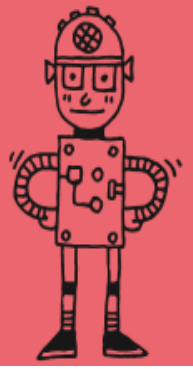
k-Means Clustering

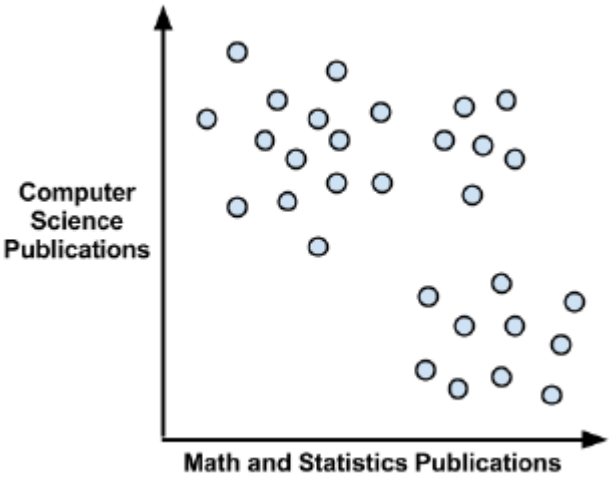
데이터사이언스 학과
김 승 환
swkim4610@inha.ac.kr

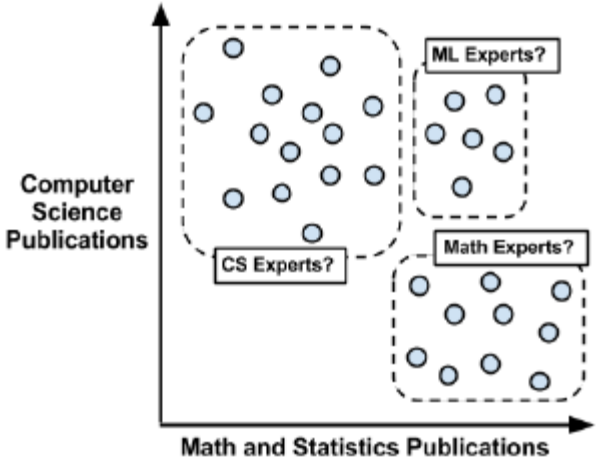


1 k-Means Clustering

군집화는 유사한 개체끼리 묶어주는 방법론이다.

군집화를 활용하여 마케팅에서는 유사한 고객끼리 묶어 세그먼트를 나누기도 하고, 복잡한 데이터를 몇 개의 범주로 간단하게 만들기도 한다. 아래의 그림은 대학교수들이 발표한 두 분야의 논문 수에 대한 도표이다.



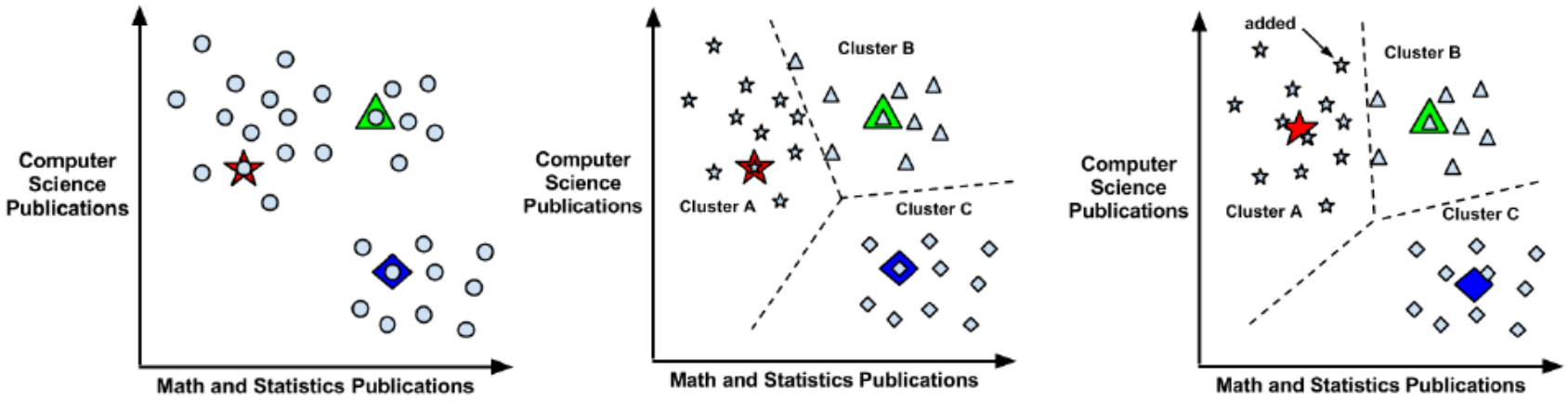


위 그림으로 우리는 각 군집의 성격을 유추하는 것이 가능하다.

이러한 유추를 Un-Supervised Learning이라고 한다.

