K-means clustering

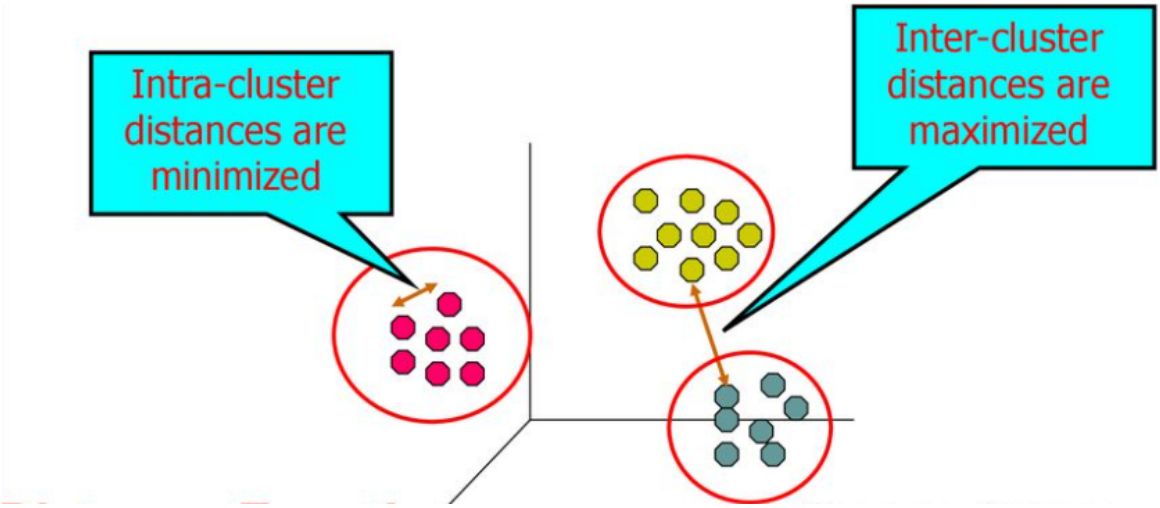
clustering?

clustering is a process of partitioning a set of data(or object) in a set of meaningful subclasses,called clusters

- cluster:a collection of data objects that are similar to one another and thus can be treated collectively as one group
- clustering: unsupervised classification:no predefined classes

Good Cluster?

