K-Means Clustering

Lab1

K-Means Clustering

- Unsupervised learning algorithm
- The objective of K-means is simple: group similar data points together and discover underlying patterns.

How it works?

- First, we'll define *k*, which refers to the number of centroids you need in the dataset. We randomly initialize the centroids.
- Second, calculate the distance between every data point and centroids, using a distance metric. Allocate every data point to the nearest cluster.
- Third, update the centroid of every cluster.
- Fourth, iterate through steps 1-3 until the centroids have been stabilized.

Hyperparameters

- There are two hyperparameters:
- Value of K? (Elbow method to find optimal value of K.)

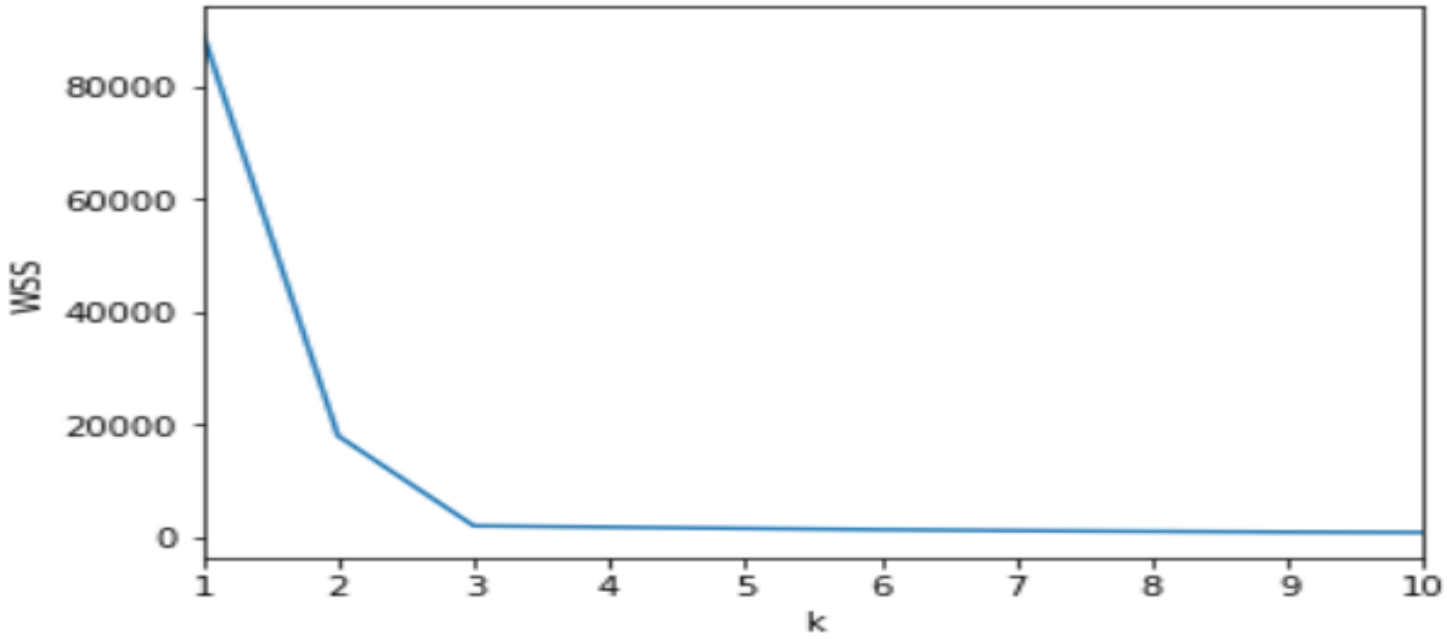
- Type of distance metric? (L1, L2 distances are good.)

- No obvious way of choosing them. Problem dependent.

