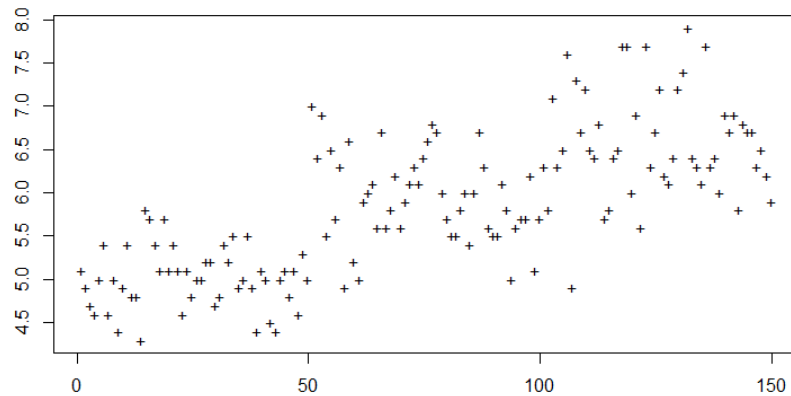


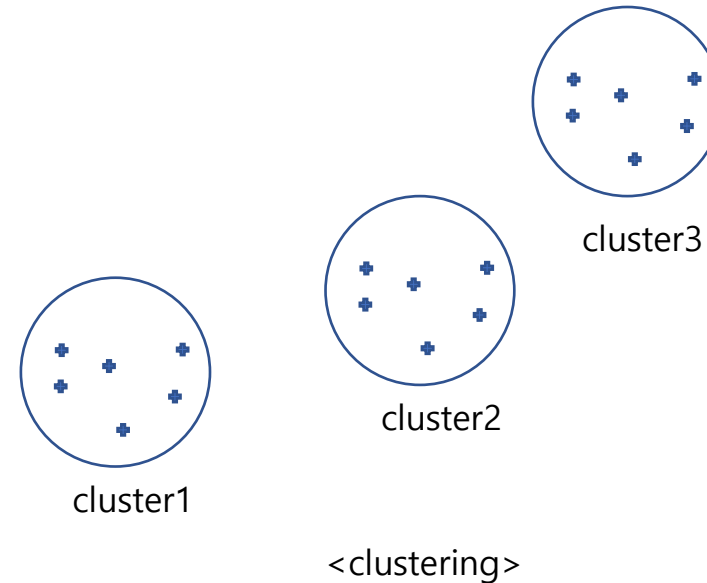
k-means data description

k-means data description

clustering



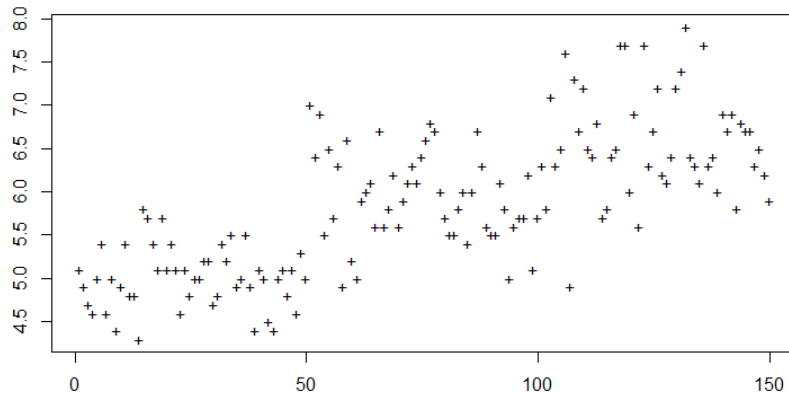
<Iris 데이터>



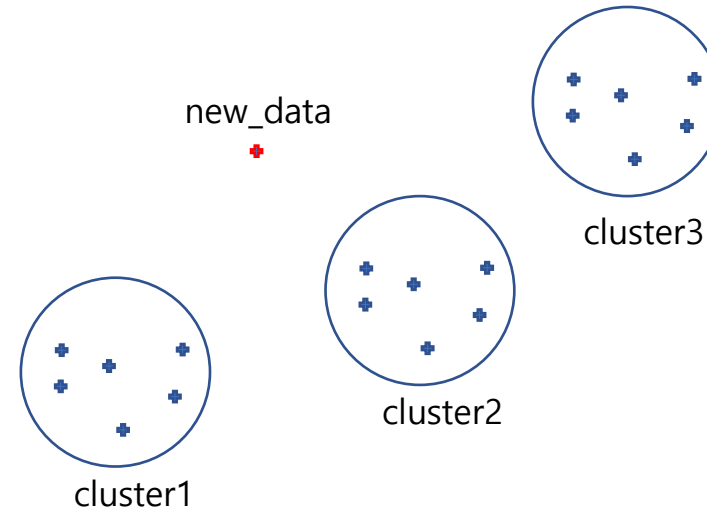
데이터를 clustering 한다는 것은 유사한 값들끼리 그룹화 시킨다는 것을 의미

k-means data description

clustering



<Iris 데이터>



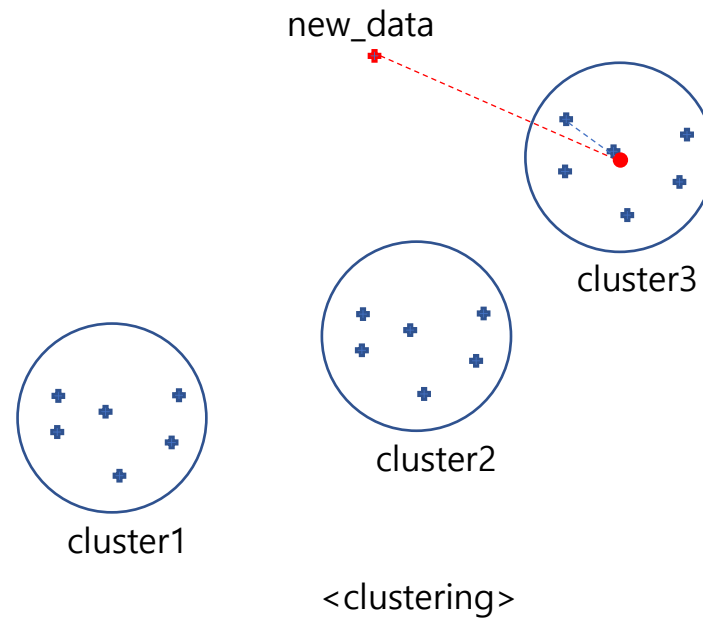