- High intra-cluster similarity
- Low inter-cluster similarity

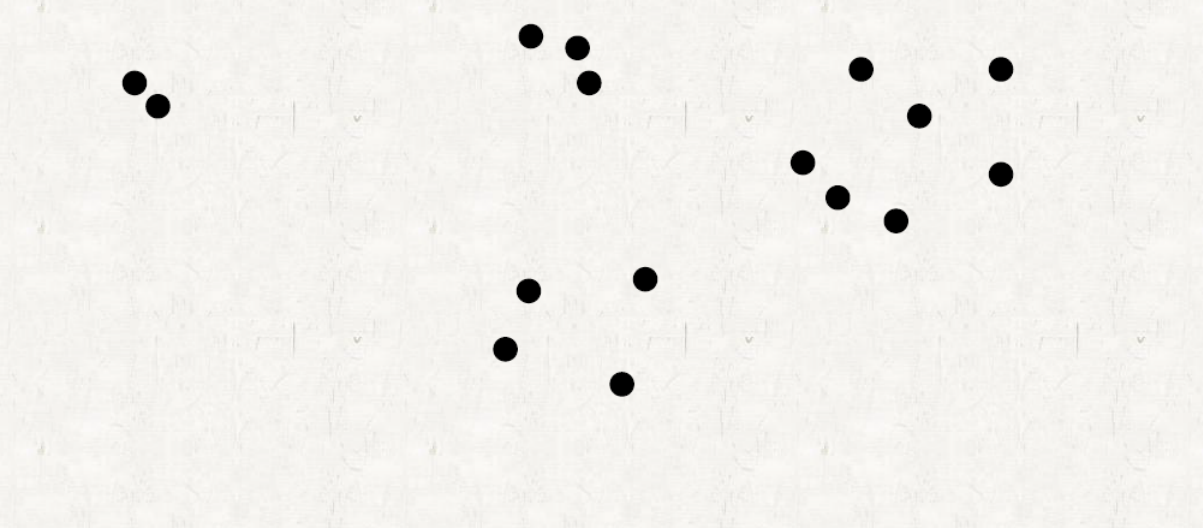


Clustering vs Classification

Clustering:Unsupervised Learning

- we do not know the class labels

- partition the data into classes

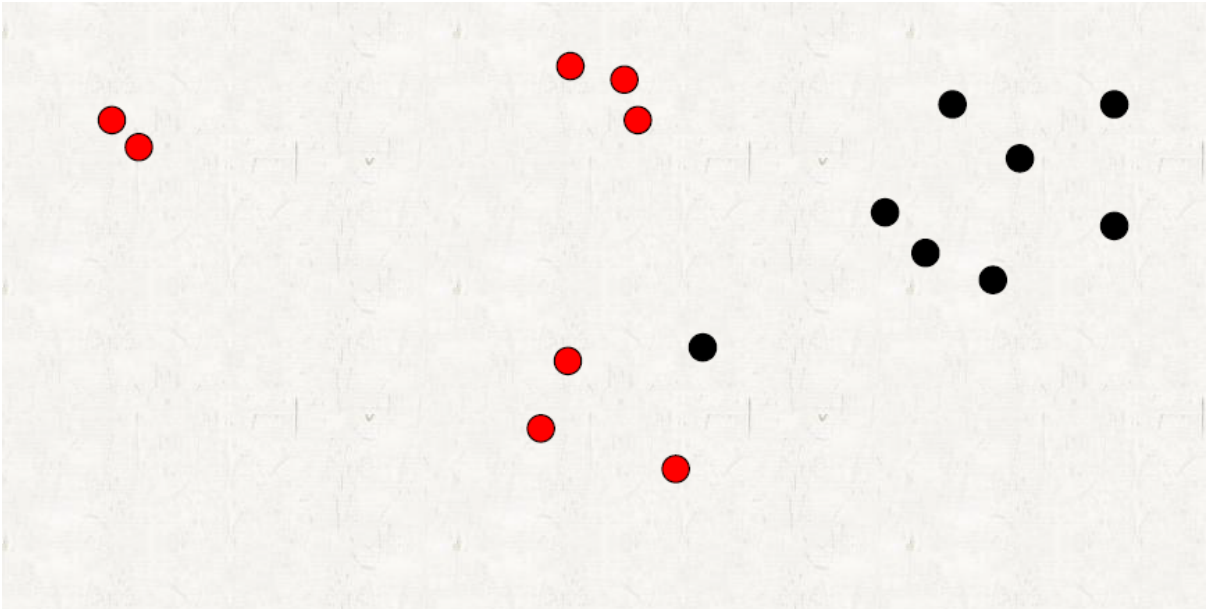


#Unsupervised Learning이란?

Unsupervised Learning은 supervised learning과 반대되는 개념으로, 맞춰야 하는 target value(label)가 없는 것을 말한다. 이를 맨 처음에 보면 맞추는 것이 없는데 뭘 learning하나.. 하는 생각을 할 수도 있다. 하지만 만약 100명의 사람들을 비슷한 사람들끼리 3개의 묶음으로 묶어야 한다고 하자. 우리는 각자 1-3으로 labeling된 사람들을 기준으로 labeling되지 않은 사람들을 묶는 것이 아니라, 아무런 label 없이 이들의 특성을 종합적으로 파악해서 묶어야할 것이다. 이런 경우와 같이 label이 없는 것에 대한 문제를 해결하는 것을 unsupervised learning이라고 한다.

Classification:Supervised Learning

- we know the class labels of each data
- find the boundary of each class



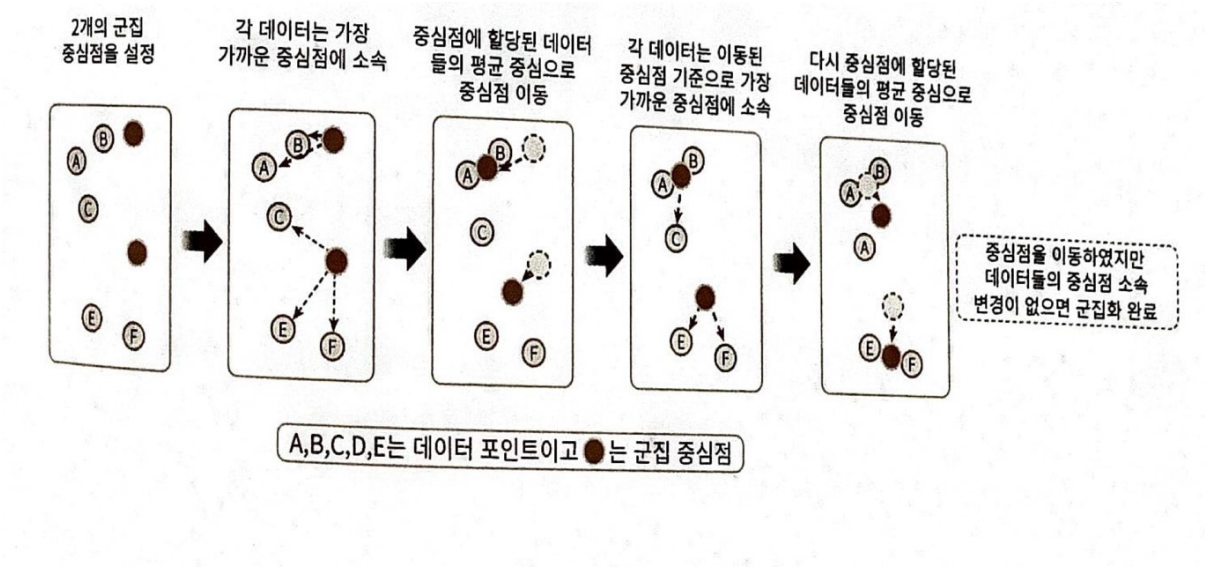
#Supervised Learning이란?

