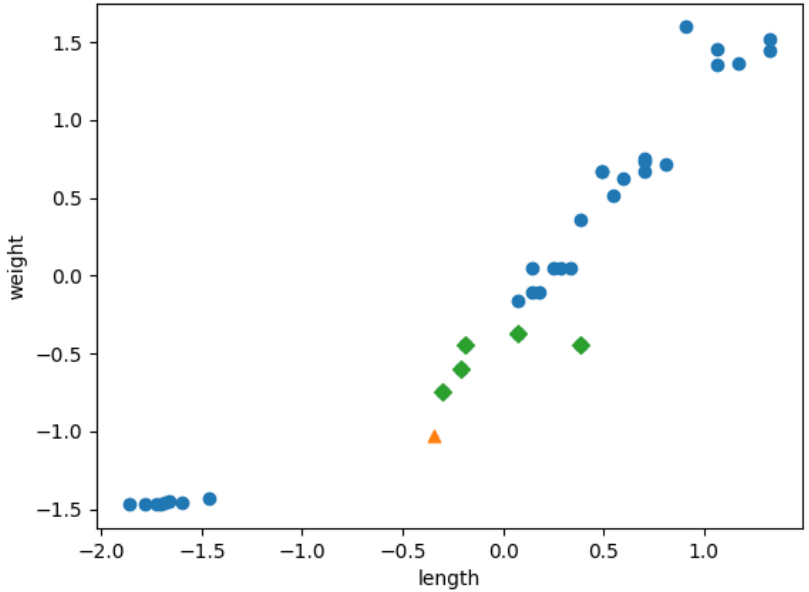
[4]K-NN_algorithm





[4]K-NN_algorithm





[4]K-NN_algorithm

Thanks!

Any questions?