<clustering>

새로운 데이터가 들어 왔을 때 아무 cluster에 속해 있지 않다는 것은 유사한 데이터가 없다는 것을 의미
이 데이터는 왕따다 -> outlier를 의미한다.

k-means data description

거리 계산

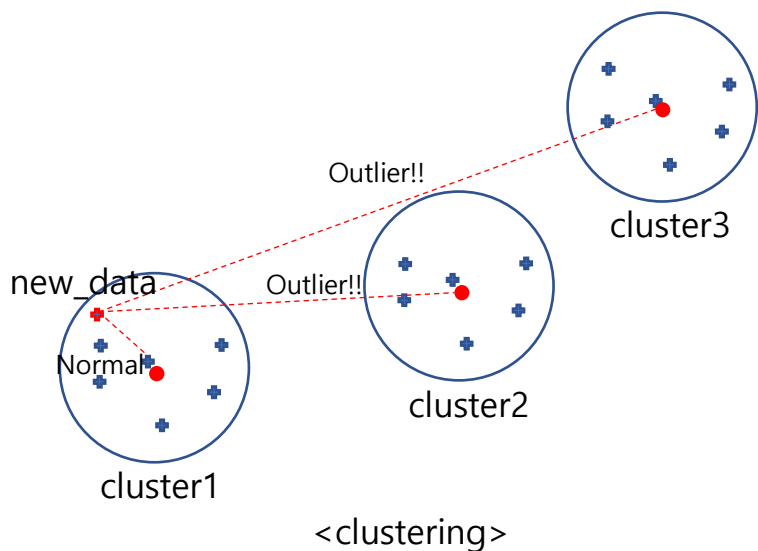
Cluster 밖에 있는 값이 이상치인 걸 판단하기 위해서는
새로운 값과 cluster의 평균 간의 거리가 해당 cluster의 범위를 넘을 경우로 판단



k-means data description

거리 계산

모든 cluster 간의 거리를 구하는 이유? 그리고 min 거리를 사용하는 이유?