Supervised Learning은 한 줄로 요약하면 맞춰야 하는 값이 있는 것을 말한다. "어떤 학생이 대학원에 합격할지의 여부를 예측해라.", "저 사람이 결혼할지 평생 혼자 살 지를 예측해라"와 같이 예/아니오의 값을 예측해야하는 것부터 "지금의 집값이 7억인데 내년에는 집값이 얼마가 될 것인지 예측해라."와 같이 정확한 값을 예측하는 것까지 모두 supervised leanring의 범주에 속한다. 이를 우리는 target value(label)가 있는 것이라고도 말할 수 있다.

K-means

k-평균은 군집화(clustering)에서 가장 일반적으로 사용되는 알고리즘이다. k-means은 군집 중심점(centroid)이라는 특정한 임의의 지점을 선택해 해당 중심에 가장 가까운 포인트들을 선택하는 군집화 기법이다.

centroid는 선택된 포인트의 평균 지점으로 이동하고 이동된 중심점에서 다시 가까운 포인트를 선택,다시 중심점을 평균 지점으로 이동하는 프로세스를 반복적으로 수행한다. 모든 데이터 포인트에서 더이상 중심점의 이동이 없을 경우에 반복을 멈추고 해당 중심점에 속하는 데이터 포인트들을 군집화하는 기법이다.



Given K,the K-means algorithm is implemented in 4 steps:

1.Randomly choose seed points(centroids)

2.Assign each object to the nearest seed point.

3.Compute the centroids(mean point) of the current clusters.

4.Go back to Step 2,stop when no more seeds change

K-means Clustering 클러스터 수 계산(최적의 k찾기)

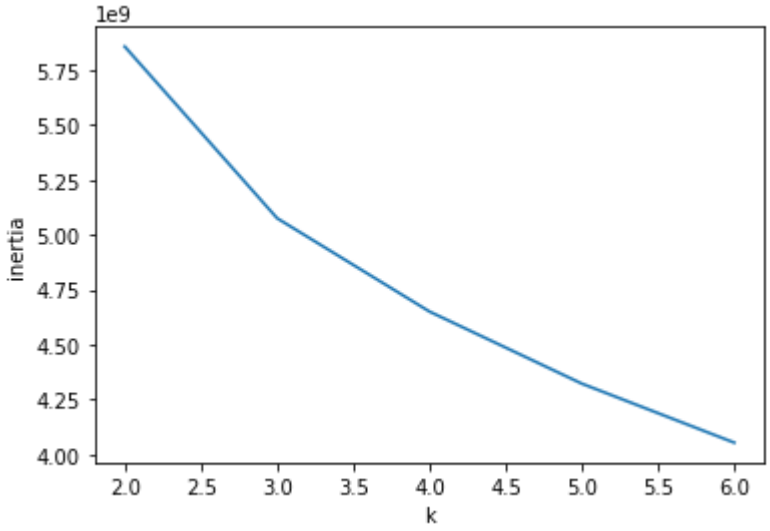
- Elbow Method
- Silhouette Method

1.Elbow Method

K-means 알고리즘은 centroid와 클러스터에 속한 샘플 사이의 거리를 잴 수 있다. 이 거리의 제곱합을 inertia라 한다. 이너셔는 클러스터에 속한 샘플이 얼마나 가깝게 모여 있는지를 나타내는 값으로 생각할 수 있다.

일반적으로 클러스터 개수가 늘어나면 클러스터의 개수가 늘어나므로 이너셔도 줄어든다.

엘보우 방법은 클러스터 개수를 늘려가면서 이너셔의 변화를 관찰하여 최적의 클러스터 개수를 찾는 방법이다.



클러스터 개수를 증가시키면서 이너셔를 그래프로 그리면 감소하는 속도가 꺾이는 지점이 있다. 이 지점부터는 클러스터 개수를 늘려도 클러스터에 잘 밀집된 정도가 개선되지 않는다.

즉 이너셔가 크게 줄어드지 않는다. 이 지점이 마치 팔꿈치 모양이여서 Elbow Method 라고 부른다.

