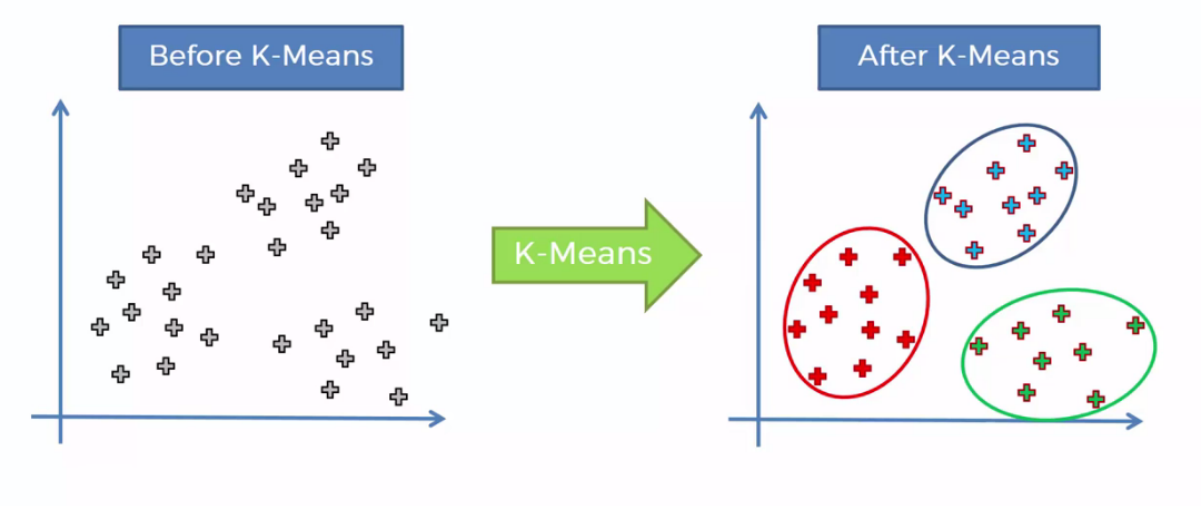
K-means clustering

K-means clustering is a popular unsupervised machine learning algorithm used for partitioning a dataset into K distinct, non-overlapping subsets (clusters). The goal is to assign each data point to a cluster in a way that minimizes the variance within each cluster.

- 비지도 머신러닝 알고리즘 중 하나.

- 각 클러스터 내의 분산이 최소화 되도록 각 데이터 포인트를 클러스터에 할당. (임의의 중심(Centroid)을 기준으로 최소 거리가 되도록 여러 그룹 또는 클러스터로 군집화)



1) K-means 알고리즘 수행 절차

1. 초기 설정: Cluster 개수(K) 결정 K = n

2. 중심 설정(Initialization of centroids): 다양한 방법으로 중심 설정 (임의로, 수동으로 등)

3. 데이터 할당(Assignment): 각 데이터 포인트를 가장 가까운 중심에 할당. 이때 거리 측정은 일반적으로 유클리드 거리 계산법을 사용

4. 중심 업데이트(Centroid Update): 각 클러스터의 중심을 해당 클러스터에 속한 모든 데이터 포인트의 평균으로 업데이트

5. 3~4 단계 반복: 클러스터 할당과 중심 업데이트가 큰 변화가 없을 때까지 계속 3, 4 단계를 반복

2) K-means clustering 강점

- In the analysis, K-means algorithm has shown higher computation speed, less clustering time and capability of handling dynamic data. (Hung, Phan Duy, Nguyen Thi Thuy Lien, and Nguyen Duc Ngoc. "Customer segmentation using hierarchical agglomerative clustering." Proceedings of the 2019 2nd International Conference on Information Science and Systems. 2019)

다양한 clustering 알고리즘(Hierarchical clustering, Density based clustering, Affinity Propagation clustering 등)과 비교하였을 때 빠른 연산 속도, 적은 클러스터링 시간 및 동적 데이터 처리 능력을 보여줌.