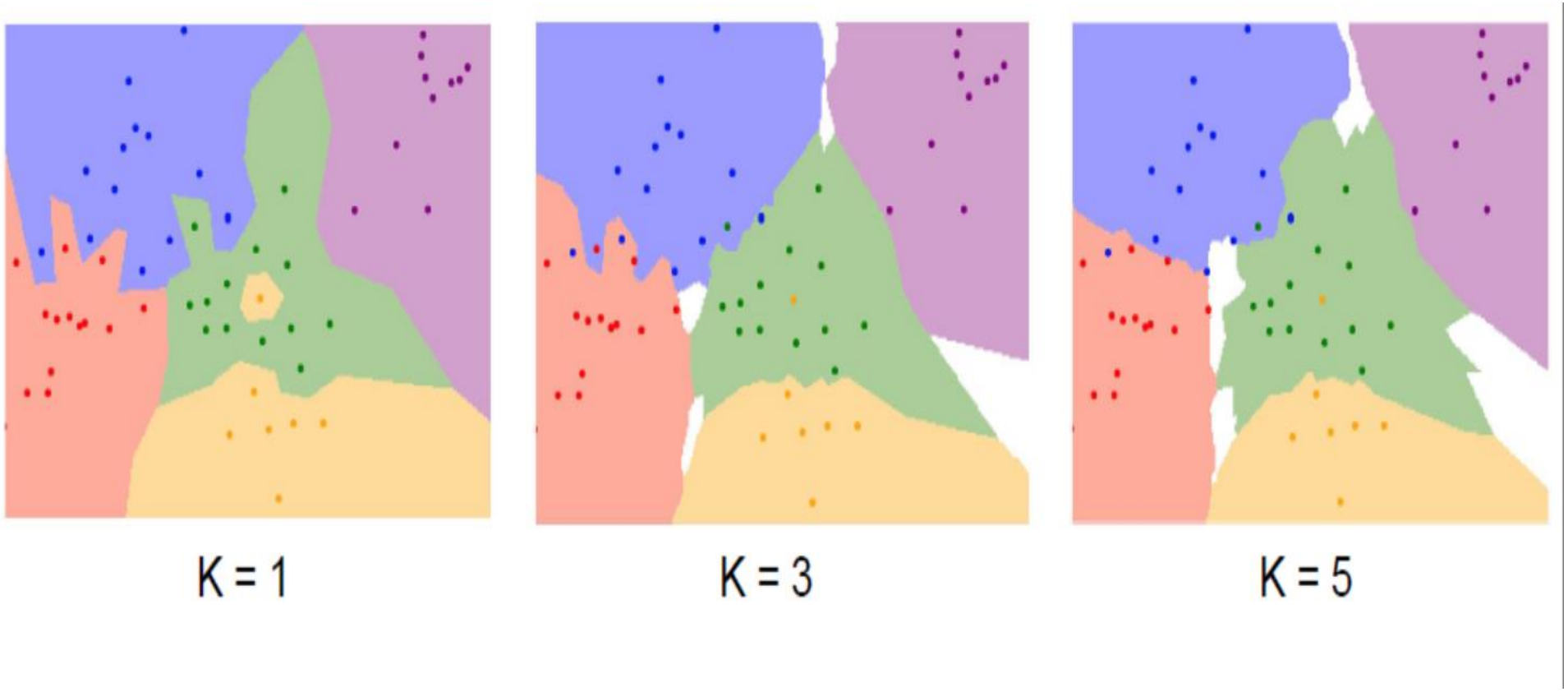
- Disadvantage: With N examples, Train O(1), Prediction O(N). Prediction is very slow.

K-Elbow Method

- Method to determine the optimal value of K or the number of clusters
- Calculate the **Within-Cluster-Sum of Squared** Errors (WSS) for **different values of k**, and choose the k for which WSS becomes first starts to diminish. In the plot of WSS-versus-k, this is visible as an **elbow.**
- **In the plot below elbow is at k=3**



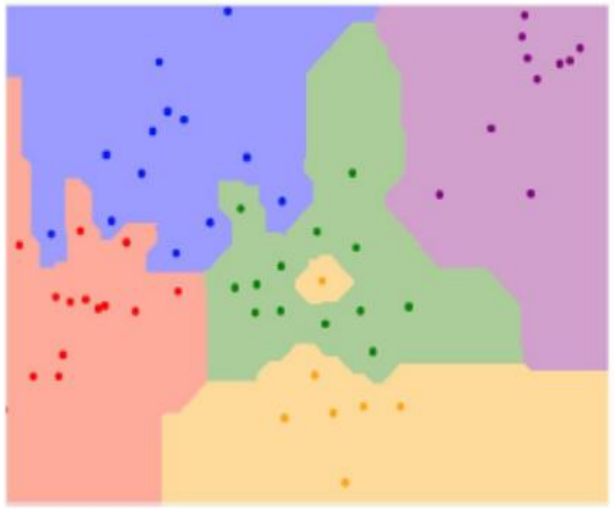
Different values of K



Distance Metric

L1 (Manhattan) distance

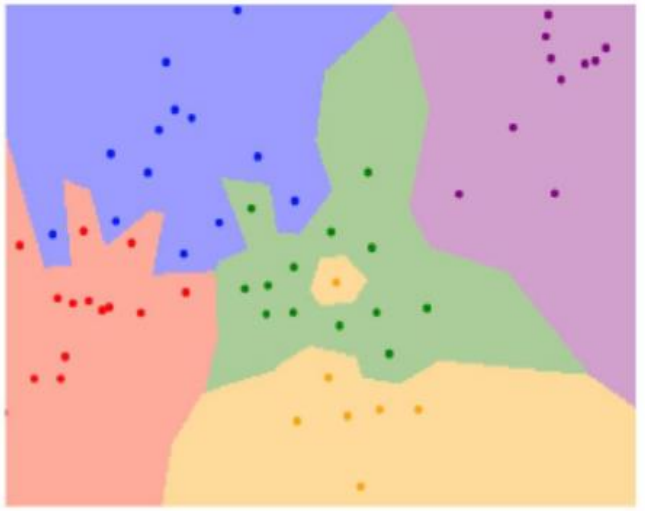
$$d_1(I_1, I_2) = \sum_p |I_1^p - I_2^p|$$



K = 1

L2 (Euclidean) distance

$$d_2(I_1, I_2) = \sqrt{\sum_p (I_1^p - I_2^p)^2}$$



K = 1

Demo

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/