K-means Strength

- Simple:easy to understand and to implement
- Efficient:Time complexity(O(tkn))—>O(n)

  n:of objects

  k:of clusters

  t:of iterations

K-means Weakness

- Clusters with non-spherical shapes(ellipsoidal shape)

  →해결방법:Mahalanobis distance를 사용한다

Mahalanobis distance는 유클리디언 거리에서 데이터의 속성들의 공분산 (covariance)을 반영하여 거리를 계산하는 방법이다. 계산값이 0에 가까울수록 유사한 것이다.

- Clusters with different sizes and densities
- Converge to a local minimum(sensitive to initial points)

  —>초기값 설정에 따라 다르다.
- Need to specift k in a advance
- Unable to handle noisy data and outliers

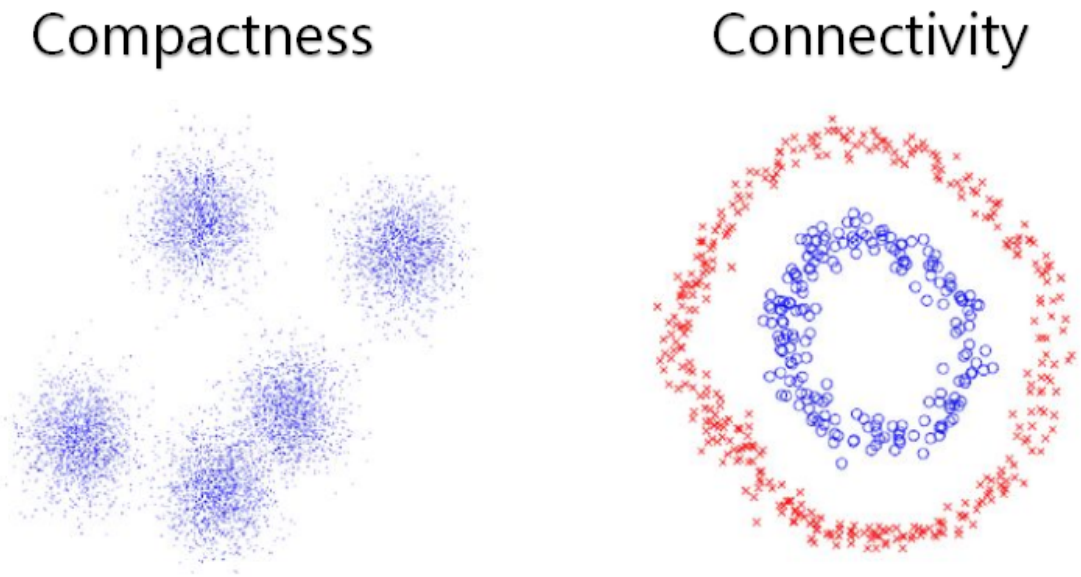
- Not suitable to discover clusters with non-convex shapes

—>해결방법:Spectral Clustering

Spectral Clustering.

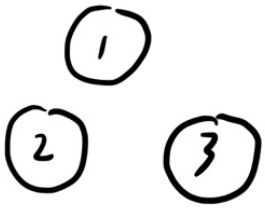
클러스터링의 종류

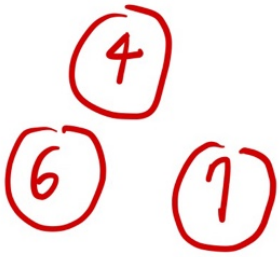
1. **Compactness:**서로 가까이 있는 대상은 같은 그룹으로 묶이고, 그 그룹의 핵을 중심으로 같은 그룹에 포함된 대상들이 밀집되어 분포하도록 하는 방식이다. 주로 두 대상간의 거리 (유클리드 거리 등) 가 대상간의 유사도를 측정하는 척도로 사용됩니다. K-Means 알고리즘이 이 분류의 알고리즘에 속한다.
2. **Connectivity:**서로 연결되어 있거나 바로 옆에 있는 대상이 같은 그룹으로 묶인다. 두 대상의 거리가 매우 가깝더라도 대상이 연결되어 있지 않다면 같은 그룹으로 묶이지 않는다. Spectral Clustering 알고리즘이 이 방법을 사용한 클러스터링 알고리즘에 해당한다.