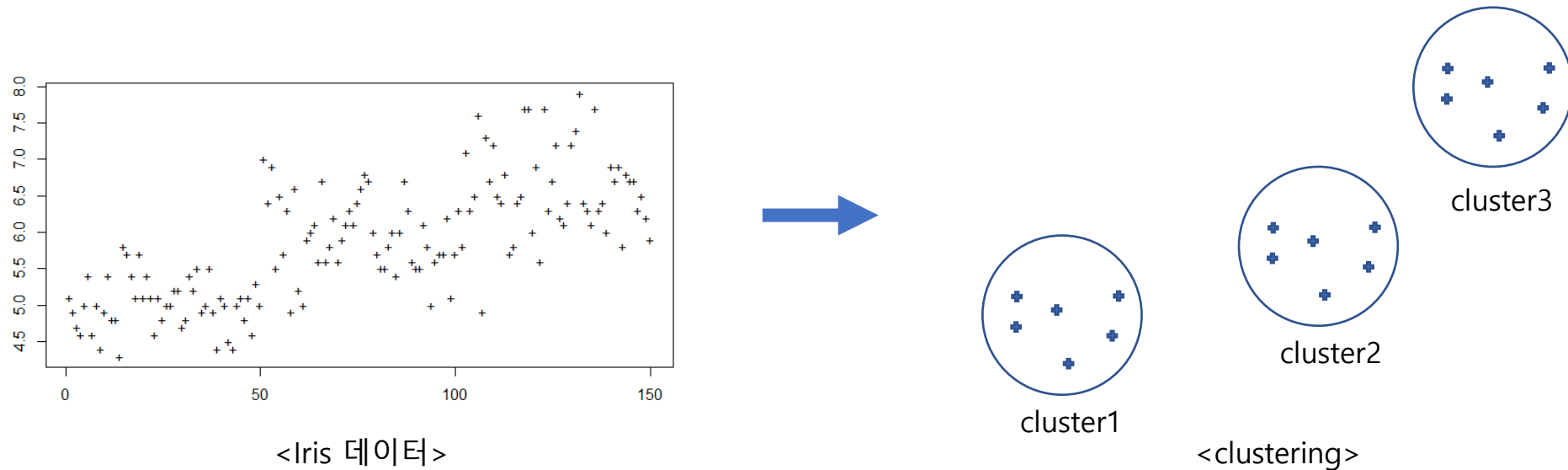
거리가 먼 클러스터와 비교할 경우 당연히 이상치로 나온다
거리가 가장 가까운 cluster와 비교했을 때 이상치 인지
아닌지를 판단해줘야 정확한 결과가 나온다.