K-means clustering을 XAI에 사용하는 법

While there are many good clustering algorithms, the resulting cluster assignments can be hard to understand because the clusters may be determined using all the features of the data, and there may be no concise way to explain the inclusion of a particular point in a cluster. (Moshkovitz, Michal, et al. "Explainable k-means and k-medians clustering." International Conference on Machine Learning. PMLR, 2020)

- 머신러닝 모델의 대부분은 그 결과를 설명하지 않는 블랙박스(black box) 형태. 데이터 clustering이 필요한 모델에서 클러스터링 결과 최종적으로 데이터가 클러스터에 할당된 이유를 이해하기 어려운 문제가 발생함. (데이터의 모든 feature을 반영하여 클러스터를 결정하고, 클러스터에 특정한 포인트가 포함된 것을 간결하게 설명할 수 있는 방법이 없음)

1) 트리 기반 Clustering 설명

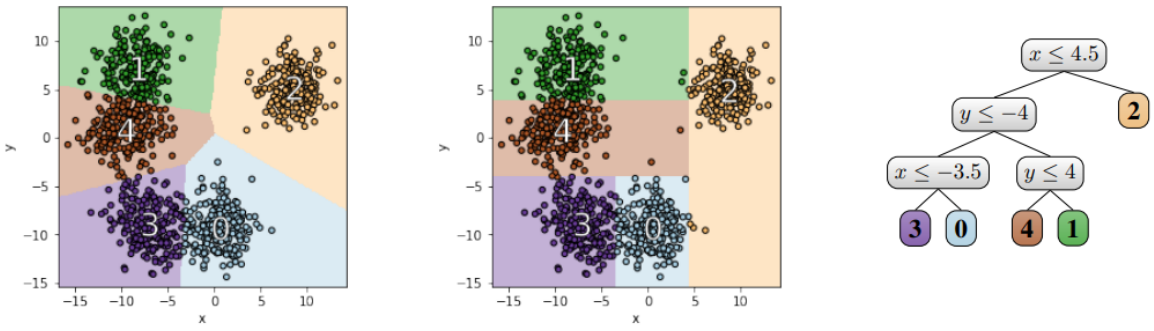
- K-means는 clustering에서 가장 자주 사용되는 방법 중 하나이지만, 단순히 결과값을 반환하므로 단독으로는 clustering 알고리즘을 쉽게 설명하지 못함.

- 각 포인트가 클러스터에 속하는 이유를 설명하기 위해서 Small decision trees를 사용할 수 있음. (Molnar, 2019; Murdoch et al., 2019)

- Threshold tree: K 개의 cluster를 찾는 것에 중점을 두므로 k 개의 잎(=cluster)을 가지는 트리 사용

Any cluster assignment is explained by computing the thresholds along the root-to-leaf path. By restricting to k leaves, we ensure that each such path accesses at most k-1 features, independent of the data dimension.

- 클러스터에 할당되는 이유를 root에서 leaf까지 경로 계산으로 설명 가능. K 개의 잎을 가지는 트리의 경우 최대 K-1 개의 Feature로 설명 가능.



(a) Optimal 5-means clusters (b) Tree based 5-means clusters (c) Threshold tree

Figure 1: The optimal 5-means clustering (left) determines uses combinations of both features. The explainable clustering (middle) uses axis-aligned rectangles summarized by the threshold tree (right). Because the clusters contain nearby points, a small threshold tree makes very few mistakes and leads to a good approximation.

To analyze clustering quality, we consider the k-means and k-medians objectives (MacQueen, 1967; Steinhaus, 1956). The goal is to efficiently determine a set of k centers that minimize either the squared L2 or the L1 distance, respectively, of the input vectors to their closest center.

- clustering 품질 분석에 K-means와 K-median objectives 고려하며, 가장 가까운 중심에 대한 입력 벡터의 제곱 L2 또는 L1 거리를 각각 최소화 하는 K centers의 집합을 효율적으로 결정하는 것을 목표로 함.(MacQueen, 1967; Steinhaus, 1956)

- Figure1에서 확인 가능하듯, 트리는 cluster를 결정 짓는 절단 축을 정의함. A와 B의 클러스터링 결과는 유사하지만 Tree를 사용한 B가 더 쉽게 설명할 수 있음.