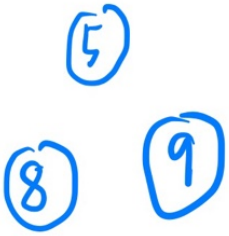


출처:https://medium.com/mathpresso/spectral-clustering-bc16ba602fb3

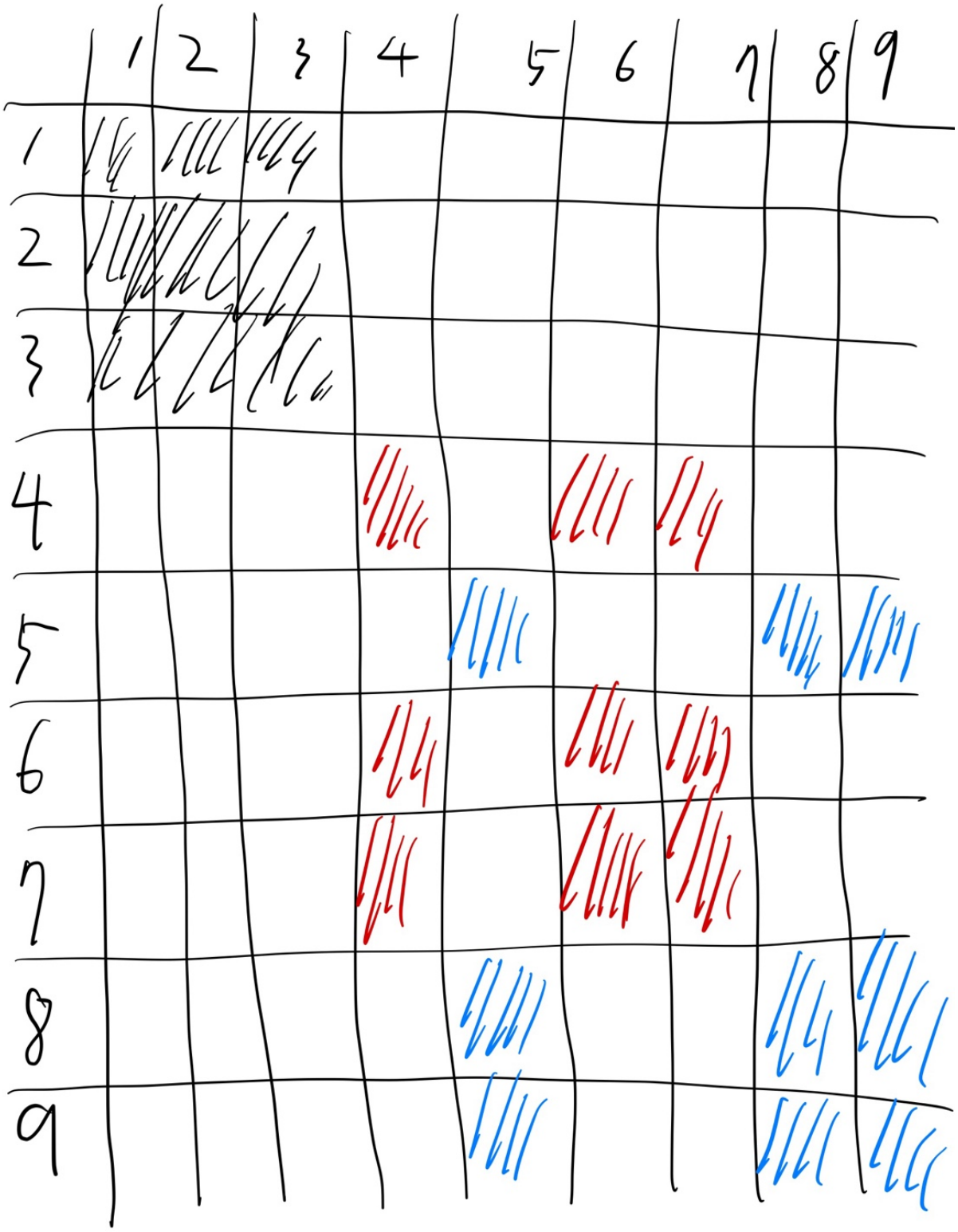
우선 데이터간의 거리에 기반한 유사도 행렬을 계산한다. 유사도 행렬을 계산한 이후 각 노드는 각 그룹을 분리하기 쉬운더 낮은 차원으로 이동. 사영된 낮은 차원의 공간에서의 거리를 기반으로 K-means와 같은 알고리즘을 통하여 각 클러스터를 생성하고 클러스터링 알고리즘은 종료된다.





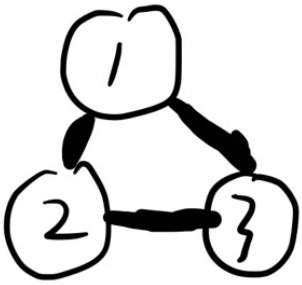


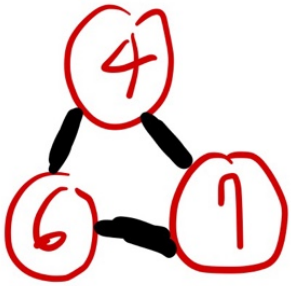
이제 spectral Clustering 알고리즘을 사용하여 위그림의 데이터들을 Graph로 표현 -> 그래프로 나타내기 위해서는 유사도 행렬 (Affinity Matrix)을 통해 이를 표현