-> 어떤 cluster를 기준으로 이상치를 판단할지 정하는 것

k-means data description step

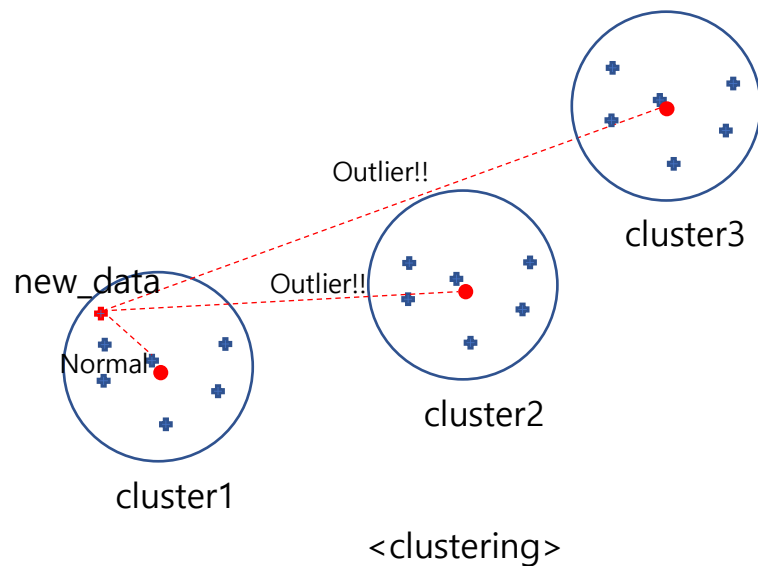
Step1 : 원 데이터를 clustering

Cluster가 대략 3개정도 나올 것 같으니까 3개의 cluster로 clustering



k-means data description

Step2: Cluster 별 각 distance를 구함

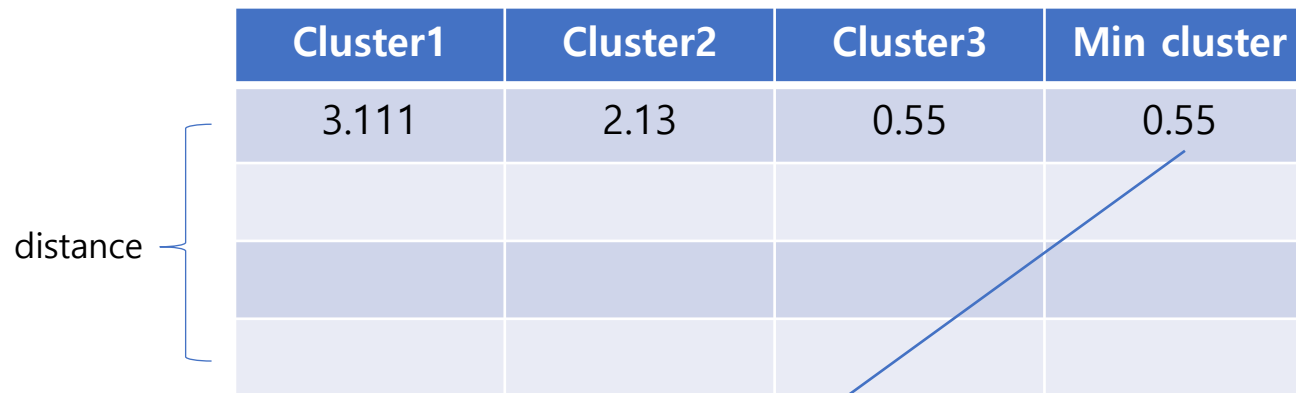


	Cluster1	Cluster2	Cluster3	Min cluster
distance				

k-means data description

Step2: Cluster 별 각 distance를 구함

min cluster 별 distance는 해당 관측치가 어디 cluster에 속해 있고 그 cluster와의 거리가 어디인지 알 수 있음.



	Cluster1	Cluster2	Cluster3	Min cluster
distance {	3.111	2.13	0.55	0.55

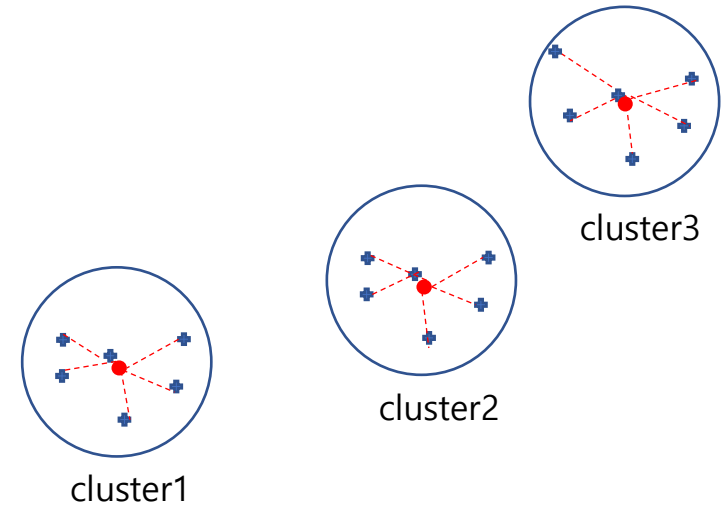
이 관측치는 3번째 cluster에 속해 있고 cluster3과의 거리가 0.55다.