1 k-Means Clustering

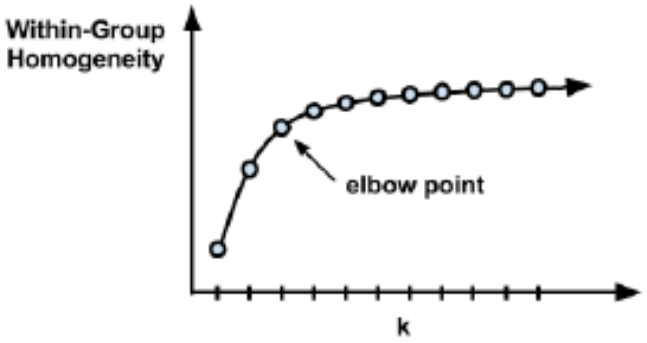
K-Means Algorithm은 원하는 군집의 수를 k개할 때, k개의 개체를 무작위로 선정하고 이를 초기 군집의 중앙값으로 선정한다. 이후, 군집중앙값과 각 개체의 거리를 계산하여 가까운 군집에 개체를 편입시킨다. 이후, 군집에 속한 개체의 중심을 구하고 그 값으로 군집중앙값을 대체한다. 이렇게 조정된 군집 중앙값을 이용하여 각 개체와의 거리를 재계산하고 가까운 군집으로 개체를 편입을 조정하는 방식이다. 이 방식으로 반복하여 더 이상 개체들이 속한 군집에 변화가 없으면 종료한다.

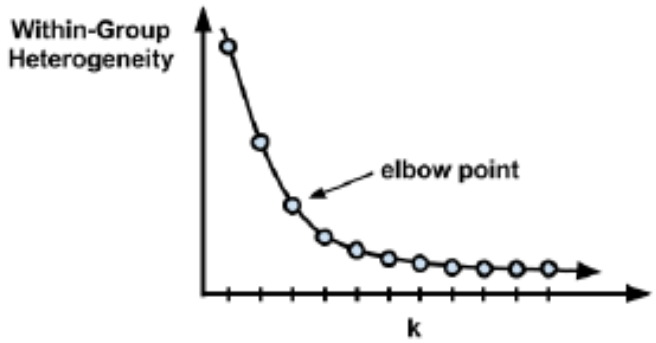


1 k-Means Clustering

Choosing the appropriate number of clusters

- 군집의 수는 다루는 문제의 성격에 따라 정해질 수도 있고, 사전 지식이 없을 경우, Data에 의해 정해질수도 있다. Data로 부터 정하는 방법으로는 그룹내 동질성 혹은 이질성을 군집의 수에 따라 plotting해서 변곡점을 찾는 방식이다.
- 또 다른 방법은 해석의 용이성이다. 대 부분 이 방법으로 클러스터의 수를 결정한다.





Finding teen market segments using k-means clustering

For this analysis, we will be using a dataset representing a random sample of 30,000 U.S. high school students who had profiles on a well-known SNS(Facebook) in 2006.

```
> str(teens)
'data.frame': 30000 obs. of 40 variables:
$ gradyear : int 2006 2006 2006 2006 2006 2006 2006 2006 ...
$ gender : Factor w/ 2 levels "F","M": 2 1 2 1 NA 1 1 2 ...
$ age : num 19 18.8 18.3 18.9 19 ...
$ friends : int 7 0 69 0 10 142 72 17 52 39 ...
$ basketball : int 0 0 0 0 0 0 0 0 0 0 ...
```

As we had expected, the data include 30,000 teenagers with four variables indicating personal characteristics and 36 words indicating interests.

2 실습

- 아래는 5개 군집의 중심좌표를 출력한 결과다.
- 0번 군집은 농구, 미식축구, 수영, 귀여움, 춤이 '+' 여서 'princess' 특성임을 알 수 있다.

	basketball	football	soccer	softball	volleyball	swimming	cheerleading	baseball
0	0.479167	0.470106	0.260635	0.359396	0.354098	0.253839	0.300328	0.331124
1	-0.160683	-0.162794	-0.086972	-0.115008	-0.114186	-0.093726	-0.110300	-0.109202
2	0.316594	0.334808	0.135711	0.188930	0.069937	0.231736	0.152551	0.267276
3	-0.339568	2.405319	-0.245497	-0.224583	-0.222700	1.645505	-0.208842	-0.205130
4	0.157756	0.246380	0.119035	0.043126	0.202017	0.212340	0.384854	0.006876

tennis	sports	cute	sex	sexy	hot	kissed	dance	band
0.143672	0.302939	0.494658	-0.006274	0.160382	0.357588	-0.041958	0.460454	0.269493
-0.051219	-0.128409	-0.182380	-0.095977	-0.077310	-0.135631	-0.133950	-0.161602	-0.098446
0.111755	0.754101	0.458089	1.944732	0.510233	0.294418	2.924140	0.404065	0.501042
-0.170662	-0.304997	0.816976	18.336008	1.615412	-0.267340	-0.204292	0.479992	0.634129
0.100686	0.086994	0.402972	0.016947	0.136403	0.427680	0.040814	0.205088	-0.102406

2 실습

Cluster 1 (N = 3,376)	Cluster 2 (N = 601)	Cluster 3 (N = 1,036)	Cluster 4 (N = 3,279)	Cluster 5 (N = 21,708)
swimming cheerleading cute sexy hot dance dress hair mall hollister abercrombie shopping clothes	band marching music rock	sports sex sexy hot kissed dance music band die death drunk drugs	basketball football soccer softball volleyball baseball sports god church Jesus bible	???
Princesses	Brains	Criminals	Athletes	Basket Cases

- 1번 군집의 이름을 공주, 2번 군집은 music, 3번 군집은 노는아이들, 4번 군집은 운동선수로 이해할 수 있다.
- 이러한 개체들의 성격은 1:1 마케팅에 중요한 정보로 사용된다.

감사합니다