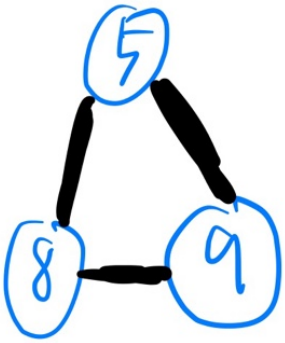

1 2 3 4 5 6 7 8 9
1
2
3
4
5
6
7
8
9

이 과정에서 주성분을 나타내는 벡터가 서로 연관된 높은 요소들로 구성되었기 때문에 같은 연관된 높은 점들을 비슷한 위치로 이동

이제 전통적 유사도 행렬에서 처음 데이터를 그래프로 나타내면.

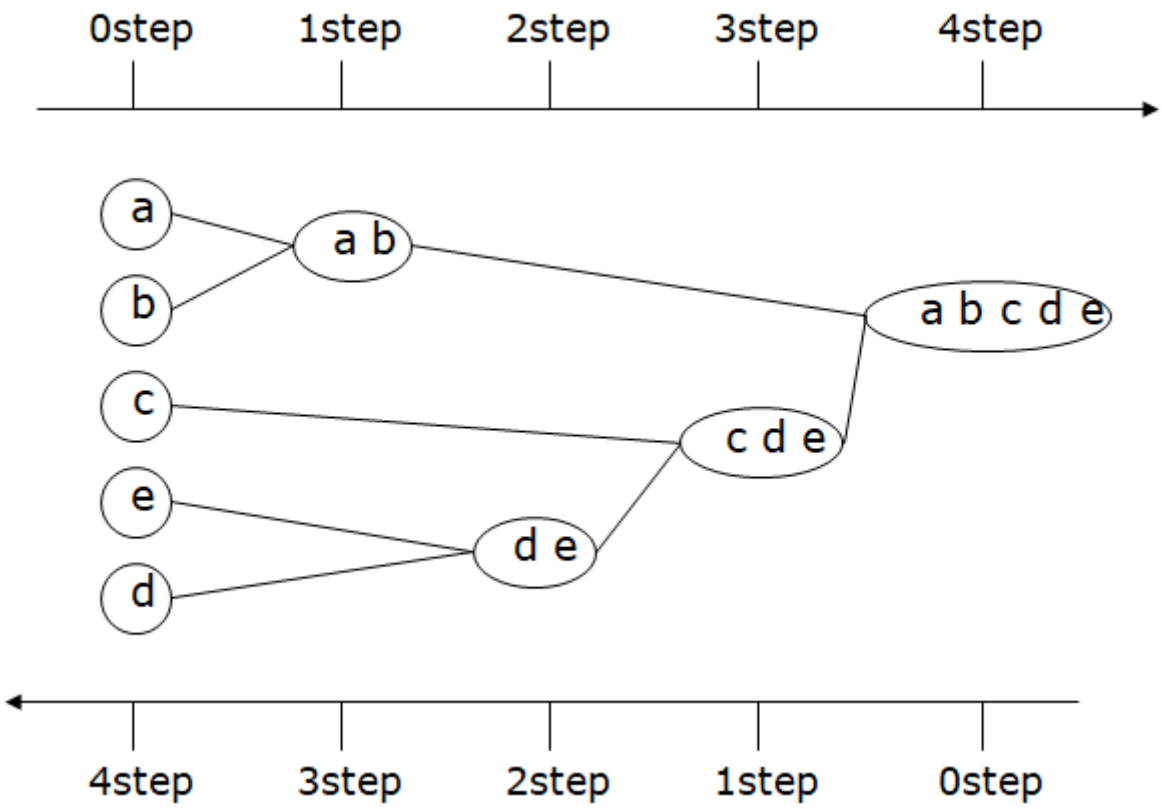






Hierarhical Method

Grouping objects into a tree of clusters



이 사진의 위쪽 방식은 Agglomerative Hierarchical Clustering아래쪽은 Disvisive Hierarchical Clustering 이다.

Agglomerative Hierarchical Clustering

이 알고리즘은 각 데이터가 모두 나눠져있는 상태에서, 작은 단위로부터 클러스터링을 시작하여 모든 데이터를 묶을 때까지 반복하는 Bottom Up 방식으로 클러스터링을 진행한다.

Similarity는 inter-cluster간의 유사성이다.