k-means data description

max (min cluster)란?

distance {

Cluster1	Cluster2	Cluster3	Max? Min cluster
3.111	2.13	0.55	0.55



Cluster1	Cluster2	cluster3

<boundary>

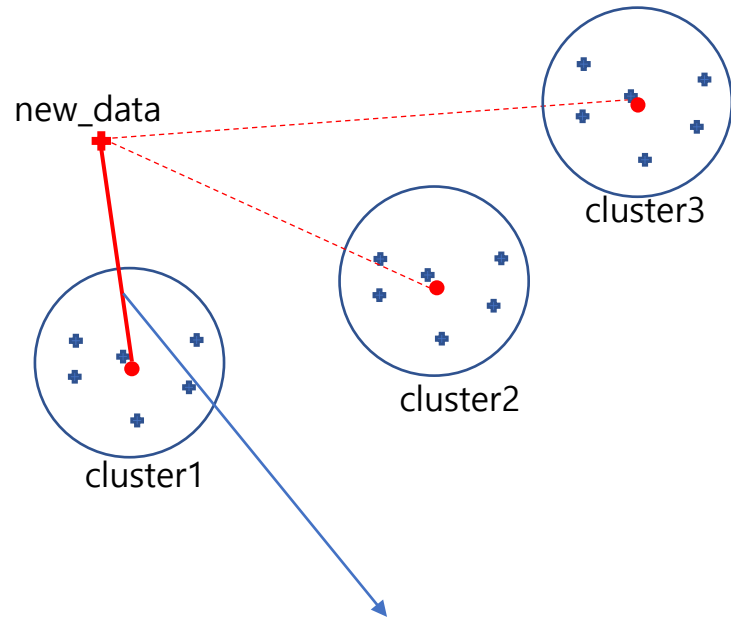
Cluster 별 최대 거리 -> boundary를 의미
이 boundary를 넘으면 이상치라는 것이다.

k-means data description

Step3: 새로운 관측치가 왔을 때 기존 cluster 별 거리를 구함

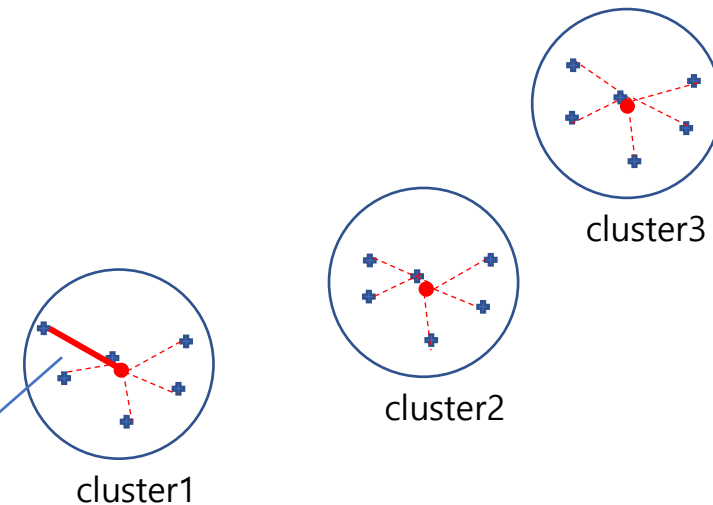
$\text{Min}(\text{distance}_{\text{cluster}}) > \text{boundary}_{\text{cluster}} \rightarrow \text{outlier}$

<test data와 기존 cluster 간의 거리>



<Train data로 학습 시킨 기존 cluster>

$>$



$\text{Min distance}_{\text{cluster1}} > \text{boundary}_{\text{cluster1}} \rightarrow \text{원 밖으로 벗어났다} \rightarrow \text{outlier}$

Thank You