Similarity

- Single-Linkage

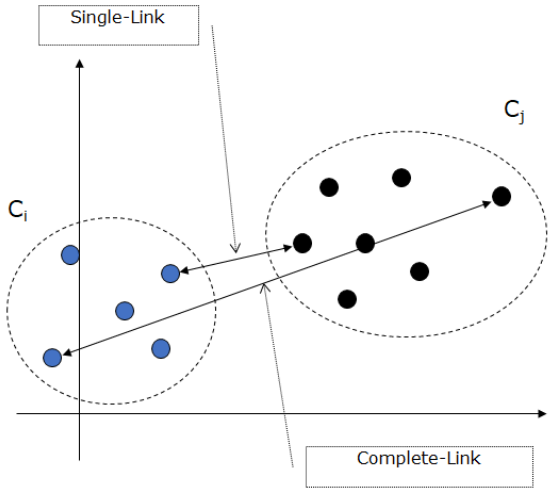
$$d_{\min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} |p - p'|$$

- Complete-Linkage

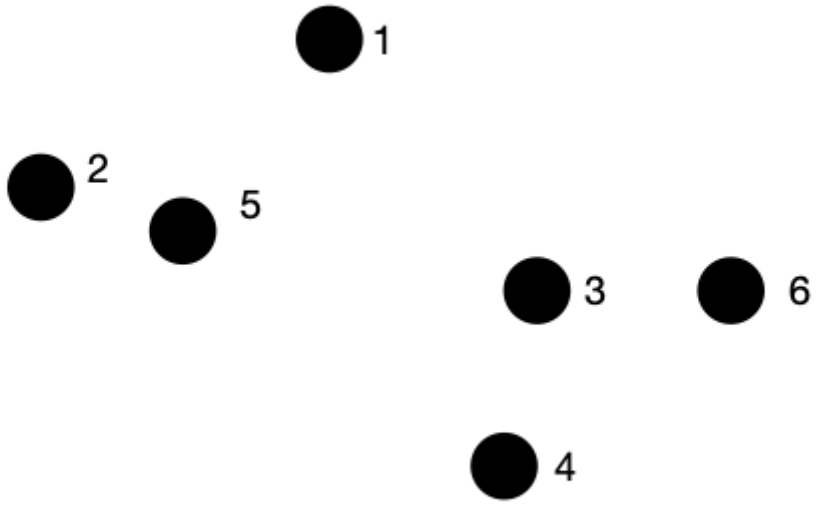
$$d_{\max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} |p - p'|$$

- Average-Linkage

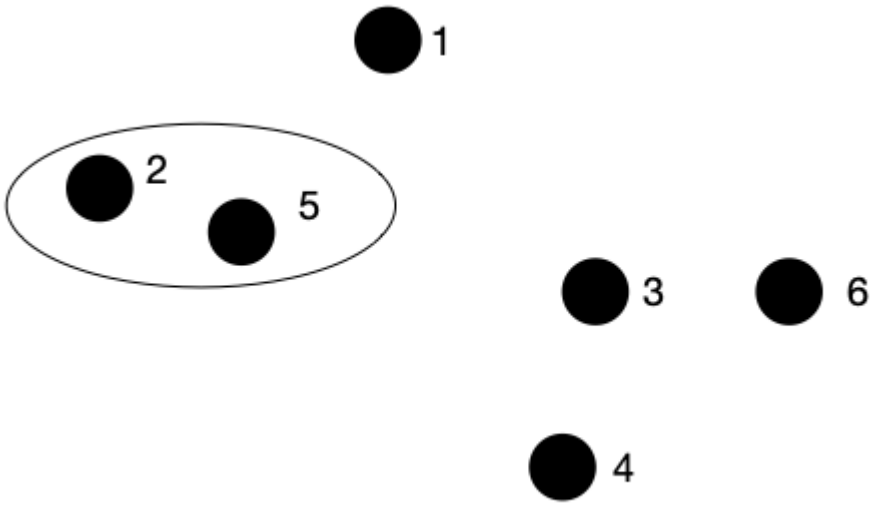
$$d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} |p - p'|$$



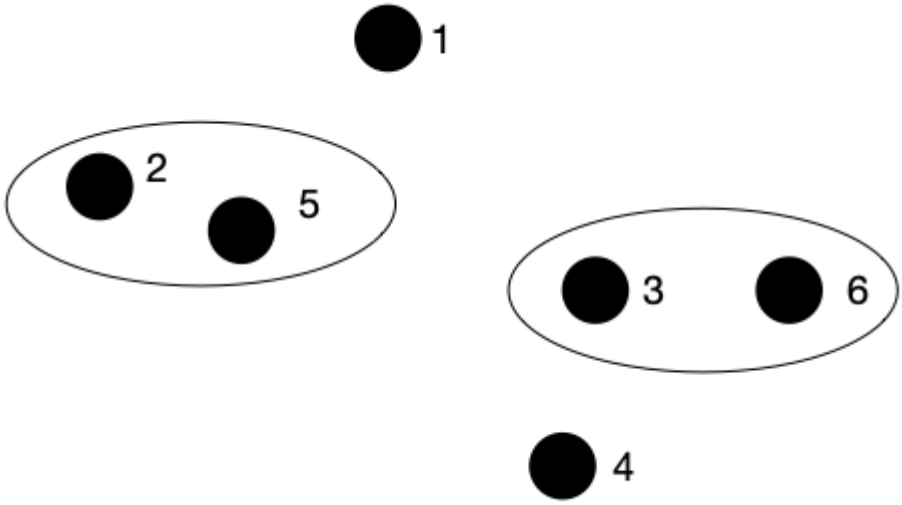
우선,single-Linkage로 설명한다.

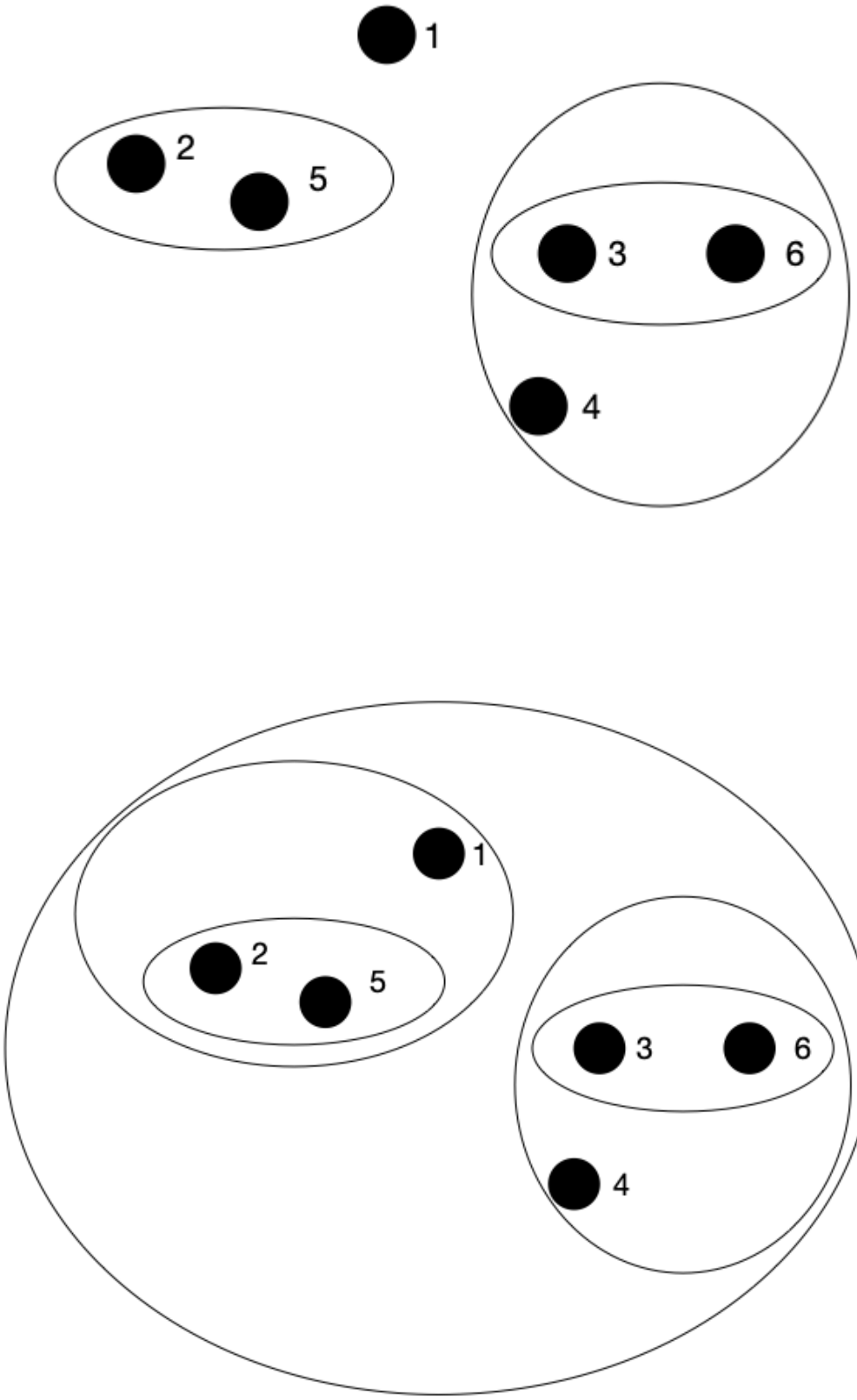


1. 우선 가장 거리가 가까운 데이터를 찾는다.



2.single-Linkage방식으로 하면




1
2
5
3
6
4
1
2
5
3
6
4

이 종료된 클러스터를 덴드로그램으로 표현하면 아래그림과 같다. 이 때, 각 덴드로그램의 수평선 높이는 클러스터가 만들어진 순서대로 정해진다.